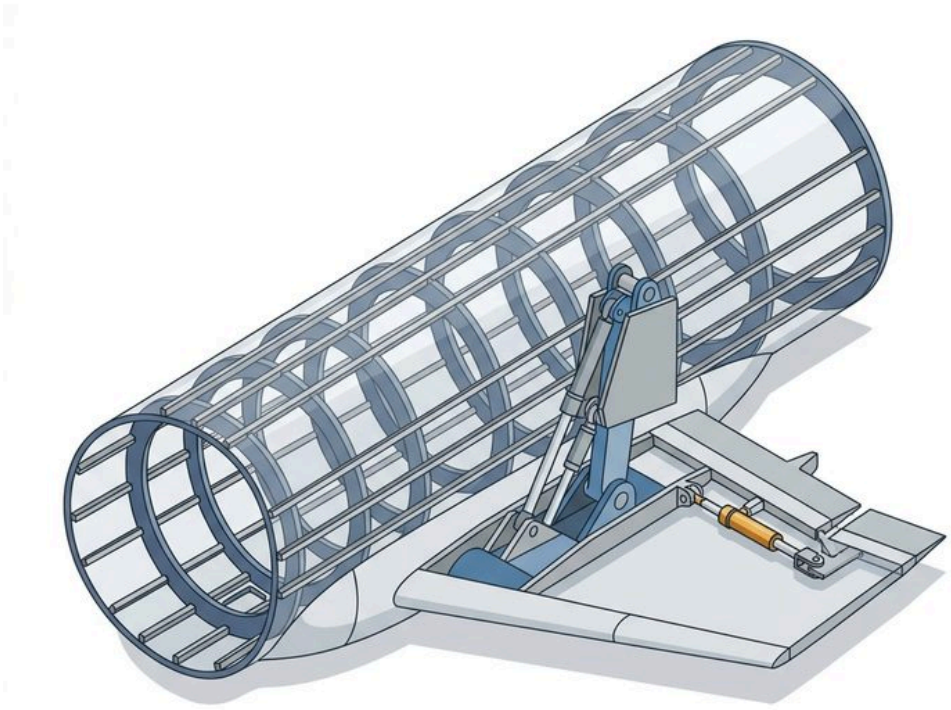
PART66ONLINE.COM

FREE SAMPLE

One chapter from each of the 17 modules



EASA Part-66 **Study Notes**

17 chapters • 356,046 words

Contents

1.1 Module 1 · Arithmetic

2.1 Module 2 · Matter

3.1 Module 3 · Electron Theory

4.1.1 Module 4 · Diodes

5.1 Module 5 · Electronic Instrument Systems

6.1 Module 6 · Aircraft Materials — Ferrous

7.1 Module 7 · Safety Precautions — Aircraft and Workshop

8.1 Module 8 · Physics of the Atmosphere

9.1 Module 9 · General

10.1 Module 10 · Regulatory Framework

11.1(a) Module 11 · Aeroplane Aerodynamics and Flight Controls

12.1 Module 12 · Theory of Flight — Rotary Wing Aerodynamics

13.1(a) Module 13 · Aeroplane Aerodynamics and Flight Controls

14.1(a) Module 14 · Turbine Engines

15.1 Module 15 · Fundamentals

16.1 Module 16 · Fundamentals

17.1 Module 17 · Fundamentals

Module 1 — Mathematics

1.1 — Arithmetic

Arithmetic is the foundation of all mathematics in EASA Part 66. This section covers the essential number skills you will use throughout your career in aircraft maintenance — from calculating fuel loads and torque values to converting between measurement systems. Category A requires Level 1 (basic familiarisation); categories B1, B2, B2L, and B3 require Level 2 (general knowledge, able to apply).

Module 1 is examined by multiple-choice questions with three options, of which exactly one is correct, and the pass mark is 75%. That format shapes how you should study arithmetic. The two wrong options are almost never random numbers: they are the answers you get from the specific slips a candidate under time pressure actually makes — dividing by 45 instead of by 0.75, adding numerators and denominators separately, dividing square centimetres by 100 instead of 10,000, or reading a fuel figure in pounds as though it were kilograms. Every section below therefore names the slip as well as the method, because recognising a wrong option for what it is will save you as many marks as knowing the right one.

Four habits do most of the work, and they carry through into the hangar as well as the exam room:

- **Read the units before the numbers.** Most arithmetic errors in aviation are unit errors wearing an arithmetic disguise. Decide what unit the answer must come out in before you start.
- **Estimate first.** A rough answer obtained in five seconds tells you the order of magnitude, and the order of magnitude is what a misplaced decimal point destroys.
- **Carry full precision, round once at the end.** Rounding part-way through and then continuing compounds the error, and on a multi-step conversion it can move the answer far enough to select the wrong option.
- **Answer the question that was asked.** A calculation can be flawless and still score zero if the question wanted litres and you gave kilograms, or wanted the radius and you gave the diameter.

The arithmetic here is not self-contained knowledge you use once and forget. It is the machinery every later module runs on: mass and balance moments in Module 7, Ohm's law and power in Module 3, stress and strain in Module 2, gear and pulley ratios in that same module, blade angles and areas in Module 17. Getting fluent here means you are never doing two hard things at once later, and it is the reason candidates who are shaky on Module 1 tend to be shaky everywhere.

> **How to use this note:** work every example with a pen before you read the solution. Arithmetic is a skill, not a body of facts — reading a worked example and understanding it is not the same as being able to reproduce it under time pressure, and only the second one is examined.

Arithmetical Terms and Signs

You must know the standard mathematical symbols and the terms associated with each operation.

Symbol	Meaning	Example
$+$	Addition (sum)	$8 + 5 = 13$
$-$	Subtraction (difference)	$15 - 7 = 8$
$\times$	Multiplication (product)	$6 \times 4 = 24$
$\div$	Division (quotient)	$20 \div 5 = 4$
$=$	Equals	$3 + 4 = 7$
$\neq$	Not equal to	$5 \neq 6$
$>$	Greater than	$10 > 7$
$<$	Less than	$3 < 8$
$\geq$	Greater than or equal to	$x \geq 0$
$\leq$	Less than or equal to	$T \leq 100$
$\approx$	Approximately equal to	$\pi \approx 3.14$
$\pm$	Plus or minus (tolerance)	50 ± 2 mm

Key terminology: In addition, the numbers being added are *addends* and the result is the *sum*. In subtraction, you subtract the *subtrahend* from the *minuend* to get the *difference*. In multiplication, the numbers are *factors* and the result is the *product*. In division, the *dividend* is divided by the *divisor* to give the *quotient* and possibly a *remainder*.

✈ **Aviation context:** Tolerances (using ±) are critical in maintenance. A torque specification of 50 ± 2 Nm means any value from 48 Nm to 52 Nm is acceptable.



A torque value is set within a tolerance: 50 +/- 2 Nm accepts anything from 48 Nm to 52 Nm.

Further Symbols You Will Meet

The symbols above are the arithmetic core. The following appear constantly in maintenance data, formulae and exam stems, and are worth recognising on sight.

Symbol	Meaning	Example
$\sqrt{a}$	Square root	$\sqrt{49} = 7$
$\sqrt[3]{a}$	Cube root	$\sqrt[3]{27} = 3$
a^n	Index, exponent or power	$4^3 = 64$
$\lvert a \rvert$	Absolute value (size, sign discarded)	$\lvert -6 \rvert = 6$
$\%$	Per cent (parts per hundred)	$25\% = 0.25$
$a : b$	Ratio of a to b	$20 : 60 = 1 : 3$
$\sum$	The sum of all the values	$\sum x = x_1 + x_2 + x_3$
$\propto$	Is proportional to	$A \propto r^2$
$\therefore$	Therefore (introduces a conclusion)	$\therefore V = 12$
$( )$ $[ ]$ $\{ \}$	Brackets: evaluate the contents first	$2 \times (3 + 4) = 14$

Round, square and curly brackets all mean exactly the same thing. Different shapes are used only so that the eye can pair them up when brackets are nested inside brackets; there is no rule that says one shape is worked before another, and the innermost pair is always evaluated first whatever its shape.

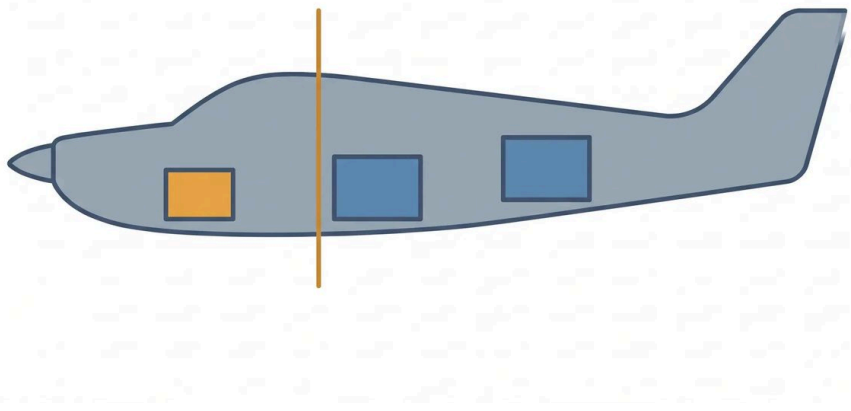
Signed (Directed) Numbers

A number can carry a direction as well as a size. On a number line, positive numbers run to the right of zero and negative numbers to the left, so –20 is *smaller* than –5 even though 20 is larger than 5. This catches candidates out with temperatures: an OAT of –40 °C is colder than –15 °C, and a limit written as $T \geq -55\,^{\circ}\mathrm{C}$ permits –50 °C but not –60 °C.

Keep the two uses of the minus sign apart. In $7 - 2$ the sign is an *operation*: take 2 away. In -2 it is part of the *number*: two units to the left of zero. When both appear together the brackets tell you which is which, as in $7 + (-2)$.

Situation	Rule	Example
Adding a negative	Becomes a subtraction	$7 + (-2) = 5$
Subtracting a negative	Becomes an addition	$(+3) - (-4) = +7$
Multiply or divide, signs alike	Result is positive	$(-4) \times (-3) = +12$
Multiply or divide, signs unlike	Result is negative	$(-4) \times 3 = -12$
Negative base, even power	Result is positive	$(-2)^4 = 16$
Negative base, odd power	Result is negative	$(-2)^3 = -8$

The "two like signs give a positive" rule is not an arbitrary convention to be memorised. Read $(-4) \times (-3)$ as "the opposite of four lots of negative three". Four lots of -3 is -12; the opposite of -12 is $+12$. The same reasoning explains subtracting a negative: taking away a debt leaves you better off, so $(+3) - (-4)$ moves you four units further to the right, to $+7$.



With the datum inside the aircraft, items ahead of it have negative arms and items behind it positive arms.

Worked Example — Signs in a Moment Calculation

Mass and balance is where signed arithmetic earns its keep. Arms are measured from the datum; the datum is frequently placed ahead of the nose precisely so that every arm comes out positive, but where it falls inside the aircraft, items ahead of it have negative arms. Three items are loaded:

80 kg at $+2.5$ m, 60 kg at -1.2 m, 25 kg at $+4.0$ m.

Moments: $80 \times 2.5 = +200$, $60 \times (-1.2) = -72$, $25 \times 4.0 = +100$ kg·m.

Total moment $= 200 - 72 + 100 = 228$ kg·m. Total mass $= 80 + 60 + 25 = 165$ kg.

Centre of gravity $= \frac{228}{165} = 1.38$ m aft of the datum.

Drop the one minus sign and the moment total becomes $200 + 72 + 100 = 372$ kg·m, putting the CG at $\frac{372}{165} = 2.25$ m — an error of nearly 0.9 m from a single missing sign. **The sign is part of the number, not decoration.**

The Order of Operations

An expression such as $8 + 4 \times 3$ has only one correct value, because there is a fixed order in which operations are carried out. Work it left to right and you get 36; obey the order and you get 20. The convention, from first to last, is:

- **Brackets** — innermost pair first, working outwards.
- **Orders** — powers and roots.
- **Division and multiplication** — these rank *equally*, worked left to right.
- **Addition and subtraction** — these also rank equally, worked left to right.

You will meet this as **BODMAS**, **BIDMAS** or **PEMDAS**. These are three mnemonics for one and the same convention, not three competing ones: they differ only in the word chosen for brackets (Brackets or Parentheses) and for powers (Orders, Indices or Exponents). Any of them, applied properly, gives the same answer.

The trap inside the mnemonic: the letters put D before M and A before S, which tempts candidates into doing all the dividing before any multiplying. That is wrong. $12 \div 2 \times 3$ is worked left to right as $(12 \div 2) \times 3 = 18$; doing the multiplication first gives $12 \div 6 = 2$. Likewise $10 - 4 + 2$ is $(10 - 4) + 2 = 8$, not $10 - 6 = 4$. Equal rank means left to right, every time.

Three symbols group their contents as though brackets were printed around them, and forgetting this is a reliable source of wrong answers:

- A **fraction bar** groups the whole numerator and the whole denominator. In $\frac{6+4}{2}$ the addition is completed first, giving $\frac{10}{2} = 5$ — not $6 + 2$.

- A **radical sign** groups everything under it, so $\sqrt{25 + 144}$ means the root of 169, which is 13. It is not the sum of the two separate roots, a point returned to in the section on roots below.
- A number written directly against a bracket means **multiply**. $2(5 - 3)$ is $2 \times 2 = 4$, and the multiplication cannot happen until the bracket is resolved.

Worked Example — Nested Brackets

Evaluate $4\{2(5 - 1) - 3\} + 8$:

Innermost bracket: $5 - 1 = 4$. Then $2 \times 4 = 8$, and $8 - 3 = 5$, which clears the curly bracket.

Now $4 \times 5 = 20$, and finally $20 + 8 = \mathbf{28}$.

Evaluate $8 + 4[5 \times 2(5 - 9 \div 3)]$:

Inside the round bracket, division outranks subtraction: $9 \div 3 = 3$, so $5 - 3 = 2$.

Inside the square bracket: $5 \times 2 \times 2 = 20$. Then $4 \times 20 = 80$, and $8 + 80 = \mathbf{88}$.

Worked Example — Order of Operations with Negatives

Evaluate $(6 - 2) - \dfrac{6 - 9}{-3 + (-3)}$:

Brackets first: $6 - 2 = 4$, $6 - 9 = -3$, and $-3 + (-3) = -6$.

The division outranks the subtraction, and two like signs give a positive: $\frac{-3}{-6} = +0.5$.

Finally $4 - 0.5 = \mathbf{3\frac{1}{2}}$.

Evaluate $15.4 \div 2 - 2(6.2 - 15.6)$:

$15.4 \div 2 = 7.7$. Inside the bracket, $6.2 - 15.6 = -9.4$, so $2 \times (-9.4) = -18.8$.

The expression is now $7.7 - (-18.8)$. Subtracting a negative adds: $7.7 + 18.8 = \mathbf{26.5}$.

Miss that last step and you get $7.7 - 18.8 = -11.1$, which is exactly the wrong option such a question will offer you.

The Laws of Arithmetic

Three laws govern which rearrangements of an expression are legitimate. They matter because rearranging is how you make arithmetic easy enough to do reliably in your head or on paper.

- **Commutative law:** order does not matter for addition or multiplication. $a + b = b + a$ and $a \times b = b \times a$. It does *not* hold for subtraction or division: $7 - 3 = 4$ but $3 - 7 = -4$, and $8 \div 2 = 4$ but $2 \div 8 = 0.25$.
- **Associative law:** grouping does not matter for addition or multiplication. $(a + b) + c = a + (b + c)$. This is why a column of masses can be totalled in any order — useful when you want to pair figures that add to round numbers.
- **Distributive law:** $a(b + c) = ab + ac$. Multiplication distributes across a sum, which is the reason long multiplication works at all.

The distributive law is the most useful of the three in practice. To find 98×7, rewrite 98 as $(100 - 2)$: $100 \times 7 = 700$, $2 \times 7 = 14$, so the answer is $700 - 14 = 686$. To find 7×43, split the 43: $7 \times 40 = 280$ and $7 \times 3 = 21$, giving 301. Neither needs paper.

Two special numbers complete the picture. Adding zero or multiplying by one changes nothing, so both can be inserted anywhere they are convenient — a fact used heavily in unit conversion, where you multiply by a fraction whose value is exactly 1. Multiplying by zero always gives zero. **Dividing by zero is not defined** and has no answer: a calculator will report an error, and in a formula it signals that the quantity in the denominator cannot physically be zero.

Kinds of Number

- **Natural numbers:** the counting numbers 1, 2, 3, ... — rivets, blades, cylinders.
- **Integers:** the whole numbers including zero and the negatives: ..., –2, –1, 0, 1, 2, ...
- **Rational numbers:** anything that can be written as one integer divided by another, such as $\frac{3}{4}$ or $\frac{7}{6}$. Every terminating decimal and every recurring decimal is rational.
- **Irrational numbers:** those that cannot. Their decimal expansion never terminates and never settles into a repeating pattern. π and $\sqrt{2}$ are the two you will meet most.
- **Real numbers:** the rationals and irrationals together — every value that can appear on a number line.

This has a practical consequence rather than being a piece of classification for its own sake. Every measurement you can take is rational, because an instrument reads to a finite number of figures. Any answer that involves π or a square root is therefore an approximation the moment you write it down as a decimal, and how far you take it decides how much rounding error you carry forward.

> **Exam technique:** when an expression mixes brackets, division and negatives, resist the urge to work it in one line. Write each stage on its own line, exactly as the worked examples above do. The marks are not awarded for speed, and almost every wrong option offered in this part of the syllabus is the answer produced by collapsing two stages into one.

Methods of Multiplication and Division

Long Multiplication

To multiply large numbers, break one factor into its place-value components and add the partial products.

Worked Example

Calculate 247×36:

```
247 × 6  = 1,482
247 × 30 = 7,410
Total = 1,482 + 7,410 = 8,892
```

Long Division

Long division breaks a division problem into a sequence of smaller steps: divide, multiply, subtract, bring down.

Worked Example

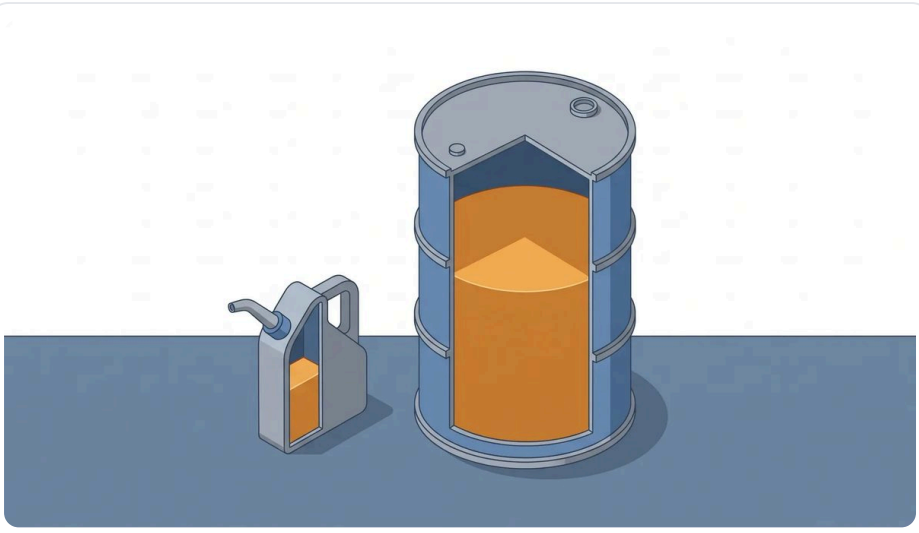
Calculate $1,596 \div 12$:

12 goes into 15 once (12), remainder 3. Bring down 9 → 39. 12 goes into 39 three times (36), remainder 3. Bring down 6 → 36. 12 goes into 36 exactly 3 times. **Answer: 133**

Estimation and Cross-checking

Always estimate before calculating. Round each number to a convenient nearby value: $247 \times 36 \approx 250 \times 40 = 10{,}000$. Your exact answer of 8,892 is reasonable. If you got 88,920, you would know something was wrong.

Common mistake: Misplacing the decimal point. A turbine oil capacity of 4.5 litres versus 45 litres is a critical difference. Always estimate first.



Misplacing the decimal point turns 4.5 litres of turbine oil into 45 litres: a jug becomes a drum.

Why Long Multiplication Works

Splitting a factor into partial products is not a trick; it is the distributive law applied to place value. The multiplier 36 is worth $30 + 6$, so

Partial Products

$$247 \times 36 = 247 \times (30 + 6) = (247 \times 30) + (247 \times 6)$$

Understanding it this way tells you where the zeros come from. The row for the tens digit is shifted one place left because you are multiplying by 30, not by 3; the row for a hundreds digit is shifted two places, because you are multiplying by 300. Candidates who memorise "put a zero on the second line" get the third line wrong the first time they meet a three-digit multiplier.

Worked Example — Three-Digit Multiplier

Calculate 428×315:

```
  428 × 300 = 128,400
  428 ×  10 =   4,280
  428 ×   5 =   2,140
  ─────────────────────
  Total = 128,400 + 4,280 + 2,140 = 134,820
```

Sanity check: $400 \times 300 = 120{,}000$, so an answer in the low hundred-thousands is right and an answer of 13,482 or 1,348,200 is not.

Multiplying and Dividing by Powers of Ten

Because our number system is base 10, multiplying by 10^n moves the decimal point n places to the *right* and dividing by 10^n moves it n places to the *left*. Nothing else changes: the digits keep their order and only their place value shifts. So $4.85 \times 1000 = 4850$ and $4.85 \div 100 = 0.0485$. Every metric conversion in this module is an application of that one fact, which is why metric arithmetic is so much less error-prone than imperial.

Multiplying Decimals

Ignore the decimal points, multiply the digits as whole numbers, then place the point so that the answer has as many decimal places as the two factors had between them. The rule follows directly from writing the decimals as fractions:

Where the Decimal Point Goes

$$0.75 \times 0.003 = \frac{75}{100} \times \frac{3}{1000} = \frac{225}{100\,000} = 0.00225$$

Two decimal places plus three decimal places gives five in the answer.

Worked Example — The Same Digits, Two Answers

Calculate $\frac{3}{4} \times 0.82$ and $\frac{3}{4} \times 82$:

Convert the fraction: $\frac{3}{4} = 0.75$. Multiply the digits: $75 \times 82 = 6150$.

For 0.75×0.82 there are $2 + 2 = 4$ decimal places, so the answer is $0.6150 = \mathbf{0.615}$.

For 0.75×82 there are $2 + 0 = 2$ decimal places, so the answer is $\mathbf{61.5}$ — exactly one hundred times the first, because the second factor is one hundred times larger.

Cross-check without decimals at all: $82 \div 4 = 20.5$, and $20.5 \times 3 = 61.5$.

Dividing by a Decimal

Long division is only comfortable when the divisor is a whole number, so make it one. Multiply the divisor *and* the dividend by the same power of ten — that leaves the quotient unchanged, because you have multiplied top and bottom of the same fraction by the same thing.

Worked Example

Calculate $8.892 \div 0.12$:

Multiply both by 100: $889.2 \div 12$.

12 into 88 goes 7 times (84), remainder 4. Bring down 9 → 49; 12 goes 4 times (48), remainder 1. Bring down 2 → 12; 12 goes exactly 1 time. Answer **74.1**.

Check by reversing: $74.1 \times 0.12 = 8.892$.

Two results that feel wrong but are not: dividing by a number smaller than 1 makes the answer *larger* than what you started with — $10 \div 0.5 = 20$, because there are twenty halves in ten. Multiplying by a number smaller than 1 makes it *smaller* — $10 \times 0.5 = 5$. If a division has produced a smaller number than the dividend, your divisor was greater than 1; if it has not, check that it was.

Division That Does Not Come Out Exactly

The long-division example above divides exactly. Most real ones do not, and you then have a choice about how to express what is left over — a choice the question usually decides for you.

Worked Example — Remainder, Fraction or Decimal

Calculate $1{,}745 \div 12$:

12 into 17 goes 1 (12), remainder 5. Bring down 4 → 54; 12 goes 4 times (48), remainder 6. Bring down 5 → 65; 12 goes 5 times (60), remainder 5.

As a whole number with a remainder: **145 remainder 5**. Check: $12 \times 145 = 1740$, and $1740 + 5 = 1745$.

As a mixed number: $145\frac{5}{12}$. As a decimal, carry on past the point: $\frac{5}{12} = 0.41\dot{6}$, so the answer is 145.4167 to four decimal places.

Which form you want depends on the question. If 1,745 fasteners are issued in packs of 12, the useful answer is 145 full packs with 5 loose — you cannot draw 0.4167 of a pack.

Short Cuts Worth Knowing

Each of these is the distributive or associative law in disguise, and each turns a written calculation into a mental one.

To do this	Do this instead	Example
$\times 5$	$\times 10$, then halve	$86 \times 5 = \frac{860}{2} = 430$
$\div 5$	Double, then $\div 10$	$93 \div 5 = \frac{186}{10} = 18.6$
$\times 25$	$\times 100$, then $\div 4$	$36 \times 25 = \frac{3600}{4} = 900$

To do this	Do this instead	Example
$\times 9$	$\times 10$, then subtract the number	$47 \times 9 = 470 - 47 = 423$
$\times 11$ (two digits)	Split the digits, insert their sum	$43 \times 11 = 4 \mid 7 \mid 3 = 473$
Awkward pair	Halve one factor, double the other	$16 \times 35 = 8 \times 70 = 560$

Checking an Answer Three Ways

Estimation, covered above, is the first line of defence. Two further checks cost seconds and between them catch most of what estimation misses.

Bracket the answer. Instead of one estimate, produce a deliberate under-estimate and a deliberate over-estimate, and confirm the exact answer sits between them. For 384×27, rounding both factors down gives $380 \times 20 = 7{,}600$ and rounding both up gives $390 \times 30 = 11{,}700$. The exact answer, $384 \times 27 = 10{,}368$, lies inside that band, so it is credible. A bracket is stronger than a single estimate because it tells you not only the order of magnitude but whether you are in the right part of it.

Reverse the operation. A product divided by one of its factors must return the other; a quotient multiplied by the divisor must return the dividend. This is the only check that verifies the answer exactly rather than approximately, and it is worth the extra few seconds on any figure you are about to enter on a worksheet.

Cast out nines. Replace each number by its repeated digit sum, do the same arithmetic on those small numbers, and the answers must agree. For $247 \times 53 = 13{,}091$: $2 + 4 + 7 = 13 \to 4$, and $5 + 3 = 8$; $4 \times 8 = 32 \to 5$. The product's digits give $1 + 3 + 0 + 9 + 1 = 14 \to 5$, which matches. Had you written 13,081 the digit sum would be 4, and the error would be exposed at once.

> ⚠ **Know what each check cannot see:** casting out nines is blind to a transposition, because swapping two digits leaves the digit sum unchanged — 13,091 and 13,901 both reduce to 5. It is equally blind to a misplaced decimal point. Estimation catches both of those but will not notice a small error in the last digit. Use the two together, or reverse the operation, which catches everything.

Worked Example — Average Fuel Consumption

An aircraft uplifts fuel into four tanks: 2,450, 2,450, 1,180 and 1,180 litres. It burns the lot over a sector of 3 hours 20 minutes. What was the average consumption in litres per hour?

Total uplift $= 2(2450) + 2(1180) = 4900 + 2360 = 7{,}260$ litres.

The time must be in hours, not hours and minutes: 20 minutes is $\frac{20}{60} = \frac{1}{3}$ hour, so the sector is $3\frac{1}{3} = \frac{10}{3}$ hours.

$7260 \div \frac{10}{3} = 7260 \times \frac{3}{10} = \mathbf{2{,}178}$ litres per hour.

Dividing by 3.20 instead — treating "3 hours 20 minutes" as a decimal — would give 2,269 L/h, an error of over 4%. Minutes are sixtieths, not hundredths.

Fractions and Decimals

Types of Fractions

- **Proper fraction:** numerator < denominator, e.g. $\frac{3}{4}$
- **Improper fraction:** numerator ≥ denominator, e.g. $\frac{7}{4}$
- **Mixed number:** whole number + fraction, e.g. $1\frac{3}{4}$

Operations with Fractions

Key Formulae

Addition/Subtraction (find a common denominator):

$$\frac{a}{b} + \frac{c}{d} = \frac{ad + bc}{bd}$$

Multiplication:

$$\frac{a}{b} \times \frac{c}{d} = \frac{a \times c}{b \times d}$$

Division (invert and multiply):

$$\frac{a}{b} \div \frac{c}{d} = \frac{a}{b} \times \frac{d}{c} = \frac{ad}{bc}$$

Worked Example

Calculate $\frac{2}{3} + \frac{3}{5}$:

Common denominator = 15. So $\frac{2}{3} = \frac{10}{15}$ and $\frac{3}{5} = \frac{9}{15}$. Sum = $\frac{10+9}{15} = \frac{19}{15} = 1\frac{4}{15}$

Decimals

Decimals are another way of expressing fractions in base 10. The digit positions to the right of the decimal point represent tenths, hundredths, thousandths, etc.

Converting fractions to decimals: Divide the numerator by the denominator. For example, $\frac{3}{8} = 3 \div 8 = 0.375$.

Converting decimals to fractions: Write the decimal over the appropriate power of 10 and simplify. For example, $0.625 = \frac{625}{1000} = \frac{5}{8}$.

Fraction	Decimal	Fraction	Decimal
$\frac{1}{2}$	0.5	$\frac{1}{8}$	0.125
$\frac{1}{3}$	0.333...	$\frac{3}{8}$	0.375
$\frac{1}{4}$	0.25	$\frac{5}{8}$	0.625
$\frac{3}{4}$	0.75	$\frac{7}{8}$	0.875
$\frac{1}{5}$	0.2	$\frac{1}{16}$	0.0625

Decimal Places and Significant Figures

Decimal places (d.p.): the number of digits after the decimal point. For example, 3.142 has 3 d.p.

Significant figures (s.f.): count all digits starting from the first non-zero digit. For example, 0.00456 has 3 s.f.; 3,050 has 3 s.f. (the trailing zero may or may not be significant depending on context).

Equivalent Fractions and Lowest Terms

A fraction is a division waiting to happen, and multiplying or dividing *both* parts of it by the same number leaves its value untouched — because that is the same as multiplying the whole fraction by 1. So $\frac{3}{4} = \frac{6}{8} = \frac{9}{12} = \frac{75}{100}$; all four are the same quantity written with different-sized pieces.

Going the other way is **cancelling**, and dividing top and bottom by their highest common factor puts the fraction in its **lowest terms** in one move. Take $\frac{84}{126}$: since $84 = 2^2 \times 3 \times 7$ and $126 = 2 \times 3^2 \times 7$, their highest common factor is $2 \times 3 \times 7 = 42$, and $\frac{84 \div 42}{126 \div 42} = \frac{2}{3}$. Cancelling in smaller steps reaches the same place, just more slowly. An answer left uncancelled is not wrong, but an exam option list will normally offer only the cancelled form, so you must be able to recognise it.

To decide which of two fractions is larger without converting either to a decimal, **cross-multiply**: compare $\frac{5}{8}$ with $\frac{7}{11}$ by forming $5 \times 11 = 55$ against $7 \times 8 = 56$. The larger product sits above the larger fraction, so $\frac{7}{11}$ is the bigger of the two. This works because you are effectively putting both over the common denominator 8×11 without bothering to write it.

Choosing the Denominator When You Add

The addition formula above always works, because bd is always *a* common denominator. It is not always the *lowest* one, so the answer it produces often needs cancelling afterwards. Using the lowest common multiple of the denominators instead keeps the numbers small and usually delivers the answer already in lowest terms.

Worked Example — Two Routes to the Same Answer
Calculate $\frac{11}{16} + \frac{5}{8}$.
Lowest common denominator: 8 divides into 16, so the LCD is 16. $\frac{5}{8} = \frac{10}{16}$, and $\frac{11}{16} + \frac{10}{16} = \frac{21}{16}$.
Formula route: $\frac{(11\times8)+(5\times16)}{16\times8} = \frac{88+80}{128} = \frac{168}{128}$. Cancelling by 8 gives $\frac{21}{16}$ — the same result, after an extra step.
As a mixed number that is $1\frac{5}{16}$. Both $\frac{21}{16}$ and $1\frac{5}{16}$ are correct; read the options to see which form is wanted.

Never add the tops and bottoms separately. $\frac{1}{2} + \frac{1}{2}$ is 1, not $\frac{2}{4}$. A second version of the same error is to *multiply* when the question says add: $\frac{11}{16} \times \frac{5}{8} = \frac{55}{128}$, and $\frac{55}{128}$ is precisely the wrong option a question on $\frac{11}{16} + \frac{5}{8}$ will put in front of you. Both slips are recognisable on sight once you know to look for them.

Mixed Numbers in a Calculation

The safe method with mixed numbers is to convert every term to an improper fraction first, work entirely in improper fractions, and convert back only at the very end. To convert, multiply the whole number by the denominator and add the numerator: $4\frac{3}{8} = \frac{(4\times8)+3}{8} = \frac{35}{8}$. To convert back, divide and keep the remainder over the denominator.

Worked Example — Mixed Numbers, Mixed Signs

Calculate $4\frac{3}{8} - 2\frac{1}{4} + \frac{1}{8}$:

Convert: $4\frac{3}{8} = \frac{35}{8}$ and $2\frac{1}{4} = \frac{9}{4} = \frac{18}{8}$. The LCD is 8 throughout.

$\frac{35}{8} - \frac{18}{8} + \frac{1}{8} = \frac{35-18+1}{8} = \frac{18}{8}$, which cancels by 2 to $\frac{9}{4} = \mathbf{2\frac{1}{4}}$.

Calculate $\frac{9}{4} + \frac{5}{12} + 5\frac{1}{8}$:

Denominators 4, 12 and 8 are 2^2, $2^2 \times 3$ and 2^3, so the LCD is $2^3 \times 3 = 24$.

$\frac{9}{4} = \frac{54}{24}$, $\frac{5}{12} = \frac{10}{24}$, and the fractional part of $5\frac{1}{8}$ is $\frac{3}{24}$.

Fractions: $\frac{54+10+3}{24} = \frac{67}{24} = 2\frac{19}{24}$. Adding the whole 5 gives $\mathbf{7\frac{19}{24}}$.

Subtraction of mixed numbers is where the converting-first rule pays for itself. $5\frac{1}{4} - 2\frac{3}{4}$ cannot be done by taking the fractions apart, because $\frac{1}{4}$ is smaller than $\frac{3}{4}$. As improper fractions it is simply $\frac{21}{4} - \frac{11}{4} = \frac{10}{4} = 2\frac{1}{2}$. If you prefer to keep the whole numbers separate you must borrow one whole from the 5, rewriting it as $4\frac{5}{4}$, and then $4\frac{5}{4} - 2\frac{3}{4} = 2\frac{2}{4} = 2\frac{1}{2}$. Both reach the same answer; the first route has fewer places to go wrong.

Multiplying and Dividing: Cancel Before You Multiply

Multiplication of fractions needs no common denominator, and cancelling diagonally before multiplying keeps the numbers small. In $\frac{3}{8} \times \frac{4}{9}$, the 4 and the 8 share a factor of 4, and the 3 and the 9 share a factor of 3, leaving $\frac{1}{2} \times \frac{1}{3} = \frac{1}{6}$. Multiplying first gives $\frac{12}{72}$, which cancels to the same $\frac{1}{6}$ with more effort.

The word **"of"** means multiply. "Three quarters of 82" is $\frac{3}{4} \times 82$, worked in the section above. This is worth fixing in your mind, because most percentage and ratio questions are phrased with "of" rather than with a multiplication sign.

Dividing by a fraction is done by inverting the second fraction and multiplying, and the reason is worth knowing rather than memorising: $\frac{c}{d} \times \frac{d}{c} = 1$, so multiplying by $\frac{d}{c}$ exactly undoes a multiplication by $\frac{c}{d}$. Read $\frac{3}{4} \div \frac{2}{5}$ as "how many two-fifths fit into three-quarters"; the answer $\frac{3}{4} \times \frac{5}{2} = \frac{15}{8} = 1\frac{7}{8}$ says a bit less than two, and the reason is worth being precise about: $\frac{2}{5}$ is smaller than $\frac{3}{4}$, so at least one of them fits, but it is also more than *half* of $\frac{3}{4}$ — half of three-quarters is $\frac{3}{8} = 0.375$, while $\frac{2}{5} = 0.4$ — so a second one cannot fit and the answer has to land between one and two.

Terminating and Recurring Decimals

Dividing the numerator by the denominator either stops or repeats — it can never wander on at random, because there are only so many possible remainders. Which of the two happens is completely predictable:

The rule: once a fraction is in its lowest terms, its decimal *terminates* if the denominator's only prime factors are 2s and 5s, and *recurs* otherwise. $\frac{3}{8}$ terminates because $8 = 2^3$; $\frac{5}{16}$ terminates because $16 = 2^4$; $\frac{7}{20}$ terminates because $20 = 2^2 \times 5$. $\frac{1}{6}$ recurs because 6 carries a factor of 3, and $\frac{1}{7}$ recurs because 7 is neither 2 nor 5. This is simply the statement that the denominator must divide into some power of 10, and $10 = 2 \times 5$.

A recurring decimal is marked with a dot over the repeating digit, or over the first and last digits of a repeating group: $\frac{1}{3} = 0.\dot{3}$, $\frac{1}{6} = 0.1\dot{6}$, and $\frac{1}{7} = 0.\dot{1}42857\dot{}$, whose block of six digits then repeats for ever. A mixed number behaves the same way: $\frac{7}{6} = 1 + \frac{1}{6} = 1.1\dot{6}$, a whole 1 followed by the recurring sixth. Recurring decimals are still rational numbers — they came from a fraction and can be turned back into one.

Turning a recurring decimal back into a fraction is a short piece of arithmetic rather than a lookup. Call the decimal x, multiply it by whichever power of ten shifts one full repeating block to the left of the point, then subtract the original — the endless tails are identical and cancel exactly. For $x = 0.\dot{3}$: $10x = 3.\dot{3}$, so $10x - x = 3$, giving $9x = 3$ and $x = \frac{3}{9} = \frac{1}{3}$. For a two-digit block, $x = 0.\dot{3}\dot{6}$: $100x = 36.\dot{3}\dot{6}$, so $99x = 36$ and $x = \frac{36}{99} = \frac{4}{11}$. Check the second one by dividing 4 by 11, which returns 0.363636... as it should.

Rounding: How and When

Rounding to a stated number of **decimal places** means looking at the digit one place beyond the last one you are keeping. If it is 5 or more, increase the last kept digit by one; if it is 4 or less, leave it alone. So $\pi = 3.14159\ldots$ is 3.14 to 2 d.p. and 3.142 to 3 d.p.

Rounding to a stated number of **significant figures** works identically, except that you start counting from the first non-zero digit rather than from the decimal point. So 0.0045678 is 0.00457 to 3 s.f., and 47,382 is 47,000 to 2 s.f. or 47,400 to 3 s.f. The zeros that replace discarded digits in a whole number are placeholders, not new information.

Simply cutting off the surplus digits is **truncation**, not rounding, and it gives a different result: 2.678 truncates to 2.67 but rounds to 2.68. When a question asks for an answer "to two decimal places" it means rounding.

Worked Example — Round Once, at the End

Convert 1,750 US gallons of fuel to litres.

Using the full factor: $1750 \times 3.785411784 = 6624.47$ litres.

Rounding the factor to 3.8 first: $1750 \times 3.8 = 6650$ litres.

The premature rounding has introduced an error of about 25 litres — only 0.4%, but it grew out of a change in the **second** significant figure of the conversion factor, because 3.785411784 was cut back to 3.8, which carries only 2 s.f. A change of one unit in the fourth significant figure would have cost nearer 0.03%. Carry the full figure through the arithmetic and round the **answer**, never the inputs.

Significant Figures as a Statement About Precision

Significant figures are not only a rounding convention; they are how a written number declares how well it is known. The rules for counting them are:

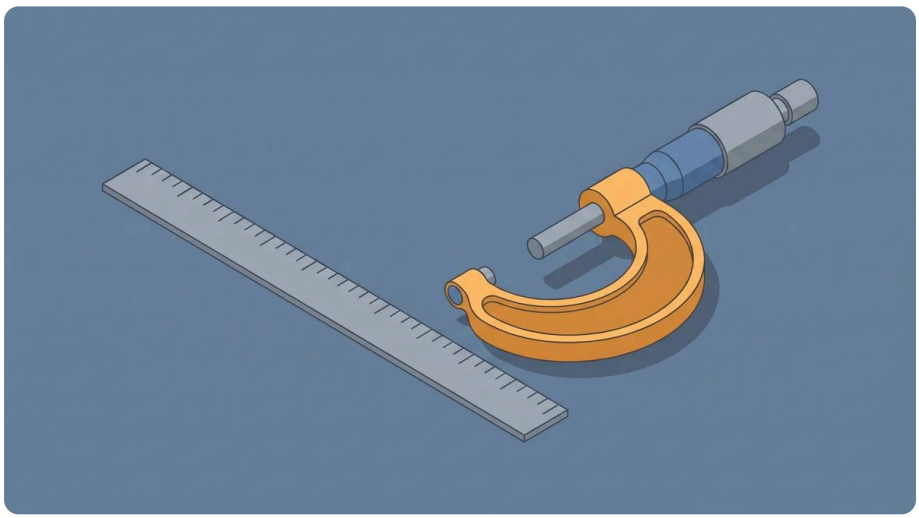
- **Leading zeros are never significant.** They only position the decimal point, which is why 0.00456 carries 3 s.f. and not 5.
- **Zeros between non-zero digits always are.** 3.05 has 3 s.f.
- **Trailing zeros after a decimal point are significant.** 2.50 mm has 3 s.f. and says the measurement was taken to hundredths of a millimetre; 2.5 mm has 2 s.f. and does not.
- **Trailing zeros in a whole number are ambiguous** unless the context settles it — which is exactly why scientific notation exists.

Writing a number in the form $a \times 10^n$ removes the ambiguity completely, because only the digits in a are shown: 3.05×10^3 is 3 s.f. while 3.050×10^3 is 4 s.f., and both are the value 3,050. The mechanics of that notation are covered in the final section of this note.

Two conventions govern how precision survives a calculation, and they are not the same rule:

- For **multiplication and division**, quote the answer to the number of *significant figures* of the least precise input. $12.34 \times 2.1 = 25.914$, but 2.1 carries only 2 s.f., so the honest answer is 26.
- For **addition and subtraction**, quote the answer to the number of *decimal places* of the least precise input. $12.34 + 2.1 = 14.44$, quoted as 14.4.

The practical version of this is that a calculator's ten digits are not ten digits of knowledge. A dimension read from a steel rule and one read from a micrometer are different measurements even when both happen to display as 2.5 mm, and multiplying a rule reading by a micrometer reading cannot produce micrometer-grade accuracy. Quoting more figures than your least precise instrument justifies is a claim you cannot support.



A steel rule and a micrometer: both may show 2.5 mm, but they are different measurements.

Fractions of an Inch in the Hangar

Imperial drill, reamer and fastener sizes step in halves, quarters, eighths, sixteenths, thirty-seconds and sixty-fourths of an inch, so converting a fraction to a decimal and on to millimetres is routine work rather than an exam exercise. The millimetre column below is exact, because the inch is *defined* as exactly 25.4 mm — the conversion arithmetic itself is set out in the units section.



Imperial drill sizes step in fractions of an inch, so converting them is routine hangar work.

Fraction of an inch	Decimal inches	Millimetres
$\frac{1}{16}$	0.0625	1.5875
$\frac{3}{16}$	0.1875	4.7625
$\frac{5}{16}$	0.3125	7.9375

Fraction of an inch	Decimal inches	Millimetres
$\frac{7}{16}$	0.4375	11.1125
$\frac{9}{16}$	0.5625	14.2875
$\frac{11}{16}$	0.6875	17.4625
$\frac{13}{16}$	0.8125	20.6375
$\frac{15}{16}$	0.9375	23.8125

The odd sixteenths are listed because the even ones reduce to eighths, quarters and halves that you already have. Notice how the pattern is generated: one sixteenth is 0.0625 in, so the whole scale can be rebuilt by repeated addition if you forget it. The same arithmetic run backwards identifies the nearest imperial size to a metric one. A 6 mm dimension is $6 \div 25.4 = 0.2362$ in, which falls between $\frac{7}{32}$ (0.21875 in) and $\frac{1}{4}$ (0.25 in) and is closest to $\frac{15}{64}$ (0.234375 in) — a difference of under 0.05 mm. That arithmetic tells you which imperial size is nearest; whether a near-size may actually be substituted for the specified one is a maintenance-data question, and the answer is found in the manual, not in the sum.

Factors and Multiples

A **factor** of a number divides into it exactly with no remainder. A **multiple** is the product of that number and any positive integer.

Examples

Factors of 24: 1, 2, 3, 4, 6, 8, 12, 24

Multiples of 6: 6, 12, 18, 24, 30, 36, ...

Prime Numbers and Prime Factorisation

A **prime number** has exactly two factors: 1 and itself. The first primes are 2, 3, 5, 7, 11, 13, 17, 19, 23, 29, ... Note that 1 is *not* prime, and 2 is the only even prime.

Any integer can be written as a product of prime factors. For example: $60 = 2^2 \times 3 \times 5$.

Highest Common Factor (HCF)

The HCF is the largest number that divides exactly into two or more numbers. To find it, identify the prime factors common to both numbers and multiply them together.

Worked Example

HCF of 36 and 48:

$36 = 2^2 \times 3^2$, $48 = 2^4 \times 3$. Common factors: $2^2 \times 3 = 12$. **HCF = 12**

Lowest Common Multiple (LCM)

The LCM is the smallest number that is a multiple of two or more numbers. Take the highest power of each prime factor present in either number.

Worked Example

LCM of 36 and 48:

$36 = 2^2 \times 3^2$, $48 = 2^4 \times 3$. Highest powers: $2^4 \times 3^2 = 16 \times 9 = 144$. **LCM = 144**

Divisibility Tests

Before you can factorise anything you have to spot what divides into it, and these tests do that by inspection rather than by trial division.

Divisor	Test	Example
2	Last digit is even	3,174
3	Digit sum divides by 3	4,713 (digits sum to 15)
4	Last two digits divide by 4	7,316 (16 divides by 4)
5	Last digit is 0 or 5	2,485
6	Passes the tests for both 2 and 3	5,214 (even; digits sum to 12)
8	Last three digits divide by 8	12,536 (536 divides by 8)
9	Digit sum divides by 9	6,633 (digits sum to 18)
10	Last digit is 0	4,730
11	Alternately subtract and add the digits	8,151 ($1 - 5 + 1 - 8 = -11$)

The tests for 4 and 8 work because 100 and 1,000 are themselves divisible by 4 and 8, so everything above the last two or three digits can be ignored. The tests for 3 and 9 are the same idea used in casting out nines, which is why that check exists at all.

Finding Every Factor Without Missing One

Factors come in pairs: if 3 divides into 84 then so does $84 \div 3 = 28$, and the two are found together. Work upwards from 1, writing down both members of each pair, and stop as soon as the two members meet — that is, once you pass $\sqrt{84} \approx 9.2$. Anything beyond that point has already been written down as the partner of a smaller factor.

Worked Example — All the Factors of 84

1×84, 2×42, 3×28, 4×21, 5 does not divide, 6×14, 7×12, 8 does not divide, 9 does not divide — and the next candidate, 10, is past the square root, so stop.

Factors: 1, 2, 3, 4, 6, 7, 12, 14, 21, 28, 42, 84 — twelve of them.

Prime Factorisation in Practice

The reliable method is repeated division by the smallest prime that will go, carrying on until you reach 1. For 360: $360 \div 2 = 180$, $\div 2 = 90$, $\div 2 = 45$; 2 no longer divides, so move to 3: $45 \div 3 = 15$, $\div 3 = 5$; then $5 \div 5 = 1$. Collecting the divisors gives $360 = 2^3 \times 3^2 \times 5$. Every whole number greater than 1 has exactly one such factorisation, which is the reason the whole method is trustworthy — and the reason 1 is excluded from the primes, since allowing it would let you write any number in endlessly many ways.

Once you have the prime factorisation you can count the factors without listing them: add one to each index and multiply. For $84 = 2^2 \times 3 \times 7$ that gives $3 \times 2 \times 2 = 12$, matching the twelve found above. For 24 it gives $4 \times 2 = 8$, matching the eight listed in the example earlier in this section. For $360 = 2^3 \times 3^2 \times 5$ it gives $4 \times 3 \times 2 = 24$ factors, which is exactly why a circle of 360° can be divided evenly so many ways — into 2, 3, 4, 5, 6, 8, 9, 10, 12 and more equal parts, all landing on whole degrees. Seven is not among the factors of 360, and that is why seven equally spaced fasteners around a circle come out at an awkward 51.43° apart rather than a whole number.

The same square-root limit tests whether a number is prime at all. To check 197, you need only try the primes up to $\sqrt{197} \approx 14$: it is odd, its digits sum to 17 so 3 does not divide it, it does not end in 0 or 5, and 7, 11 and 13 all leave remainders. There is nothing left to test, so 197 is prime.

Euclid's Method for the HCF

Prime factorisation is the clearest route to the HCF, but it becomes slow with large or awkward numbers. Euclid's method avoids factorising anything: divide the larger number by the smaller, then replace the pair with the divisor and the remainder, and repeat until the remainder is zero. The last non-zero remainder is the HCF.

Worked Example

HCF of 36 and 48, by Euclid:

$48 = (1 \times 36) + 12$; $36 = (3 \times 12) + 0$. Last non-zero remainder: **12**, agreeing with the prime-factor method.

HCF of 1,071 and 462, where factorising by eye is not inviting:

$1071 = (2 \times 462) + 147$; $462 = (3 \times 147) + 21$; $147 = (7 \times 21) + 0$. **HCF = 21**.

Check: $1071 \div 21 = 51$ and $462 \div 21 = 22$, both exact, and 51 and 22 share no factor — confirming 21 was the highest.

The Link Between HCF and LCM

For any two numbers

$$\mathrm{HCF}(a, b) \times \mathrm{LCM}(a, b) = a \times b$$

Test it on the pair worked above: $12 \times 144 = 1728$, and $36 \times 48 = 1728$. The identity gives you a free cross-check on both answers at once, and it also gives you the second quantity as soon as you have the first — having found the HCF by Euclid, the LCM is $\frac{a \times b}{\mathrm{HCF}}$.

This identity holds for two numbers only. It does not extend to three. For 6, 7 and 8 the HCF is 1 and the LCM is 168, but the product of the three numbers is 336. With three or more numbers, use the prime-factor method: highest power of every prime present for the LCM, lowest power of every prime common to all of them for the HCF.

Worked Example — LCM of Several Numbers

Find the LCM of 6, 7 and 8.

$6 = 2 \times 3$, 7 is prime, $8 = 2^3$. Highest power of each prime: $2^3 \times 3 \times 7 = \mathbf{168}$.

84 is a tempting wrong answer because it is a multiple of both 6 and 7 — but $84 \div 8 = 10.5$, so it fails on the 8. Always test a candidate against *every* number in the list.

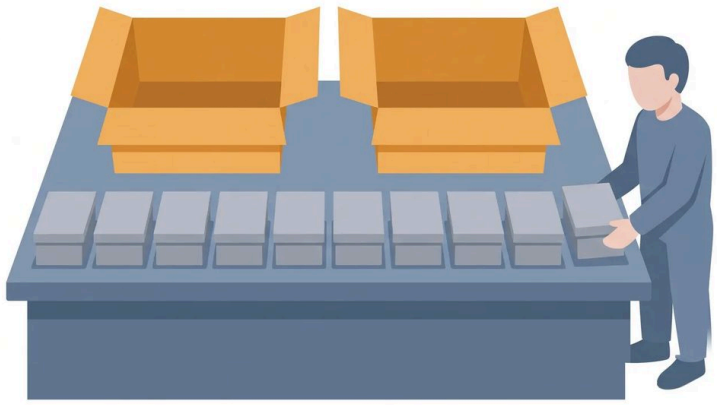
Find the lowest common denominator for $\frac{1}{6} + \frac{1}{5} + \frac{1}{17} + \frac{1}{2}$.

This is the LCM of 6, 5, 17 and 2. Here $6 = 2 \times 3$, and 5, 17 and 2 are all prime. The 2 adds nothing new, because 6 already contains one.

LCM $= 2 \times 3 \times 5 \times 17 = \mathbf{510}$. Note that 1,020 *is* a common denominator — it is simply twice as large as it needs to be, which is what "lowest" rules out.

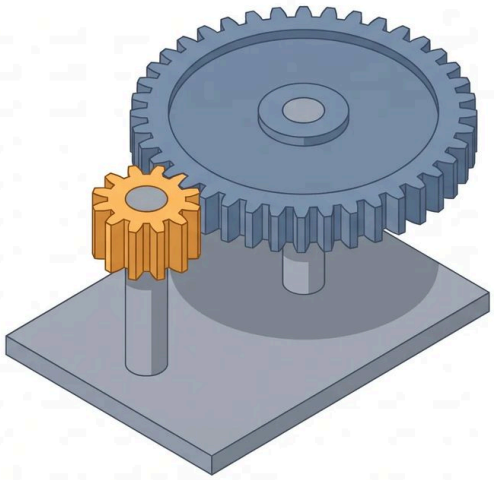
Where Factors and Multiples Are Actually Used

These are not classroom exercises. Each one solves a recurring practical problem.



84 rivets and 126 washers make 42 identical kits with none left over: the HCF gives the largest equal share.

- **The LCM tells you when repeating events coincide.** If one task falls due every 100 flight hours and another every 150, the two fall together every $\mathrm{LCM}(100, 150) = 300$ hours — useful for grouping work into a single visit rather than two.
- **The HCF tells you the largest equal share.** Making up identical kits from 84 rivets and 126 washers, the HCF of 42 says you can make 42 kits, each holding 2 rivets and 3 washers, with nothing left over. That is the same 42 that reduced $\frac{84}{126}$ to $\frac{2}{3}$ in the fractions section — cancelling a fraction and dividing stock into equal shares are the same arithmetic.
- **The HCF also predicts wear patterns in a gear train.** If a 20-tooth pinion drives a 60-tooth gear, the HCF is 20, so every tooth on the pinion meets the same three gear teeth for the life of the assembly and any local wear or damage is repeatedly hammered on the same partners. Choose tooth counts with no common factor — a 23-tooth pinion against a 60-tooth gear, HCF 1 — and the pattern does not repeat until the pinion has turned $\mathrm{LCM}(23, 60) \div 23 = 60$ times, by which point every tooth has met every other and wear is spread across the whole set. This is the reasoning behind a hunting-tooth ratio.



A small pinion driving a larger gear: the tooth counts decide which teeth keep meeting as they turn.

Weights, Measures and Conversion Factors

Aviation uses a mixture of SI (metric) and imperial/US customary units. You must be able to convert between them confidently.

Length

Conversion	Factor
1 inch (in)	25.4 mm (exact)
1 foot (ft) = 12 in	0.3048 m
1 yard (yd) = 3 ft	0.9144 m
1 mile (statute)	1.609 km
1 nautical mile (NM)	1.852 km (exact)

Mass

Conversion	Factor
1 pound (lb)	0.4536 kg
1 kilogram (kg)	2.205 lb
1 US ton (short ton)	907.2 kg
1 Imperial ton (long ton)	1,016 kg

Conversion	Factor
1 metric tonne	1,000 kg

Volume

Conversion	Factor
1 US gallon	3.785 litres
1 Imperial gallon	4.546 litres
1 litre	1,000 cm³ = 0.001 m³

Pressure

Conversion	Factor
1 atmosphere (atm)	101.325 kPa = 14.696 psi
1 bar	100 kPa = 14.504 psi
1 psi	6.895 kPa
1 atm	29.92 inHg = 760 mmHg

Temperature

Temperature Conversions

$$^\circ\text{C} = \frac{5}{9}\,(^\circ\text{F} - 32) \qquad ^\circ\text{F} = \frac{9}{5}\,^\circ\text{C} + 32$$

$$\text{K} = ^\circ\text{C} + 273.15$$

Aviation context: ISA (International Standard Atmosphere) sea-level temperature is 15 °C (59 °F, 288.15 K). Temperature decreases at approximately 2 °C per 1,000 ft of altitude in the troposphere.

Worked Example

Convert 68 °F to Celsius:

$^\circ\text{C} = \frac{5}{9}(68 - 32) = \frac{5}{9} \times 36 = 20\,^\circ\text{C}$

SI Base Units and Prefixes

The SI system is built on seven base units — the metre (length), kilogram (mass), second (time), ampere (electric current), kelvin (thermodynamic temperature), mole (amount of substance) and candela (luminous intensity). Everything else is derived from them: the newton is $\mathrm{kg \cdot m/s^2}$, the pascal is $\mathrm{N/m^2}$, the volt is $\mathrm{W/A}$. Prefixes then scale any unit by a power of ten, so no new conversion factors are needed.

Prefix	Symbol	Multiplier	Typically seen in
giga	G	10^{9}	GHz (radio frequency)
mega	M	10^{6}	MΩ (insulation resistance), MPa
kilo	k	10^{3}	kg, kN, kPa
—	—	$10^{0} = 1$	m, s, A, N
centi	c	10^{-2}	cm
milli	m	10^{-3}	mm, mA, mV
micro	*μ*	10^{-6}	*μ*m (surface finish), *μ*F
nano	n	10^{-9}	nF, ns
pico	p	10^{-12}	pF

⚠ **Capitals are not decoration in unit symbols.** A capital M is mega and a lower-case m is milli, so mA and MA differ by a factor of a thousand million. The same discipline applies to the units themselves: K is kelvin, k is kilo; N is newton, n is nano. Transcribing a value from maintenance data with the wrong case changes what it means.

The Unit-Factor Method

Guessing whether to multiply or divide is the single largest source of conversion errors, and there is a method that removes the guess entirely. Every conversion factor can be written as a fraction whose value is exactly 1. Since 1 inch *is* 25.4 mm, both of these are equal to one:

Two Ways to Write the Same Fact

$$\frac{25.4\ \mathrm{mm}}{1\ \mathrm{in}} = 1 \qquad\qquad \frac{1\ \mathrm{in}}{25.4\ \mathrm{mm}} = 1$$

Multiplying by 1 never changes a quantity, so choose whichever version puts the unit you want to **get rid of** on the bottom, and cancel it.

Worked Example — Both Directions

Convert 6 inches to millimetres: $6\ \text{in} \times \frac{25.4\ \text{mm}}{1\ \text{in}} = 152.4$ mm. The inches cancel and millimetres are left, so the arrangement was right.

Convert 6 millimetres to inches: $6\ \text{mm} \times \frac{1\ \text{in}}{25.4\ \text{mm}} = \frac{6}{25.4} = 0.236$ in. This time the millimetres cancel.

Independent check on the direction: a millimetre is a *smaller* unit than an inch, so the same length must be a *bigger* number when expressed in millimetres. 6 in becoming 152.4 mm obeys that; 6 mm becoming 0.236 in obeys it too. If a conversion has produced a bigger number in the bigger unit, it is upside down.

This also answers the question the exam actually asks, which is usually about the direction rather than the arithmetic: to convert inches to millimetres you *multiply* by 25.4, and to convert inches to centimetres you multiply by 2.54. Pounds to kilograms is a multiplication by 0.4536, and kilograms to pounds a multiplication by 2.205 — each pair being the two ways up of the same fact.

Exact Definitions Behind the Rounded Factors

Several of the factors in the tables above are not measurements at all but definitions, agreed internationally and exact by construction. Knowing which is which tells you how far you may safely round.

Quantity	Exact definition	Common working value
Inch	1 in = 25.4 mm exactly	25.4 mm
Pound (avoirdupois)	1 lb = 0.45359237 kg exactly	0.4536 kg
Kilogram to pounds	Reciprocal of the above	2.20462 lb
US gallon	1 US gal = 3.785411784 L exactly	3.785 L
Imperial gallon	1 imp gal = 4.54609 L exactly	4.546 L
Nautical mile	1 NM = 1852 m exactly	1.852 km
Statute mile	1 mile = 1609.344 m exactly	1.609 km

Worked Example — Fuel in Pounds to Kilograms

A loadsheet calls for 12,500 lb of fuel. What is that in kilograms?

$12{,}500 \times 0.45359237 = 5669.90$ kg, so **5,670 kg** to the nearest kilogram.

Reverse to check: $5669.90 \times 2.20462 = 12{,}500$ lb. The round trip returns the original figure, so neither step was inverted.

Two Gallons, and Why It Matters

The US gallon and the imperial gallon are genuinely different volumes, not two names for one thing. At 4.54609 litres the imperial gallon is about 20% larger than the US gallon's 3.785411784 litres. Ordering 500 gallons of fuel therefore means 1,893 litres if the figure is American and 2,273 litres if it is British — a difference of 380 litres, which is not a rounding matter. Whenever a volume in gallons appears in maintenance or fuelling data, establish which gallon it is before converting anything.

Fuel: Volume Becomes Mass Only Through Density

Fuel is uplifted by volume, because that is what a bowser meter measures, but it is planned and flown by mass, because that is what affects the aircraft's performance and centre of gravity. Density is the bridge between the two, and it is *not* a constant to be memorised: it depends on the fuel grade and changes with temperature, and for aviation turbine fuel it typically lies somewhere in the region of 0.78 to 0.84 kilograms per litre at 15 °C. The figure you use must come from the fuelling documentation or from the density actually measured at uplift.



Fuel is uplifted by volume through the bowser meter, but its density must be read before it becomes mass.

Mass, Volume and Density

$$\text{mass} = \text{volume} \times \text{density} \qquad \text{volume} = \frac{\text{mass}}{\text{density}}$$

Worked Example — Uplift to Loadsheet

7,000 litres are uplifted and the fuelling docket states a density of 0.80 kg/litre. Find the mass in kilograms and in pounds.

Mass $= 7000 \times 0.80 = 5600$ kg.

In pounds: $5600 \times 2.20462 = 12{,}346$ lb.

Now suppose the density had actually been 0.76 kg/litre and 0.80 was assumed. The true mass would be $7000 \times 0.76 = 5320$ kg — 280 kg less than the figure on the loadsheet. A 5% error in the density becomes a 5% error in the aircraft's fuel mass, which is why the density is read rather than remembered.

Areas and Volumes: the Factor Gets Squared or Cubed

This is the most reliably examined trap in the whole of unit conversion. A linear conversion factor of k becomes k^2 for areas and k^3 for volumes, because an area is a length multiplied by a length and a volume is a length multiplied by a length multiplied by a length. Since 1 m = 100 cm:

- $1\ \text{m}^2 = 100^2 = 10{,}000\ \text{cm}^2$
- $1\ \text{m}^3 = 100^3 = 1{,}000{,}000\ \text{cm}^3$
- $1\ \text{in}^2 = 25.4^2 = 645.16\ \text{mm}^2$
- $1\ \text{in}^3 = 25.4^3 = 16{,}387.064\ \text{mm}^3$

Worked Example — A Patch and a Block

A rectangular repair patch measures 11 cm by 120 cm. What is its area in square metres?

$11 \times 120 = 1320\ \text{cm}^2$. Dividing by 10,000: $\frac{1320}{10{,}000} = \mathbf{0.132\ m^2}$.

Dividing by 100 instead gives 13.2, and dividing by nothing leaves 1,320 — both of which will be sitting in the option list waiting for you.

A cuboid block measures 4 cm by 6 cm by 12 cm. What is its volume in cubic metres?

$4 \times 6 \times 12 = 288\ \text{cm}^3$. Dividing by 1,000,000: $\mathbf{0.000288\ m^3}$.

A useful alternative is to convert the sides first: 0.04 m × 0.06 m × 0.12 m = 0.000288 m³, which reaches the same answer and never asks you to remember the cubed factor at all.

One consequence of the cubed factor is worth stating on its own, because it is used constantly and rarely explained: a litre is defined as a cubic decimetre, and since a decimetre is 10 cm, that is $10^3 = 1000$ cubic centimetres. The abbreviation "cc" that appears on engine displacement figures is simply cm³, so an 11,946 cc engine and an 11.9 litre engine are the same engine. A cubic metre is 1,000 litres by the same reasoning.

Speed and Rate of Climb

A knot is one nautical mile per hour, and because the nautical mile (1.852 km) is longer than the statute mile (1.609344 km), a speed in knots is always a *smaller* number than the same speed in miles per hour. Dividing one definition by the other gives the factor: $1.852 \div 1.609344 = 1.15078$, so 1 knot = 1.15078 mph. Converting to kilometres per hour is easier still, because 1 knot is 1.852 km/h by definition.

Worked Example

Convert 162 knots to miles per hour and to kilometres per hour.

$162 \times 1.15078 = 186.4$, so approximately **186 mph**.

$162 \times 1.852 = 300.0$, so **300 km/h**.

The mph figure is larger than the knots figure, as it must be. An answer below 162 would have been the reciprocal factor applied by mistake.

Vertical speeds convert the same way. Since 1 ft = 0.3048 m and there are 60 seconds in a minute, 1 ft/min $= \frac{0.3048}{60} = 0.00508$ m/s, so a rate of descent of 1,000 ft/min is 5.08 m/s.

Pressure: Gauge Against Absolute

A pressure figure is incomplete until you know what it is measured against. Most workshop instruments — tyre gauges, hydraulic gauges, oxygen trolley gauges — read **gauge pressure**, which takes the surrounding atmosphere as zero. **Absolute pressure** takes a vacuum as zero, so at any positive gauge reading the absolute figure is the larger of the two:

Gauge and Absolute

$$p_{\text{absolute}} = p_{\text{gauge}} + p_{\text{atmospheric}}$$

At ISA sea level, $p_{\text{atmospheric}} = 14.696$ psi = 101.325 kPa = 1013.25 hPa.

A tyre gauge reading 200 psi therefore indicates 200 psig, which is about 214.7 psia at sea level. The distinction is not academic: gas-law calculations and anything involving compression ratios must use absolute pressures, because a gas at "zero" gauge pressure is not a gas that has ceased to exist. Note also that the hectopascal and the millibar are the same size, so 1013.25 hPa and 1013.25 mbar are two names for one pressure.

Worked Example

A hydraulic system operates at 3,000 psi. Express that in bar and in kPa.

1 psi = 6.895 kPa, so $3000 \times 6.895 = 20{,}685$ kPa, which is **20.7 MPa**.

Since 1 bar = 100 kPa, that is $\frac{20{,}685}{100} = \mathbf{206.9}\ \text{bar}$.

Cross-check with the other factor in the table: 1 bar = 14.504 psi, so $3000 \div 14.504 = 206.8$ bar. The two routes agree to within the rounding of the factors, which is what you want to see.

Temperature: Scales, Intervals and Absolute Zero

Celsius and Fahrenheit differ in two ways at once — the size of the degree and the position of zero — which is why their conversion needs both a factor and an offset. Kelvin shares the Celsius degree size but places zero at absolute zero, so only an offset is needed. Kelvin is written without a degree symbol and never takes a negative value.

Setting the two Celsius-to-Fahrenheit expressions equal shows there is exactly one temperature where the scales read the same number: it is –40, so –40 °C is –40 °F. That is a genuinely useful memory anchor, and it is also a fast check on any conversion — below –40 the Fahrenheit figure is the more negative of the two, above it the Fahrenheit figure is the larger.

A temperature and a temperature difference convert differently. The 32 in the formula fixes where zero sits, so it applies only when converting an actual temperature. For a *change* in temperature there is no offset: a rise of 10 °C is a rise of 18 °F, not 50 °F. Check it: 20 °C is 68 °F and 30 °C is 86 °F, and $86 - 68 = 18$. A change of 1 °C is also a change of exactly 1 K.

Worked Example — ISA Temperature at Altitude

Using the ISA sea-level value and lapse rate given above, what is the standard temperature at 10,000 ft, in Celsius and Fahrenheit?

Drop $= 2 \times 10 = 20$ °C, so $15 - 20 = \mathbf{-5}\ ^\circ\text{C}$.

$^\circ\text{F} = \frac{9}{5}(-5) + 32 = -9 + 32 = \mathbf{23}\ ^\circ\text{F}$.

In absolute terms, $-5 + 273.15 = 268.15$ K. The two ways of stating the ISA lapse rate agree with each other: 6.5 °C per kilometre works out at $6.5 \times 0.3048 = 1.98$ °C per 1,000 ft, which is the "approximately 2" quoted above.

Force, Mass and Torque

Mass and weight are different quantities and, in SI, take different units: mass is in kilograms, and weight is the force gravity exerts on that mass, in newtons. With $g \approx 9.81\ \text{m/s}^2$, a 100 kg component has a weight of

$100 \times 9.81 = 981$ N. The imperial system uses the word "pound" for both ideas, which is why you will see **lb** for mass and **lbf** where a force is specifically meant.

Torque is a force multiplied by the perpendicular distance from the axis, so its units are a force unit times a length unit — N·m, lb·ft or lb·in. Because a foot is twelve inches, 1 lb·ft is 12 lb·in, and 1 lb·ft is 1.356 N·m.

Worked Example — Torque in Three Units

A force of 50 lb is applied 2 ft from the axis of a nut. What torque is that, and what is it in the other common units?

$\text{Torque} = 50 \times 2 = \mathbf{100}\ \text{lb} \cdot \text{ft}$. The distance was already in feet, so nothing had to be converted first.

In pound-inches: $100 \times 12 = 1{,}200\ \text{lb} \cdot \text{in}$.

In newton-metres: $100 \times 1.356 = 135.6\ \text{N} \cdot \text{m}$.

Pound-inches and pound-feet are not interchangeable. They differ by a factor of twelve, so a fastener specified at 100 lb·in and tightened to 100 lb·ft receives twelve times the intended torque. Read the unit on the torque figure, the unit on the wrench scale and the unit in the maintenance data every single time — the numbers on their own tell you nothing.

Ratio and Proportion

A **ratio** compares two quantities of the same kind. It is written as $a : b$ and can also be expressed as the fraction $\frac{a}{b}$.

Worked Example — Gear Ratio

A gear train has a driving gear with 20 teeth and a driven gear with 60 teeth. The gear ratio is:

$20 : 60 = 1 : 3$

This means the driven gear turns once for every 3 turns of the driving gear.

Direct Proportion

Two quantities are in direct proportion if when one increases, the other increases by the same factor. If y is directly proportional to x, then $y = kx$ for some constant k.

Inverse Proportion

Two quantities are in inverse proportion if when one increases, the other decreases by the same factor. If y is inversely proportional to x, then $y = \frac{k}{x}$.

> ✈ **Aviation context:** Boyle's Law states that gas pressure and volume are inversely proportional at constant temperature: $P_1V_1 = P_2V_2$. This is fundamental to understanding pneumatic and pressurisation systems.

> **Worked Example — Mixing Ratio**
> An oil-fuel mixture requires a ratio of 1 : 40. How much oil is needed for 20 litres of mixture?
> Total parts = 1 + 40 = 41. Oil = $\frac{1}{41} \times 20 = 0.488$ litres ≈ **0.49 litres**.

Reading a Ratio Correctly

Two rules govern every ratio question, and almost every wrong answer comes from breaking one of them.

First, both quantities must be in the same unit before the ratio is formed. A 50 cm link and a 2 m rod are not in the ratio 50 : 2. Converting the rod to 200 cm gives 50 : 200, which simplifies to 1 : 4. A ratio is a pure number with no units of its own, and it only becomes one once the units have been made to match.

Second, decide whether the ratio compares a part with another part, or a part with the whole. The ratio 1 : 4 describes a mixture in which one component is a quarter of the size of the other. That same mixture has five parts in total, so the first component is one *fifth* — 20% — of the whole mixture, while being 25% of the second component. Both statements are true and they are different numbers; which one a question wants depends entirely on whether it gives you a component quantity or a total quantity.

Simplifying and Scaling a Ratio

Multiplying or dividing every term of a ratio by the same number leaves it unchanged — the same rule that governs equivalent fractions. So $6 : 5$ is equally $12 : 10$, $18 : 15$ or $24 : 20$. To test whether a given ratio is equivalent to another, reduce both to their lowest terms and compare: $10 : 16$ reduces to $5 : 8$ and $20 : 25$ reduces to $4 : 5$, so neither is a form of $6 : 5$, while $24 : 20$ divides through by 4 to give exactly $6 : 5$.

Worked Example — Dividing a Quantity in a Given Ratio

Share 96 kg of ballast between two stations in the ratio 3 : 5.

Total parts $= 3 + 5 = 8$. One part $= \frac{96}{8} = 12$ kg.

Shares: $3 \times 12 = \mathbf{36}$ kg and $5 \times 12 = \mathbf{60}$ kg. Check: $36 + 60 = 96$, and $36 : 60$ reduces to $3 : 5$.

A three-term ratio works identically. To make up 40 kg of a compound whose constituents are in the ratio 2 : 3 : 5, there are 10 parts, so one part is 4 kg and the constituents are 8 kg, 12 kg and 20 kg.

The Two Question Shapes

Ratio questions arrive in exactly two forms, and reading which one you have been given is most of the work.

- **You are told one component.** Scale directly from that component's own share of the ratio. If $A : B = a : b$ and you know A, then $B = A \times \frac{b}{a}$. The total plays no part.
- **You are told the total.** Add the terms to get the number of parts, divide the total by that number to find one part, then scale each component — the method used in the example above.



Resin to hardener at 1000 : 45 means 60 grams of resin needs only about 2.7 grams of hardener.

Worked Example — Resin and Hardener

A composite repair scheme specifies resin to hardener in the ratio 1000 : 45. How much hardener goes with 60 grams of resin?

You have been given a *component* — the resin — so scale from the resin's term: hardener $= 60 \times \frac{45}{1000} = \mathbf{2.7}$ grams.

Check the size of the answer before accepting it: hardener is 45 parts against resin's 1000, so it must come out at roughly one twentieth of 60 grams, i.e. around 3 grams. An answer of 47 g or 145 g would be nonsense on its face.

Had the question instead given 60 grams as the *total mixed weight*, the working would start from $1000 + 45 = 1045$ parts, giving $60 \times \frac{45}{1045} = 2.58$ grams. Same ratio, different question, different answer.

Telling Direct from Inverse

There is a simple test that distinguishes the two kinds of proportion from a table of values, and it is worth applying rather than judging by feel.

Relationship	What stays constant	Form	Effect of doubling x
Direct proportion	The quotient $y \div x$	$y = kx$	y doubles
Inverse proportion	The product $x \times y$	$y = \frac{k}{x}$	y halves

In direct proportion the graph is a straight line through the origin, and the practical route to an answer is the **unitary method**: find the value for one, then multiply. If seven identical brackets have a total mass of 2.8 kg, one bracket is $2.8 \div 7 = 0.4$ kg, so twelve of them are $12 \times 0.4 = 4.8$ kg. Note that direct proportion requires the line to pass through the origin: doubling the number of brackets doubles the mass because zero brackets have zero mass. Quantities that merely increase together are not necessarily in proportion.

Worked Example — Inverse Proportion

A task is estimated at 8 hours for 3 technicians. How long should it take 4 technicians?

The constant here is the total effort: $3 \times 8 = 24$ technician-hours.

With four technicians, $\frac{24}{4} = \mathbf{6}$ hours. More people, less time — the product stays at 24.

Treat the arithmetic as an ideal rather than a promise. It assumes the work divides perfectly, which real maintenance tasks often do not: access may allow only one person at the panel, and some steps must be completed before others can begin.

Gear and Pulley Ratios

Meshing gears turn at speeds inversely proportional to their tooth counts, because every tooth that leaves one gear must have entered the other:

Gears and Belt Drives

$$N_1 T_1 = N_2 T_2 \qquad\qquad N_1 D_1 = N_2 D_2$$

N = rotational speed, T = number of teeth, D = pulley diameter. Gears use tooth counts; flat and vee-belt drives use pitch diameters.

Take the 20-tooth driver and 60-tooth driven gear described above. The driven gear turns at $\frac{20}{60} = \frac{1}{3}$ of the driver's speed — a **reduction** drive. Because power is torque multiplied by rotational speed, and an ideal gear train neither creates nor destroys power, the torque available at the slower shaft is correspondingly *higher*: three times the input torque, less whatever friction takes. Speed down, torque up. A drive arranged the other way round, with a large gear driving a small one, does the reverse: the output shaft turns faster and its torque is lower.

Worked Example — Belt Drive

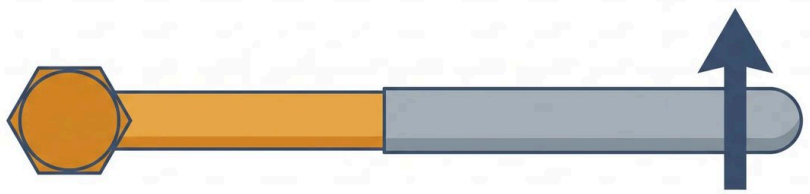
A 100 mm drive pulley turning at 3,000 rpm is belted to a 250 mm driven pulley. Find the driven speed.

$N_2 = N_1 \times \frac{D_1}{D_2} = 3000 \times \frac{100}{250} = \mathbf{1{,}200}$ rpm.

Direction check: the driven pulley is the larger of the two, so it must turn more slowly than the driver — and 1,200 rpm is indeed less than 3,000 rpm. Getting 7,500 rpm would mean the ratio had been used upside down.

Proportion at the Torque Wrench

Adding an adaptor that lengthens a torque wrench along its own axis changes the lever arm at the fastener but not the arm the wrench itself is measuring. The torque delivered to the fastener is therefore *greater* than the wrench indicates, in direct proportion to the lengths involved, so the wrench must be set to a *lower* figure than the specification if the fastener is to receive the specified torque.



An in-line extension lengthens the lever arm at the bolt, but the wrench still measures only its own length.

Torque Wrench with an In-Line Extension

$$T_{\text{indicated}} = T_{\text{required}} \times \frac{L}{L + E}$$

L = effective length of the wrench, E = added length of the extension, both measured along the same line.

Worked Example

A bolt is specified at 50 lb·in. An 8-inch torque wrench is used with a 2-inch in-line extension. What should the wrench read?

$T_{\text{indicated}} = 50 \times \frac{8}{8+2} = 50 \times \frac{8}{10} = \mathbf{40}\ \text{lb} \cdot \text{in}.$

Sense-check the direction from first principles. With a force F at the handle, the fastener sees $F \times 10$ while the wrench measures $F \times 8$, so the fastener gets 1.25 times what is displayed. Reading 40 therefore delivers $40 \times 1.25 = 50$ lb·in, as required. Setting the wrench to 50 with the extension fitted would apply $50 \times 1.25 = 62.5$ lb·in — 25% over.

The correction applies to an adaptor that extends the wrench along its own axis. An adaptor set at right angles to the wrench axis does not lengthen the measured arm, so no correction is needed. Always confirm the arrangement in the maintenance data before applying any formula.

Scale Drawings

A drawing scale is a ratio of drawing size to real size, so 1 : 50 means every millimetre on the paper stands for 50 mm on the aircraft. A line measuring 24 mm on such a drawing represents $24 \times 50 = 1200$ mm, or 1.2 m. A detail drawn at 2 : 1 is the other way round — drawn twice life size, so a 24 mm feature on paper is a 12 mm feature in reality.

Areas on a scale drawing do not scale by the same factor. They scale by the *square* of the linear ratio, for exactly the reason set out for area conversion factors in the units section above: at 1 : 50, one square centimetre of drawing represents $50^2 = 2500$ square centimetres, which is 0.25 m².

⚠ **Measuring off a drawing is not a source of dimensions.** Scaling arithmetic works only where the drawing is genuinely to scale and the dimension is not marked otherwise. Reproduced, photocopied or rescaled prints are frequently no longer at their stated scale, and a drawing annotated "not to scale" carries no scaling information at all. Take dimensions from the figures, not from a rule laid on the paper.

Averages and Percentages

Averages

- **Mean** (arithmetic average): sum of all values divided by the number of values.

$$\text{Mean} = \frac{\sum x}{n}$$

- **Median:** the middle value when data is arranged in order. For an even count of values, take the mean of the two middle values.
- **Mode:** the most frequently occurring value.

Worked Example
Five temperature readings (°C): 18, 22, 19, 22, 24
Mean = $\frac{18+22+19+22+24}{5} = \frac{105}{5} = 21\,^{\circ}\text{C}$
Ordered: 18, 19, 22, 22, 24 → Median = 22 °C
Mode = 22 °C (appears twice)

Percentages

A percentage is a fraction expressed out of 100. The symbol is %.

Percentage Formulae

Finding a percentage of a value:

$$\text{Result} = \frac{\text{percentage}}{100} \times \text{value}$$

Expressing one quantity as a percentage of another:

$$\text{Percentage} = \frac{\text{part}}{\text{whole}} \times 100\%$$

Percentage change:

$$\text{Percentage change} = \frac{\text{new value} - \text{old value}}{\text{old value}} \times 100\%$$

Worked Example — Efficiency

A generator produces 920 W of electrical power from 1,150 W of mechanical input. What is its efficiency?

$\text{Efficiency} = \frac{920}{1150} \times 100\% = 80\%$

Choosing Between the Three Averages

The three averages are not interchangeable, and each answers a different question about the same data. The mean uses every value and is the one wanted whenever a total matters — total fuel, total mass, total time. The median reports the middle of the distribution and ignores how extreme the extremes are. The mode reports what happens most often, and is the only one of the three that works on data which are not numbers at all, such as the most frequently recurring defect on a fleet.

The practical difference shows up as soon as one reading is wrong. Take the five temperature readings above and suppose a faulty probe had added a sixth reading of 90 °C. The mean would leap from 21 °C to $\frac{105+90}{6} = 32.5$ °C — higher than every genuine reading in the set. The median would be recomputed as the mean of the two middle values of 18, 19, 22, 22, 24, 90 — still 22 °C. One rogue value drags the mean a long way and leaves the median almost untouched, which is exactly why a median is preferred when a data set may contain a spurious reading.

The Weighted Mean

Averaging a set of averages by adding them up and dividing is wrong unless each of them was based on the same amount of data. The correct method multiplies each value by its weight — the number of items, hours or units it represents — before dividing by the total weight.

Weighted Mean

$$\bar{x}_w = \frac{\sum(w \times x)}{\sum w}$$

Worked Example — Averaging Two Fuel Flows

An aircraft cruises for 3 hours at an average 2,100 litres per hour, then flies 1 hour at 2,900 litres per hour. What is the average consumption over the whole flight?

The tempting answer is $\frac{2100+2900}{2} = 2500$ L/h. It is wrong, because the two rates did not apply for equal times.

Weight each rate by its hours: $\frac{(3\times2100)+(1\times2900)}{3+1} = \frac{6300+2900}{4} = \frac{9200}{4} = \mathbf{2{,}300}$ litres per hour.

Sense-check: three quarters of the flight was at the lower rate, so the true average must sit closer to 2,100 than to 2,900 — and 2,300 does.

Alongside the average it is worth quoting the **range**, the difference between the largest and smallest values, because two sets can share a mean and be nothing alike. The five temperature readings run from 18 °C to 24 °C, a range of 6 °C; a set of readings all at 21 °C would have the same mean and a range of zero. When you are looking for a drifting sensor or an unstable system, the spread is often the more informative number.

Average Speed and Average Rate

An average rate is always the total quantity divided by the total of whatever it is per — never the average of the individual rates. Speed is the case that catches people out most reliably.

Worked Example — Two Ways to Get Speed Wrong

A vehicle covers 24 miles in 45 minutes. What is its average speed?

Speed must come out in miles per *hour*, so the time has to be in hours: 45 minutes $= \frac{45}{60} = 0.75$ hour. $\frac{24}{0.75} = \mathbf{32}$ mph. Dividing 24 by 45 instead gives 0.533, and multiplying that by 60 recovers the same 32 — but dividing by 45 and stopping gives an answer of a different order entirely.

Now a return trip: 60 miles out at 60 mph, and the same 60 miles back at 30 mph. What is the average speed for the round trip?

Not 45 mph. The outbound leg takes 1 hour and the return leg takes 2 hours, so the trip is 120 miles in 3 hours: $\frac{120}{3} = \mathbf{40}$ mph.

The slower leg occupies more of the time, so it carries more weight in the average — the same principle as the weighted mean above. Averaging the two speeds directly would only be correct if equal *times* had been spent at each, not equal distances.

Moving Between Fractions, Decimals and Percentages

A percentage is just a fraction with 100 fixed as the denominator, so all three notations describe the same quantity and converting between them needs no new arithmetic:

- **Fraction to percentage:** multiply by 100. $\frac{9}{20} \times 100 = \frac{900}{20} = 45\%$, and $\frac{5}{8} \times 100 = \frac{500}{8} = 62.5\%$.
- **Decimal to percentage:** multiply by 100, which moves the point two places right. $0.45 \rightarrow 45\%$.
- **Percentage to decimal:** divide by 100, moving the point two places left. $7.5\% \rightarrow 0.075$.
- **Percentage to fraction:** write it over 100 and cancel. $45\% = \frac{45}{100} = \frac{9}{20}$.

Converting a percentage to a decimal before calculating with it is almost always the quickest route, because "12% of" then becomes a single multiplication by 0.12 rather than a multiplication and a division.

Increase, Decrease, and Getting Back to the Original

An increase of $p\%$ multiplies a quantity by $\left(1 + \frac{p}{100}\right)$ and a decrease multiplies it by $\left(1 - \frac{p}{100}\right)$. Handling it as one multiplication rather than as "work out the change, then add it on" removes a step and the error that goes with it.

Worked Example — Percentage Change and Its Reverse

A component measured 4.80 mm when new and now measures 4.32 mm. What percentage has it lost? $\frac{4.32 - 4.80}{4.80} \times 100\% = \frac{-0.48}{4.80} \times 100\% = \mathbf{-10\%}$. The change is always divided by the *original* value, not by the new one.

A measured dimension is reported as 3% over nominal, at 51.5 mm. What is the nominal dimension? The measurement is 1.03 times nominal, so nominal $= \frac{51.5}{1.03} = \mathbf{50.0}$ mm.

Taking 3% off the 51.5 instead gives $51.5 \times 0.97 = 49.96$ mm, which is close but wrong — because 3% of the larger number is not the same quantity as 3% of the smaller one. To undo a percentage increase you divide; you do not subtract the same percentage.

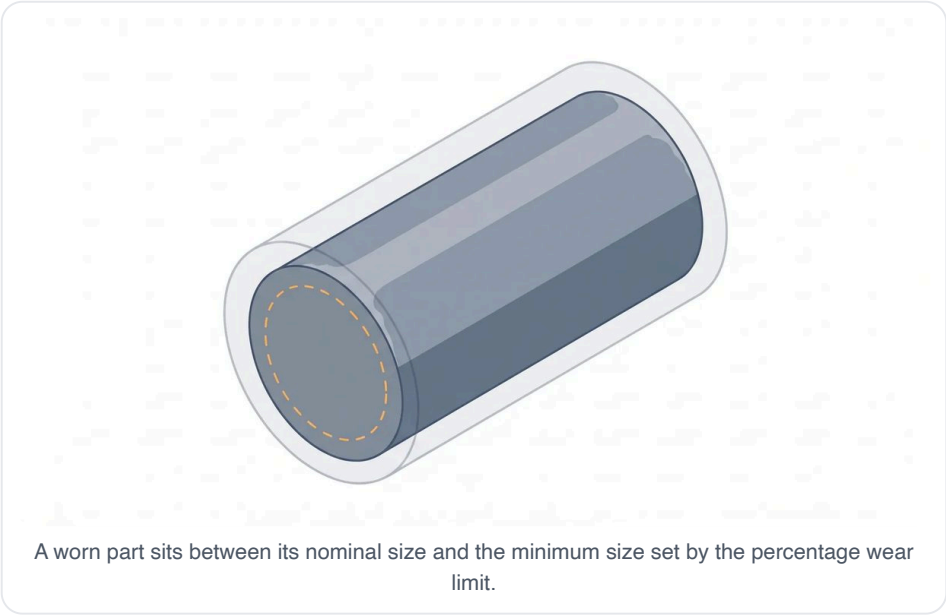
⚠ **Successive percentages do not add.** A quantity increased by 10% and then decreased by 10% does not return to where it started: $100 \rightarrow 110 \rightarrow 99$. The reason is visible in the algebra, since $(1 + p)(1 - p) = 1 - p^2$, so there is always a shortfall of p^2 — here $0.10^2 = 1\%$. For the same reason, two stages of a system with efficiencies of 90% and 80% give an overall efficiency of $0.90 \times 0.80 = 72\%$, not 85%. Percentages applied in sequence multiply.

A related distinction is worth getting right in words as well as in figures. If a pass rate rises from 60% to 75%, it has risen by 15 **percentage points**, but the relative increase is $\frac{15}{60} \times 100\% = 25\%$. Both descriptions are

correct and they are different numbers; saying "a 15% rise" when you mean 15 percentage points misstates the change by a wide margin.

Percentages as Tolerances and Wear Limits

A great deal of maintenance data expresses acceptability as a percentage rather than as an absolute figure, because the same rule then applies across a range of sizes. Turning that percentage into the two numbers you can actually measure against is straightforward, but it must be done from the correct base.



A worn part sits between its nominal size and the minimum size set by the percentage wear limit.

Worked Example — A Percentage Tolerance

A resistor is marked 470 Ω with a tolerance of ±5%. What range of measured values is acceptable?

$470 \times 0.05 = 23.5\ \Omega$, so the band is 470 ± 23.5, that is **446.5 Ω to 493.5 Ω**.

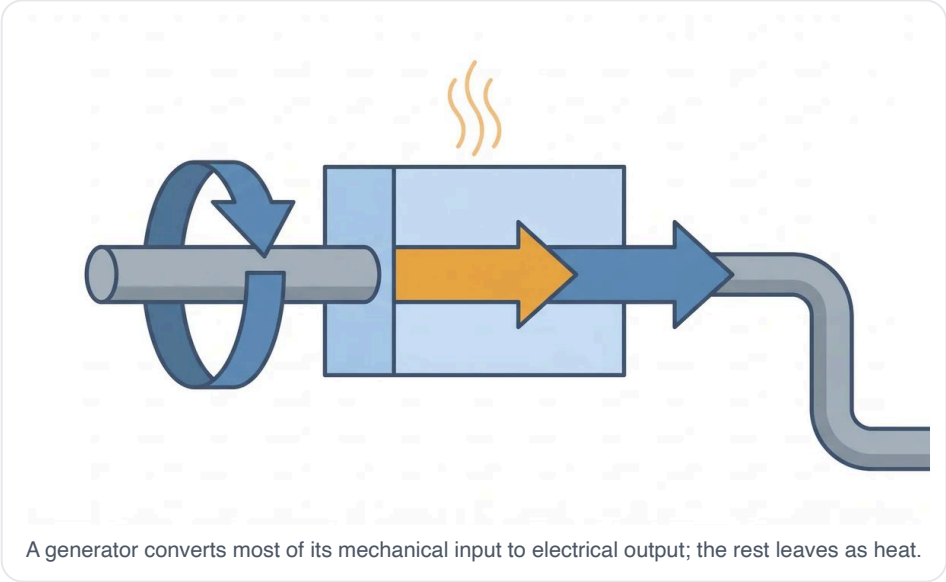
Suppose maintenance data permits a nominally 12.00 mm part to wear by up to 20%. A measurement of 10.20 mm is found. Is it serviceable?

The limit is $12.00 \times 0.20 = 2.40$ mm of wear, so the minimum acceptable dimension is $12.00 - 2.40 = 9.60$ mm.

Actual wear is $12.00 - 10.20 = 1.80$ mm, which is $\frac{1.80}{12.00} \times 100\% = 15\%$. That is inside the 20% allowance, so the part is within limits — but it has used three quarters of its allowance, which is the more useful thing to report.

The figures above are illustrative arithmetic. Actual tolerances and wear limits are type-specific and come from the approved maintenance data, never from a worked example.

Efficiency cannot exceed 100%. The generator example above turns 1,150 W of mechanical input into 920 W of electrical output, so 230 W — the other 20% — leaves as heat, friction and windage rather than disappearing. If an efficiency calculation ever comes out above 100%, the arithmetic is not the problem: an input and an output have been swapped, or the two figures were not in the same units.



A generator converts most of its mechanical input to electrical output; the rest leaves as heat.

Areas and Volumes

Area Formulae (2D Shapes)

Shape	Formula
Rectangle	$A = l \times w$
Square	$A = s^2$
Triangle	$A = \frac{1}{2} \times b \times h$
Parallelogram	$A = b \times h$
Trapezoid (Trapezium)	$A = \frac{1}{2}(a + b) \times h$
Circle	$A = \pi r^2$

Volume Formulae (3D Shapes)

Shape	Volume	Surface Area
Cube	$V = s^3$	$SA = 6s^2$
Rectangular prism	$V = l \times w \times h$	$SA = 2(lw + lh + wh)$
Cylinder	$V = \pi r^2 h$	$SA = 2\pi r(r + h)$
Cone	$V = \frac{1}{3}\pi r^2 h$	$SA = \pi r(r + l)$, l = slant
Sphere	$V = \frac{4}{3}\pi r^3$	$SA = 4\pi r^2$

Worked Example — Cylinder Volume

A piston engine cylinder has a bore (diameter) of 130 mm and a stroke of 150 mm. Find the swept volume of one cylinder.

Radius $r = 65$ mm = 0.065 m. Height $h = 150$ mm = 0.150 m.

$V = \pi r^2 h = \pi \times 0.065^2 \times 0.150 = \pi \times 0.004225 \times 0.150 \approx 0.001991$ m³ = **1,991 cm³ ≈ 1.99 litres**

Perimeter and Circumference

Area answers "how much surface", perimeter answers "how far round". They are different quantities with different units — square millimetres against millimetres — and confusing them is the most heavily exploited error in this part of the syllabus.

Around the Edge

$$\text{Rectangle: } P = 2(l + w) \qquad \text{Circle: } C = 2\pi r = \pi d$$

The circle formulae are where the marks are lost. $2\pi r$ is the distance round the rim; πr^2 is the surface enclosed. And because the diameter is twice the radius, $d^2 = (2r)^2 = 4r^2$ — so putting the diameter into the area formula in place of the radius produces an answer exactly **four times too large**. If a question gives you a diameter, halve it before it goes anywhere near πr^2.

Working Backwards from the Circumference

A circle is fully described by any one of its radius, diameter, circumference or area, so any of them can be recovered from any other. Given the circumference, rearrange $C = 2\pi r$ to $r = \frac{C}{2\pi}$, and substituting that into

$A = \pi r^2$ gives a direct route:

Area from Circumference

$$A = \pi \left(\frac{C}{2\pi} \right)^2 = \frac{C^2}{4\pi}$$

Worked Example — Two Routes, One Answer

A circular inspection panel has a circumference of 250 mm. Find its area.

Route 1 — find the radius first: $r = \frac{250}{2\pi} = 39.79$ mm. Then $A = \pi \times 39.79^2 = \pi \times 1583.1 = 4974$ mm².

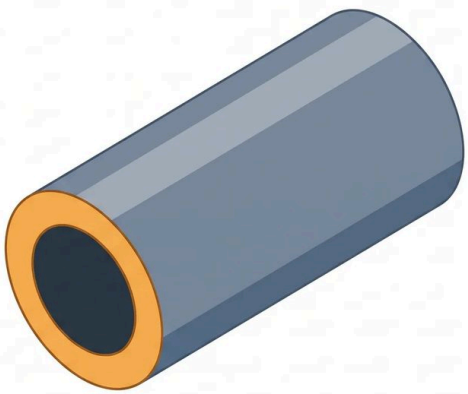
Route 2 — direct: $A = \frac{250^2}{4\pi} = \frac{62{,}500}{12.566} = 4974$ mm².

Both give **4,974 mm²**, which is 49.74 cm² — remembering that a square centimetre is 100 square millimetres, not 10.

Rings, Sectors and Shapes with Holes in Them

Real components are rarely one clean shape. Almost all of them can be handled by adding and subtracting simple ones.

An **annulus** — the cross-section of a tube, a washer, a bearing race — is the difference of two circles, $A = \pi(R^2 - r^2)$. For an outer radius of 30 mm and an inner radius of 22 mm, that is $\pi(900 - 484) = \pi \times 416 = 1307$ mm². The difference of two squares gives an equivalent form, $\pi(R - r)(R + r) = \pi \times 8 \times 52$, which reaches the same 416π and is easier to do in your head.



The cross-section of a tube is an annulus: the outer circle's area minus the area of the hole.

A **sector** is the fraction of a circle cut off by an angle, so both its area and its arc length are simply that fraction of the whole:

Sector of a Circle

$$A_{\text{sector}} = \frac{\theta}{360} \times \pi r^2 \qquad \text{arc} = \frac{\theta}{360} \times 2\pi r$$

Worked Example — Equally Spaced Fasteners

A circular patch is secured by seven equally spaced rivets. What is the angular spacing, and what area of the patch does each rivet "own"?

A full circle is 360°, so the spacing is $\frac{360}{7} = \mathbf{51.43}^\circ$ between adjacent rivets. Seven is not a factor of 360, which is why this does not come out as a whole number.

If the patch has a radius of 100 mm, each of the seven sectors covers $\frac{51.43}{360} \times \pi \times 100^2 = \frac{31{,}416}{7} = 4488$ mm².

Check: seven sectors of 4,488 mm² total 31,416 mm², which is $\pi \times 100^2$ — the whole circle, as it must be.

A **composite shape** is handled by subtraction. A rectangular doubler 200 mm by 150 mm with a 60 mm diameter lightening hole has an area of $(200 \times 150) - (\pi \times 30^2) = 30{,}000 - 2827 = 27{,}173$ mm². Note the halving of the 60 mm diameter to a 30 mm radius before squaring. Using 60 as the radius would subtract $\pi \times 60^2 = 11{,}310$ mm² instead of 2,827 mm², leaving a doubler of 18,690 mm² — an answer over 30% adrift that looks entirely plausible on the page, which is what makes this particular slip worth guarding against.

Triangles: Which Height?

In $A = \frac{1}{2} \times b \times h$, the height must be measured *perpendicular* to the base you have chosen, not along a sloping side. In a right-angled triangle the two legs are already perpendicular to each other, so either can serve as the base and the other is then the height — which is why the area of a right-angled triangle is simply half the product of its two shorter sides.

When all three sides are known but no perpendicular height is, Heron's formula gives the area without needing one. With s as half the perimeter:

Heron's Formula

$$s = \frac{a+b+c}{2} \qquad A = \sqrt{s(s-a)(s-b)(s-c)}$$

Worked Example

Find the area of a triangle with sides 13, 14 and 15 cm.

$s = \frac{13+14+15}{2} = 21$.

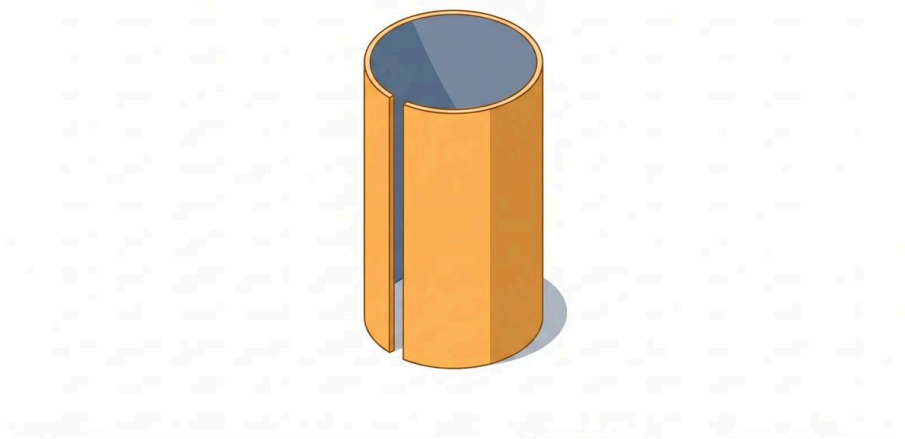
$A = \sqrt{21 \times 8 \times 7 \times 6} = \sqrt{7056} = \mathbf{84}$ cm².

Curved Surface Against Total Surface

A surface-area question is only answerable once you know which surfaces are wanted, and the cone is where this bites. Its **curved** (lateral) surface is $\pi r l$, where l is the *slant* height measured up the sloping side — not the vertical height. Adding the circular base, πr^2, gives the **total** surface $\pi r(r + l)$. If a question asks only for the curved area, including the base makes the answer too large; if it asks for the total and you omit the base, too small.

The slant height is rarely given directly, but it follows from the vertical height and the base radius, since they form a right-angled triangle: $l^2 = r^2 + h^2$. A cone of base radius 30 mm and vertical height 40 mm therefore has $l = \sqrt{900 + 1600} = \sqrt{2500} = 50$ mm, so its curved surface is $\pi \times 30 \times 50 = 4712$ mm², its base is $\pi \times 30^2 = 2827$ mm², and its total surface is 7,539 mm² — which $\pi r(r + l) = \pi \times 30 \times 80$ confirms directly.

The cylinder decomposes the same way, and its formula stops being something to memorise once you see it. Unroll the curved surface and it is a rectangle whose width is the circumference and whose height is the cylinder's height, giving $2\pi r h$; add the two circular ends, $2 \times \pi r^2$, and you have $2\pi r(r + h)$.



The curved wall of a cylinder is a rectangle rolled up: its width goes once round the circumference.

What Happens When You Scale a Shape

Double every linear dimension of a cube of side 1: the surface area goes from 6 to 24 and the volume from 1 to 8. That is the same k^2 and k^3 rule set out for conversion factors in the units section, now applied to a shape rather than to a unit. Its physical consequence is the part worth carrying: volume grows faster than surface area, so a large tank holds far more per unit of skin than a small one, which is why large volumes lose heat comparatively slowly and why doubling a duct's dimensions has a much bigger effect on the air it carries than on the material it is made from.

Capacity, Displacement and Compression

Volume calculations become capacity calculations as soon as the units change: 1 litre is 1,000 cm³, and 1 m³ is 1,000 litres.

Worked Example — Tank Capacity and Contents

A rectangular tank measures 1.2 m by 0.8 m by 0.5 m. What is its capacity, and what mass of fuel does it hold at a stated density of 0.80 kg per litre?

$V = 1.2 \times 0.8 \times 0.5 = 0.48\ \text{m}^3$, which is $0.48 \times 1000 = \mathbf{480}$ litres.

Mass $= 480 \times 0.80 = \mathbf{384}$ kg.

Worked Example — From One Cylinder to the Whole Engine

The cylinder worked above has a swept volume of 1,991 cm³. If the engine has six such cylinders, its total displacement is $6 \times 1991 = 11{,}946\ \text{cm}^3$, or about **11.9 litres**.

Compression ratio compares the total volume above the piston at the bottom of its stroke with the volume left at the top:

$\text{CR} = \frac{\text{swept volume} + \text{clearance volume}}{\text{clearance volume}}$

Suppose the clearance volume of this cylinder is 250 cm³. Then $\text{CR} = \frac{1991+250}{250} = \frac{2241}{250} = 8.96$, quoted as about **9 : 1**. The 250 cm³ is an assumed figure for the arithmetic; a real clearance volume is a type-specific number taken from the engine data.

Prisms, Tapers and Awkward Tank Shapes

Real tanks and ducts are rarely boxes or perfect cylinders, but most of them are **prisms** — solids of constant cross-section — and every prism obeys one rule regardless of how complicated that cross-section is:

Any Prism, Any Cross-Section

$$V = A_{\text{cross-section}} \times \text{length}$$

The cuboid and the cylinder are just the two commonest cases of it: $l \times w \times h$ is a rectangular cross-section times a length, and $\pi r^2 h$ is a circular one. So a tank whose section is a trapezium, a composite shape, or a rectangle with a rounded corner is handled by finding that one area and multiplying — there is no separate formula to learn.

Worked Example — A Tapered Tank Bay

A tank bay has a trapezoidal cross-section: the two parallel faces measure 400 mm and 600 mm, separated by a perpendicular depth of 300 mm. The bay is 1.2 m long. What does it hold?

Cross-sectional area $= \frac{1}{2}(400 + 600) \times 300 = \frac{1}{2} \times 1000 \times 300 = 150{,}000\ \text{mm}^2$, which is $0.15\ \text{m}^2$.

Volume $= 0.15 \times 1.2 = 0.18\ \text{m}^3 = \mathbf{180}$ litres.

Note the perpendicular distance between the parallel faces is what the formula wants, not the length of a sloping side — the same requirement as the height of a triangle.

Solids that taper to a point rather than keeping their section are the other family, and they follow a single rule too: a pyramid or a cone occupies exactly **one third** of the prism or cylinder that shares its base and height. That is why the cone formula in the table above is $\frac{1}{3}\pi r^2 h$ — the cylinder of the same base radius and height holds three times as much. A cone of base radius 30 mm and vertical height 40 mm therefore holds $\frac{1}{3} \times \pi \times 900 \times 40 = 37{,}699\ \text{mm}^3$, against the 113,097 mm³ of the cylinder that encloses it.

Areas in Sheet-Metal and Pipework

Two everyday uses of area close this section, and both turn a geometric answer into something you can act on.

Mass of a sheet. Area multiplied by thickness gives volume, and volume multiplied by density gives mass. A piece of 2 mm aluminium-alloy sheet 1.5 m by 1.0 m has a volume of $1.5 \times 1.0 \times 0.002 = 0.003\ \text{m}^3$, which is 3,000 cm³. Airframe aluminium alloys have a density of roughly 2.7 grams per cubic centimetre, so the sheet weighs about $3000 \times 2.7 = 8100$ g, or 8.1 kg. The same three-step chain — area, then volume, then mass — estimates the weight of any repair patch or doubler before it is cut.

Bore and flow velocity. The cross-sectional area of a pipe is $A = \frac{\pi d^2}{4}$, which is just πr^2 written in terms of the bore. Because the diameter is squared, doubling the bore *quadruples* the area: a 10 mm bore gives $\pi \times 5^2 = 78.5\ \text{mm}^2$, while a 20 mm bore gives $\pi \times 10^2 = 314\ \text{mm}^2$. For a given volumetric flow, velocity is flow divided by area, so that same doubling of the bore reduces the flow velocity to a quarter of what it was. A small change in a pipe or restrictor size has a much larger effect than its dimensions suggest.

Squares, Cubes, Square and Cube Roots

Squaring a number means multiplying it by itself: $n^2 = n \times n$. The **square root** reverses this: $\sqrt{n^2} = n$.

Cubing means raising to the power of 3: $n^3 = n \times n \times n$. The **cube root** reverses it: $\sqrt[3]{n^3} = n$.

n	n^2	n^3	$\sqrt{n}$	$\sqrt[3]{n}$
1	1	1	1.000	1.000
2	4	8	1.414	1.260
3	9	27	1.732	1.442
4	16	64	2.000	1.587
5	25	125	2.236	1.710
10	100	1,000	3.162	2.154

Remember: The square root of a negative number is not a real number. However, squaring a negative number gives a positive result: $(-5)^2 = 25$.

Perfect Squares and Perfect Cubes Worth Recognising

Recognising a perfect square or cube on sight turns a root question into a recall question. The table above covers the smallest cases; these are the others you will meet most often.

n	**6**	**7**	**8**	**9**	**11**	**12**	**15**	**20**
n^2	36	49	64	81	121	144	225	400
n^3	216	343	512	729	1,331	1,728	3,375	8,000

Note that 64 appears as both a square and a cube, being 8^2 and 4^3, and that 729 is 9^3 as well as 27^2. A number can be a perfect square and a perfect cube at the same time, which is worth knowing before you assume a question has only one sensible reading.

Two Square Roots, but One Radical Sign

Every positive number has two square roots, one positive and one negative, because both 5×5 and $(-5) \times (-5)$ give 25. The radical sign is defined to mean the **principal** root — the non-negative one — so $\sqrt{25} = 5$

and not ± 5. But solving the equation $x^2 = 25$ is a different question, and its answer is $x = \pm 5$, because both values satisfy it.

The distinction matters as soon as you take a square root while rearranging a formula. Recovering a radius from an area, or a velocity from a kinetic energy, produces two mathematical answers of which only the positive one is physically meaningful — there is no such thing as a radius of –30 mm. Discarding the negative root is a decision you make about the physics, not something the arithmetic does for you.

Roots of Negative Numbers

Squaring a negative gives a positive, so no real number squared can produce a negative — hence there is no real square root of a negative number. **Cube roots behave differently, and this is not an exception to the rule but a consequence of it.** Cubing a negative keeps it negative, since $(-2) \times (-2) \times (-2) = -8$, so every negative number has exactly one real cube root and it is negative: $\sqrt[3]{-8} = -2$ and $\sqrt[3]{-125} = -5$.

The general pattern is that an *even* root of a negative number has no real value, while an *odd* root always does. If you are ever unsure, cube the candidate and see whether it reproduces the original.

Estimating a Root Without a Calculator

Bracket first. To find $\sqrt{50}$, note that $7^2 = 49$ and $8^2 = 64$, so the answer lies between 7 and 8 and must be very close to 7, since 50 is barely above 49. That alone is enough to reject a wrong option in a multiple-choice question.

To sharpen it, use the **divide-and-average** method: divide the number by your current guess, then average the guess with the result. Each pass roughly doubles the number of correct digits.

Worked Example — $\sqrt{50}$ by Hand

Start with the guess 7.

$50 \div 7 = 7.142857$; average with 7: $\frac{7+7.142857}{2} = 7.071429$.

$50 \div 7.071429 = 7.070707$; average again: $\frac{7.071429+7.070707}{2} = 7.071068$.

Two passes have produced **7.071068**, which is correct to every digit shown. Why it works: if your guess is too small, dividing by it overshoots, so the true root always lies between the guess and the quotient — and the midpoint is a far better guess than either.

The same method handles cube roots by bracketing and refining. For $\sqrt[3]{100}$: $4^3 = 64$ and $5^3 = 125$, so the answer is between 4 and 5. Trying 4.6 gives 97.3 and 4.7 gives 103.8, so it is between those; 4.64 gives 99.9, so $\sqrt[3]{100} \approx 4.64$.

Splitting a Root into Factors

For non-negative numbers, a root distributes across a product and a quotient:

Roots of Products and Quotients

$$\sqrt[n]{a \times b} = \sqrt[n]{a} \times \sqrt[n]{b} \qquad \sqrt[n]{\frac{a}{b}} = \frac{\sqrt[n]{a}}{\sqrt[n]{b}}$$

Pulling out a perfect-square factor simplifies awkward roots. Since $50 = 25 \times 2$, $\sqrt{50} = \sqrt{25} \times \sqrt{2} = 5\sqrt{2}$; using $\sqrt{2} = 1.414$ from the table above, that gives $5 \times 1.414 = 7.07$, agreeing with the hand calculation. The same idea makes $\sqrt{0.04}$ easy: it is $\sqrt{\frac{4}{100}} = \frac{2}{10} = 0.2$.

⚠ **A root does not distribute across a sum.** $\sqrt{9+16} = \sqrt{25} = 5$, whereas $\sqrt{9} + \sqrt{16} = 3 + 4 = 7$. The two are not equal and never will be. Splitting only ever works across multiplication and division, never across a sum or a difference.

Squaring Numbers That End in 5

Multiply the leading digits by the next whole number up, then write 25 on the end. So 35^2: $3 \times 4 = 12$, giving **1225**. 65^2: $6 \times 7 = 42$, giving **4225**. 85^2: $8 \times 9 = 72$, giving **7225**. It follows from the distributive law rather than being a coincidence, and it is quick enough to be worth using as a cross-check on a calculator entry.

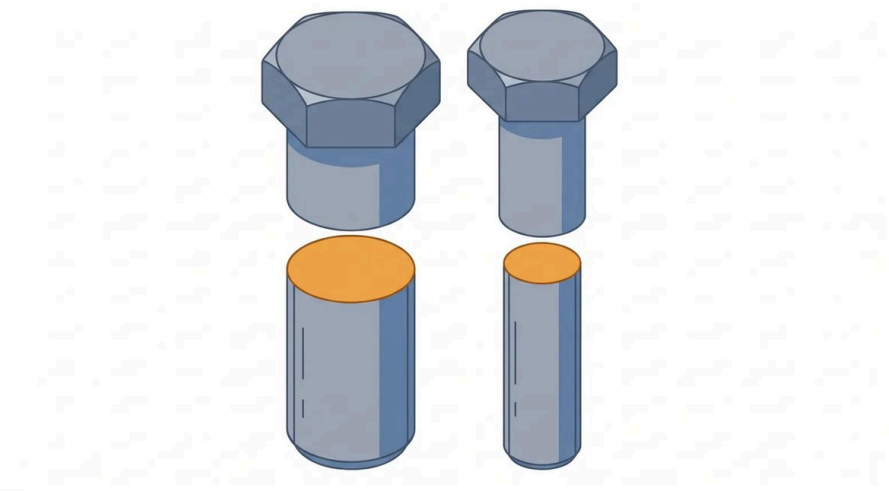
Units Are Squared and Cubed Along with the Numbers

Squaring a quantity squares its unit as well as its number, which is why an area comes out in mm² and a volume in mm³. That single fact is the origin of the area and volume conversion factors listed in the units section. Taking a root goes the other way and halves or thirds the unit: the square root of an area in mm² is a length in mm.

This gives you a free check on any rearranged formula. If you have taken the square root of something and the units that come out are not a length, the rearrangement is wrong.

Why Squares and Cubes Dominate Engineering Numbers

Squares and cubes are not decoration in engineering formulae; they describe how quantities really behave, and each of the following is a direction you should be able to state without hesitating.



Halving a bolt's diameter leaves a quarter of the cross-section to carry the same load.

- **Area against diameter.** Because $A = \frac{\pi d^2}{4}$, halving a bolt's diameter leaves a quarter of the cross-section. For the same applied load the stress, $\frac{\text{force}}{\text{area}}$, is therefore four times higher — which is why an undersized fastener is a far more serious substitution than its dimensions suggest.
- **Kinetic energy against speed.** Since $E_k = \frac{1}{2}mv^2$, doubling the speed does not double the energy that has to be absorbed, it quadruples it. A touchdown at 1.4 times the intended speed arrives with roughly twice the kinetic energy to be dissipated.
- **Volume against bore.** The same squared diameter met under bore and flow velocity above also sits inside swept volume, $\pi r^2 h$: for a given stroke, doubling the bore quadruples a cylinder's displacement.
- **Intensity against distance.** Energy radiating outwards spreads over a spherical surface of area $4\pi r^2$, so at twice the distance the same energy covers four times the area and the intensity falls to a quarter. This inverse-square behaviour applies to radiated signal strength and to radiation dose alike.

Squares and cubes are the special cases $n = 2$ and $n = 3$ of a general system of indices, and roots are the same system with fractional indices — $\sqrt{a}$ is $a^{1/2}$ and $\sqrt[3]{a}$ is $a^{1/3}$. The rules that govern all of them together are the subject of the final section.

Powers and Roots (Laws of Indices)

An **index** (plural: indices), also called an exponent or power, indicates how many times a base is multiplied by itself: $a^n = \underbrace{a \times a \times \cdots \times a}_{n \text{ times}}$

Laws of Indices

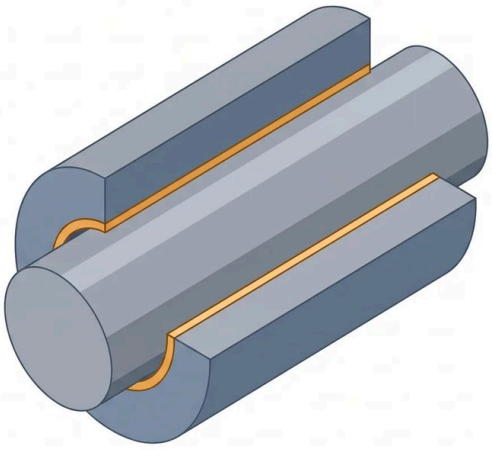
Rule	Law	Example
Multiplication	$a^m \times a^n = a^{m+n}$	$2^3 \times 2^4 = 2^7 = 128$
Division	$a^m \div a^n = a^{m-n}$	$5^6 \div 5^2 = 5^4 = 625$
Power of a power	$(a^m)^n = a^{mn}$	$(3^2)^3 = 3^6 = 729$
Zero index	$a^0 = 1$	$7^0 = 1$
Negative index	$a^{-n} = \frac{1}{a^n}$	$2^{-3} = \frac{1}{8}$
Fractional index	$a^{1/n} = \sqrt[n]{a}$	$27^{1/3} = 3$
General fractional	$a^{m/n} = \sqrt[n]{a^m}$	$8^{2/3} = (\sqrt[3]{8})^2 = 4$

Worked Example

Simplify $\frac{3^5 \times 3^{-2}}{3^2}$:

Numerator: $3^{5+(-2)} = 3^3$. Then $\frac{3^3}{3^2} = 3^{3-2} = 3^1 = \mathbf{3}$

Aviation context: Scientific notation uses powers of 10 to express very large or small numbers. The speed of light is 3×10^8 m/s. A micrometre is 1×10^{-6} m. This is essential when working with electrical values, wavelengths, and material tolerances.



Material tolerances are tiny lengths, the kind that powers of ten such as ten to the minus six express neatly.

Where the Index Laws Come From

The three laws in the table are not conventions to be taken on trust. Writing the powers out shows why they must hold: $a^3 \times a^4$ is three as multiplied by four as, which is seven as, so the indices add. Division cancels as from top and bottom, so the indices subtract. A power of a power repeats the multiplication, so the indices multiply.

The remaining four entries look arbitrary until you notice that each is forced by the first three — they are the only definitions that keep the laws working.

- **Zero index.** $\frac{a^m}{a^m}$ is obviously 1. The division law says it is $a^{m-m} = a^0$. Both must be true, so $a^0 = 1$ for every a except zero.
- **Negative index.** $\frac{a^0}{a^n}$ is $\frac{1}{a^n}$. The division law says it is $a^{0-n} = a^{-n}$. Hence $a^{-n} = \frac{1}{a^n}$ — a negative index means a reciprocal, never a negative number.
- **Fractional index.** By the power-of-a-power law, $\left(a^{1/n}\right)^n = a^{n \times \frac{1}{n}} = a$. Something which gives a when raised to the power n is by definition the nth root, so $a^{1/n} = \sqrt[n]{a}$.
- **General fractional index.** Combining the last two, $a^{m/n}$ is the nth root raised to the power m, in either order.

Root first, then power. $8^{2/3}$ can be worked as $\sqrt[3]{8^2} = \sqrt[3]{64} = 4$ or as $\left(\sqrt[3]{8}\right)^2 = 2^2 = 4$. The answers are identical, but taking the root first keeps the numbers small enough to handle without a calculator — the difference between cube-rooting 64 and cube-rooting 8. On something like $16^{3/4}$ the saving is larger still: $\left(\sqrt[4]{16}\right)^3 = 2^3 = 8$, where the other order would ask you for the fourth root of 4,096.

What the Laws Do Not Say

Every index law above concerns a *product or quotient of powers of the same base*. Reading them more widely than that is the standard way to lose a mark here.

- **Different bases do not combine by adding indices.** $2^3 \times 5^2$ cannot be simplified to any single power; it is just $8 \times 25 = 200$.
- **But equal indices do combine across different bases.** $a^m \times b^m = (ab)^m$, which is why $2^3 \times 5^3 = 10^3 = 1000$. The matching pair is $\frac{a^m}{b^m} = \left(\frac{a}{b}\right)^m$, so $\frac{6^2}{3^2} = 2^2 = 4$.
- **Sums of powers do not combine at all.** $2^3 + 2^4 = 8 + 16 = 24$, which is not 2^7. The addition law is for multiplication, not addition.
- **A power does not distribute across a bracketed sum.** $(3 + 4)^2 = 49$, not $3^2 + 4^2 = 25$.

⚠ **The index binds tighter than a leading minus sign.** -3^2 means "the negative of 3^2", which is –9. $(-3)^2$ means "negative three, squared", which is +9. The brackets are the whole of the difference, and the same applies on a calculator — entering a negative value without bracketing it is one of the most common ways a keyed calculation goes wrong.

Worked Examples — Simplifying with Indices

$\dfrac{2^3 \times 2^5}{2^4} = 2^{3+5-4} = 2^4 = \mathbf{16}$. Check the long way: $\frac{8\times32}{16} = \frac{256}{16} = 16$.

$\left(5^2\right)^3 \div 5^4 = 5^6 \div 5^4 = 5^2 = \mathbf{25}$.

$16^{3/4} = \left(\sqrt[4]{16}\right)^3 = 2^3 = \mathbf{8}$.

$32^{-2/5}$: the minus sign says take the reciprocal, so it is $\dfrac{1}{32^{2/5}} = \dfrac{1}{\left(\sqrt[5]{32}\right)^2} = \dfrac{1}{2^2} = \mathbf{\dfrac{1}{4}}$.

Standard Form

A number is in **standard form** (also called scientific notation) when it is written as $a \times 10^n$ with exactly one non-zero digit before the decimal point — that is, with $1 \leq a < 10$ — and n a whole number. To convert, count how many places the point has to move to leave one digit in front of it: moving it *left* gives a positive exponent, moving it *right* gives a negative one.

- $4{,}730{,}000 = 4.73 \times 10^6$ — the point moved six places left.
- $0.0000413 = 4.13 \times 10^{-5}$ — the point moved five places right.

The sign of the exponent is therefore a reliable size check on its own: a positive exponent means a number of 10 or more, a negative exponent means a number below 1, and 10^0 covers everything between. If you have written a very small number with a positive exponent, the point went the wrong way.

Standard form is a format, not a value. The same quantity has endlessly many correct power-of-ten expressions, and only one of them is standard form. 0.0000413 is equally 4.13×10^{-5}, 41.3×10^{-6} and 413×10^{-7} — every one of those multiplies out to the same number, and you can confirm it by moving the point back. Only the first is *in standard form*, because only there is the leading value between 1 and 10. So a question asking what a number "can be written as" and a question asking for it "in standard form" are not the same question, and the wrong options are usually the right digits attached to the wrong power of ten.

Calculating in Standard Form

Standard form turns awkward arithmetic into two easy pieces: handle the leading numbers, and handle the powers of ten with the index laws.

Operation	Method	Example
Multiply	Multiply the leading numbers, add the exponents	$(3 \times 10^4)(2 \times 10^{-7}) = 6 \times 10^{-3}$
Divide	Divide the leading numbers, subtract the exponents	$\frac{4.8 \times 10^6}{1.6 \times 10^2} = 3 \times 10^4$
Raise to a power	Raise the leading number, multiply the exponent	$(2 \times 10^3)^4 = 16 \times 10^{12} = 1.6 \times 10^{13}$
Add or subtract	Make the exponents equal first	$3.2 \times 10^3 + 0.45 \times 10^3 = 3.65 \times 10^3$

After multiplying or dividing, the leading number may fall outside the 1-to-10 range and the answer then has to be renormalised. $(6 \times 10^4)(5 \times 10^3) = 30 \times 10^7$, and since $30 = 3 \times 10^1$, that becomes 3×10^8. The addition row is the one candidates skip: $3.2 \times 10^3 + 4.5 \times 10^2$ cannot be added as it stands, because the two terms are not counting the same thing. Rewriting 4.5×10^2 as 0.45×10^3 makes them comparable, and only then do the leading numbers add.

Worked Example — Roots of Numbers in Standard Form

Square-rooting halves the exponent, so it is straightforward when the exponent is even: $\sqrt{4 \times 10^6} = 2 \times 10^3$.

When the exponent is odd, shift the point to make it even first. For $\sqrt{9 \times 10^{-5}}$, rewrite 9×10^{-5} as 90×10^{-6}:

$\sqrt{90 \times 10^{-6}} = \sqrt{90} \times 10^{-3} = 9.487 \times 10^{-3}$.

Check it against the plain decimal: $9 \times 10^{-5} = 0.00009$, whose square root is 0.009487 — the same number.

Engineering Notation and the SI Prefixes

Engineering notation is standard form with one extra restriction: the exponent must be a multiple of three, so the leading number sits anywhere from 1 up to 1,000. The point of the restriction is that every allowed exponent corresponds to an SI prefix, so a value written this way can be read straight off in the units an engineer actually uses.

- 4.7×10^{-5} farads is 47×10^{-6} F, read as 47 μF.
- 2.2×10^6 ohms is already at a multiple of three, and reads as 2.2 MΩ.
- 1.5×10^9 hertz reads as 1.5 GHz.
- 8.2×10^{-11} farads becomes 82×10^{-12} F, or 82 pF.

This is why component markings, test-set ranges and maintenance data step in thousands rather than in arbitrary powers, and it is worth being fluent in both directions — a resistance quoted as 2,200,000 Ω and one quoted as 2.2 MΩ are the same reading, and an insulation-test limit is far easier to check against a meter when both are expressed the same way.



A meter reading is easiest to check against maintenance data when both are expressed the same way.

Module 2 — Physics

2.1 — Matter

This section covers the fundamental nature of matter — what everything in the physical world is made of. Understanding atomic structure, chemical bonding, and the states of matter is essential for aircraft maintenance engineers, since the materials used in aviation (metals, composites, fluids, gases) all behave according to these principles.

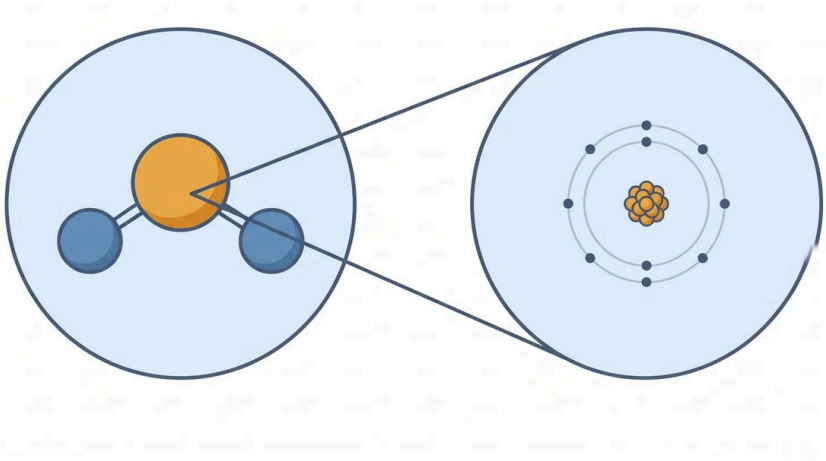
Nature of Matter: Elements, Atoms, and Molecules

What Counts as Matter

Matter is anything that has mass and occupies space. Both halves of that test have to be met, and it is the definition examiners lean on: a forged undercarriage leg, a litre of hydraulic fluid and the nitrogen charge in a tyre are all matter, because each has mass and each takes up volume. Heat, light, sound and electrical energy are *not* matter. They have no mass and occupy no space, although they routinely change the temperature, motion or state of matter that does — which is exactly what the rest of this module is about.

Two quantities are easy to confuse at this point. **Mass** is the quantity of matter in a body. It is the same on the ramp, at 40,000 ft and in orbit, and it is what resists a change of motion. **Weight** is the force gravity exerts on that mass, so it varies with the strength of the gravitational field. An engineer reading a spring balance is measuring weight and inferring mass from it; the mechanics section of this module develops the relationship properly. **Volume** is the space the matter occupies, and mass divided by volume gives **density** — the property that decides whether a material earns its place in an airframe, and the reason two metals of comparable strength are not equally attractive to a designer.

Matter is organised in a hierarchy, and a surprising number of exam questions are really asking which level of that hierarchy a term belongs to:



One molecule of bonded atoms, and one of those atoms enlarged to show its nucleus and electrons.

- **Sub-atomic particles** — protons, neutrons and electrons. They exist and they determine everything above them, but on their own they are not a chemical substance you could weigh out or react.
- **Atom** — the smallest particle of an element that can exist on its own and still be that element. This is the answer the exam wants when it asks for the smallest particle of matter that can exist in nature. Physicists can name smaller objects — quarks inside the proton, for instance — but a quark is never found in isolation, so it is not something that can exist by itself as matter.
- **Molecule** — two or more atoms chemically bonded together. A molecule built from identical atoms is still an element; one built from different atoms is a compound.
- **Element, compound and mixture** — the bulk substances an engineer actually handles, stores and signs for.

One rule underpins the whole of this section: in any process an engineer meets, matter is neither created nor destroyed, only rearranged. Melt a kilogram of ice and you have a kilogram of water. Burn a kilogram of kerosene and the exhaust products weigh *more* than the fuel did, because oxygen drawn from the air has combined with it — nothing has vanished, and the extra mass came in through the intake. Whenever a question seems to hinge on something appearing or disappearing, the answer is almost always an audit of where the particles went.

Chemical Elements

An **element** is a pure substance that cannot be broken down into simpler substances by chemical means. There are 118 known elements, each identified by an **atomic number** (the number of protons in its nucleus). Elements are organised in the **Periodic Table**.

Elements important in aviation include:

Element	Symbol	Atomic No.	Aviation Use
Aluminium	Al	13	Primary airframe material (lightweight, corrosion-resistant)
Titanium	Ti	22	Engine components, high-strength fasteners
Iron	Fe	26	Steel alloys for landing gear, engine parts
Chromium	Cr	24	Plating, stainless steel alloys
Copper	Cu	29	Electrical wiring and connectors
Nitrogen	N	7	Tyre inflation, fuel tank inerting
Oxygen	O	8	Crew and passenger breathing systems
Hydrogen	H	1	Lightest element; component of fuels and water

Reading the Periodic Table

The Periodic Table is not simply a list of the elements — it is arranged so that an element's *position* predicts its behaviour, and that is what makes it worth learning rather than memorising. The elements are laid out in order of **increasing atomic number**, one proton at a time, reading left to right along each row and then down to the next. (Early versions ordered them by atomic weight, which put a handful of pairs in the wrong sequence; ordering by proton count fixed that and is the arrangement in use today.)

- A **period** is a horizontal row. Each new period begins a new outer electron shell, so an element's period number tells you how many shells are in use.
- A **group** is a vertical column. Every member of a group has the same number of electrons in its outer shell, which is why they behave chemically alike: sodium and potassium sit in the same group and both form chlorides of the form NaCl and KCl.

The horizontal position separates the two great families. **Metals** occupy the left and centre. They hold few outer electrons, part with them readily to form positive ions, and in bulk are lustrous, malleable, ductile and good conductors of both heat and electricity. **Non-metals** occupy the upper right. Their outer shells are nearly full, so they gain or share electrons instead of losing them; as solids they tend to be dull and brittle, several are gases at room temperature, and as a family they are poor conductors of heat and electricity. Between the two runs a diagonal staircase of **metalloids** — boron, silicon, germanium, arsenic — whose conductivity sits between the extremes and which are the raw material of the semiconductor industry.

The extreme right-hand column is the special case: the **noble gases** (helium, neon, argon, krypton, xenon and radon), whose outer shells are already full. They are non-metals, so the general statement that the elements at the far right of the table are non-metals holds all the way to the last column.

Region of the table	Outer-shell electrons	What the atom does with them	Typical members	Where the engineer meets them
Groups 1–2 (reactive metals)	1 or 2	Loses them, forming a positive ion	Lithium, sodium, magnesium, calcium	Lithium cells; magnesium alloy castings
Transition metals (centre block)	Variable	Loses a variable number; alloys readily	Titanium, chromium, iron, nickel, copper	Engine and airframe alloys, plating, wiring
Group 13	3	Loses three, forming Al^{3+}	Aluminium	The principal airframe metal
Group 14	4	Shares four rather than losing them	Carbon, silicon	Composite fibres, semiconductors, carbides
Groups 15–17 (non-metals)	5 to 7	Gains or shares to complete the octet	Nitrogen, oxygen, sulphur, chlorine	Inerting gas, breathing oxygen, corrosive salts
Group 18 (noble gases)	8 (2 for helium)	Nothing — the shell is already full	Helium, neon, argon	Argon shielding for welding; helium leak testing

Because every member of a group carries the same outer-electron count, a group also shares a combining power, which is what lets you predict a compound's formula from the table alone. That connection is worked through under valency below.

Structure of the Atom

Every atom consists of three sub-atomic particles:

Particle	Location	Charge	Relative Mass
Proton	Nucleus	Positive (+1)	1
Neutron	Nucleus	Neutral (0)	1
Electron	Shells (orbits)	Negative (–1)	≈ 1/1836

Key definitions:

- **Atomic number (Z)** — the number of protons in the nucleus. This defines the element.
- **Mass number (A)** — the total number of protons + neutrons in the nucleus.
- **Isotopes** — atoms of the same element with different numbers of neutrons (same Z, different A).
- **Ion** — an atom that has gained or lost electrons, giving it a net electrical charge.

Atomic Notation

$${}^{A}_{Z}X \qquad \text{e.g. } {}^{27}_{13}\text{Al} \text{ — aluminium with 13 protons and 14 neutrons}$$

The Scale of an Atom and Its Particles

Relative mass and charge are what an exam question normally asks for, but the measured values behind them explain several answers that otherwise look arbitrary:

Particle	Rest mass (kg)	Mass in atomic mass units (u)	Electric charge (C)
Proton	1.673×10^{-27}	1.0073	$+1.602 \times 10^{-19}$
Neutron	1.675×10^{-27}	1.0087	0
Electron	9.109×10^{-31}	0.000549	-1.602×10^{-19}

Three consequences follow directly from those numbers.

A proton and a neutron weigh essentially the same. The neutron is heavier by about one part in a thousand, which is why both are given a relative mass of 1 and why the mass of a proton is correctly described as equal to the mass of a neutron. That approximation is what makes the mass number a simple count of nucleons rather than a sum that has to be looked up.

The electron is negligible by comparison. Dividing the proton mass by the electron mass gives a ratio of about 1,836, which is where the fraction 1/1836 comes from; rounded to one significant figure that ratio becomes "about 2,000 times", so an answer option built on 2,000 and a textbook figure of 1,836 are describing the same ratio. The practical consequence is that virtually all of an atom's mass sits in its nucleus. Take an aluminium-27 atom: its 13 electrons together weigh about 1.2×10^{-29} kg, against a nucleus of roughly $4.5 \times$

10^{-26} kg, so the electrons contribute less than 0.03 % of the atom's mass. Strip them all off and the atom's weight is essentially unchanged — but its chemistry is destroyed.

The proton and electron charges are exactly equal and opposite. An atom is therefore electrically neutral only when it holds exactly as many electrons as protons, which is why the number of electrons in a stable, un-ionised atom always equals the number of protons in its nucleus. An atom with two protons carries two electrons; one with 13 protons carries 13. Break that equality by adding or removing electrons and you no longer have a neutral atom but an ion.

The dimensions are as lopsided as the masses. A whole atom is of the order of 10^{-10} m across — a tenth of a nanometre — while its nucleus measures only about 10^{-15} to 10^{-14} m. The nucleus is therefore some ten thousand to a hundred thousand times smaller in diameter than the atom that contains it, which means it occupies less than a million-millionth of the atom's volume. An atom is overwhelmingly empty space, with more than 99.9 % of its mass concentrated in a speck at the centre and the electrons sweeping out everything else.

That raises an obvious question: if protons all carry the same positive charge, and like charges repel, why does a nucleus not fly apart? The answer is the **strong nuclear force**, an attraction between nucleons that is far stronger than the electrostatic repulsion but which acts only over distances of about 10^{-15} m — roughly the width of the nucleus itself. Neutrons contribute that attraction without adding any repulsion, which is why they act as the nuclear "glue" and why larger nuclei need proportionally more neutrons than protons to stay together. Where the balance is wrong, the nucleus is unstable.

Counting Protons, Neutrons and Electrons

Almost every atomic-structure question you will meet reduces to the same three-step count, so it is worth having it as a reflex:

- The **atomic number Z** is the number of protons, and nothing else. It is what identifies the element — change it and you have a different element altogether.
- The **mass number A** is the number of protons *plus* neutrons, i.e. the total number of nucleons. Electrons are never included; they weigh too little to matter.
- The **neutron count** is therefore $A - Z$, and in a neutral atom the **electron count** equals Z.

Worked Example

Aluminium-27 has $Z = 13$ and $A = 27$. Neutrons $= 27 - 13 = 14$. Electrons in the neutral atom $= 13$. Those are the figures quoted in the notation box above.

Now work backwards from a picture. A shell diagram shows three particles marked p+ and two marked n inside the nucleus. Protons $= 3$, so $Z = 3$. Nucleons $= 3 + 2 = 5$, so $A = 5$. A neutral atom of it would carry three electrons — two on the first ring and one on the second.

Two traps recur. The first is offering "protons plus electrons" for the mass number. In a neutral atom the electron count equals the proton count, so what that option actually adds up is $Z + Z$, twice the atomic number, when the mass number A is protons plus neutrons. The two agree only for an atom that happens to hold exactly as many neutrons as protons, and neither worked example above is such an atom: aluminium-27 would come out at $13 + 13 = 26$ rather than 27, and the shell-diagram nucleus at $3 + 3 = 6$ rather than 5. Electrons are not nucleons and never enter a mass number at all, so the option is wrong whether the atom is neutral or ionised, and whichever isotope it is. The second is confusing mass number with relative atomic mass. The mass number is always a whole number, because it is a count of particles; relative atomic mass usually is not, for the reason set out under isotopes below.

The atomic-structure diagram above is drawn on exactly this principle. Its nucleus is labelled only "p+ n" rather than with a count, but the electrons can be counted: two occupy the innermost ring and five the next, and the third ring is drawn empty. Read as a real atom, that arrangement is nitrogen — seven electrons in a neutral atom imply seven protons, so $Z = 7$. The empty outer ring is there to make the point that an energy level exists whether or not anything occupies it. The rings themselves are schematic: neither their radii nor the spacing between them is drawn to scale, and the electrons are spaced evenly around each ring for legibility rather than being pinned to any real position.

Models of the Atom

Every picture of an atom is a model, chosen because it predicts something useful, and it is worth knowing which model is being used and what it is good for.

- The earliest useful picture treated the atom as a solid indivisible sphere. It explains why elements combine in fixed ratios, and nothing else.
- The discovery of the electron produced an atom of positive material with electrons embedded in it. That model was killed by experiment: firing alpha particles at thin gold foil sent a few of them straight back, which is impossible unless the positive charge and nearly all the mass are concentrated in a very small central body. That experiment is where the nucleus comes from.
- The **shell model** that followed puts electrons in fixed circular orbits at discrete energy levels, and this is the model used throughout this section, throughout the syllabus and in the diagram above. Its virtue is that it

gets the chemistry right: the electron count per shell, the octet rule, valency, ionisation and the emission of light at definite wavelengths all fall straight out of it.

- The modern picture replaces the orbit with an **orbital** — a region in which an electron of a given energy is likely to be found, with no defined path at all. The s orbital is spherical, the three p orbitals are dumbbell-shaped and mutually perpendicular, and the electron's position within any of them is a matter of probability rather than position.

The two later models do not contradict each other on anything an engineer needs. Energy levels, capacities and the drive towards a full outer shell are identical in both; only the picture of where the electron is at a given instant differs. That is why the shell model remains the working tool, and why a shell diagram should be read as a bookkeeping device for electron counts rather than as a photograph of an orbit.

Isotopes and Relative Atomic Mass

Because chemical behaviour is decided by the electrons, and the electron count follows the proton count, adding or removing a neutron changes an atom's mass without changing its chemistry at all. Atoms of one element that differ only in neutron count are **isotopes** of that element: same atomic number, different mass number, different number of nucleons in the nucleus. Carbon-14 and carbon-12 have the same number of protons and the same number of electrons; only the neutrons differ, and both are unambiguously carbon.

Hydrogen is the textbook case because it has so few particles that the effect is obvious:

Isotope	Protons	Neutrons	Mass number	Abundance and notes
Protium (ordinary hydrogen)	1	0	1	Over 99.9 % of natural hydrogen; the only common nuclide with no neutron at all
Deuterium	1	1	2	About 0.02 %; combined with oxygen it gives heavy water, $\mathrm{D_2O}$
Tritium	1	2	3	Trace amounts only; radioactive, and forms tritiated water rather than heavy water

Hydrogen is also the lightest and structurally simplest element there is — a single proton with a single electron — and the most abundant in the universe, which is why it is the element questions reach for when they ask which material is the smallest or simplest.

Carbon behaves the same way: carbon-12 accounts for about 98.9 % of natural carbon and carbon-13 for most of the rest, while carbon-14 exists only in traces, is radioactive, and decays at a known rate — the basis

of radiocarbon dating. The **half-life** of a radioactive isotope is simply the time taken for half of a given quantity of it to decay.

Isotopes are also why relative atomic masses are rarely whole numbers. The value quoted for an element is the *weighted average* over its natural isotopes. Chlorine is quoted as 35.5 not because any chlorine atom weighs 35.5 u, but because about 75 % of chlorine atoms are chlorine-35 and about 25 % are chlorine-37:

Weighted Average Atomic Mass

$$(0.75 \times 35) + (0.25 \times 37) = 26.25 + 9.25 = 35.5$$

The unit those masses are quoted in is fixed by international agreement against one particular nuclide: the **atomic mass unit** is defined as exactly one twelfth of the mass of a carbon-12 atom, which works out at about 1.661×10^{-27} kg. Every atomic weight on the Periodic Table is therefore referred to carbon. Hydrogen was the original reference and oxygen was used for a time afterwards, but carbon-12 is the standard in force today, which is the point behind questions asking which element atomic weights are related to.

Two practical consequences. Radioactive isotopes are used deliberately in maintenance: iridium and cobalt gamma sources are standard for radiographic inspection, and they work precisely because an unstable nucleus is indifferent to the chemistry going on around it. And an isotope is not an ion — the distinction is worth nailing down now, because the wrong answers on isotope questions are often built out of electron counts, which is the ion's defining property and not the isotope's. An isotope differs in *neutrons* and is electrically neutral; an ion differs in *electrons* and carries a net charge.

Unstable Nuclei and Ionising Radiation

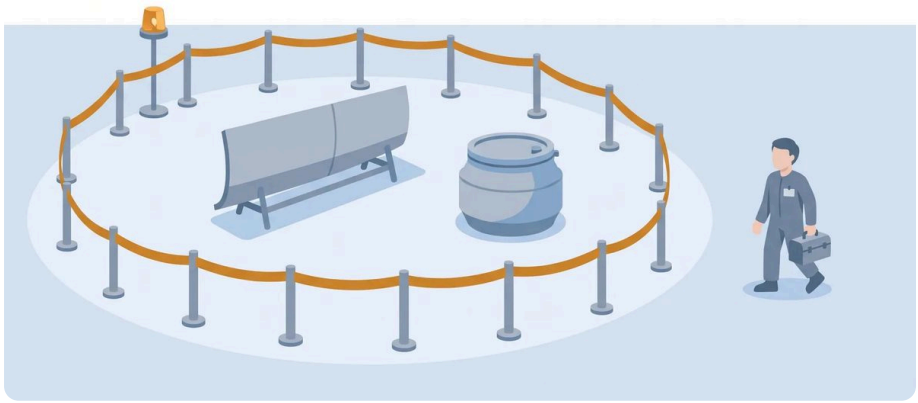
A nucleus is stable only when the strong nuclear force and the electrostatic repulsion between its protons are in balance, and that balance depends on the neutron-to-proton ratio. Where the ratio is wrong — too few neutrons, too many, or simply too many nucleons in total — the nucleus sheds particles or energy until it reaches a stable arrangement. That is **radioactive decay**, and it is a property of the nucleus alone: heating, cooling, dissolving or chemically combining a radioactive isotope changes its decay rate not at all.

Emission	What it is	Charge	Stopped by
Alpha	A helium nucleus — two protons and two neutrons	Positive	A sheet of paper or the outer layer of skin; dangerous chiefly if inhaled or ingested
Beta	A high-speed electron ejected as a neutron converts to a proton	Negative	A few millimetres of aluminium

Emission	What it is	Charge	Stopped by
Gamma	Electromagnetic radiation of very short wavelength — no particle at all	None	Only substantial thicknesses of lead or concrete; never absorbed completely

All three are described as **ionising** radiation because each carries enough energy to knock electrons out of the atoms it passes through, leaving a trail of ions behind. That is what makes it useful and what makes it hazardous, and the two are the same mechanism: the ionisation that exposes a radiographic film is the ionisation that damages living tissue.

The practical consequence in maintenance is gamma radiography. A sealed isotope source is used to inspect welds, castings and structure that cannot be reached with an X-ray set, and because gamma radiation cannot be switched off, control is by distance, shielding and time: a controlled area is established and cleared, warning signs and audible or visual indicators are posted, the source is kept in its shielded container until the moment of exposure, and personnel carry dose monitoring. Anyone working near an active radiography operation must respect the barriers even if their own task is unrelated — the source has no way of knowing who is in the beam.



A gamma radiography controlled area: source shielded, barrier and beacon set, people kept well outside.

Ions and Ionisation

The direction of the charge is the part that catches people out, so reason it through rather than memorising it. Electrons are negative. Remove one and the atom is left with more protons than electrons, so it becomes a **positive ion** — a positive ion is one that has *lost* electrons, not gained them. Add an electron and the atom has more negative charge than positive, so it becomes a **negative ion**. A positive ion is also called a cation and a negative ion an anion, and the charge is written as a superscript: Na^{+}, Mg^{2+}, Al^{3+}, Cl^{-}, O^{2-}.

Which way an atom goes is decided by its outer shell. Metals hold one to three outer electrons loosely and shed them to reach the full shell underneath, so metals form positive ions. Non-metals are within one to three electrons of a full shell and capture what they need, so they form negative ions. Noble gases do neither.

The energy needed to pull the most loosely held electron away from a neutral atom is its **ionisation energy**. It is lowest for the outermost electron of a large metal atom, because that electron is furthest from the nucleus and partly screened from it by the shells in between, and highest for a small atom with a full shell. In practice atoms are ionised by any of the following:

- **Friction** — charge separation as one material rubs or flows over another. Fuel moving through a hose and filter, and air flowing over an airframe in flight, both do this.
- **Heat** — a flame or an electric arc is hot enough to strip electrons from gas atoms.
- **Radiation** — ultraviolet and X-rays carry enough energy per photon to eject an electron. Solar UV is what maintains the ionosphere.
- **Dissolving** — an ionic compound such as sodium chloride splits into free ions when it dissolves, which is what makes the solution an electrolyte.

Ions matter to an engineer for several reasons at once. An aircraft accumulates static charge in flight through friction with the air and precipitation, and that charge is bled back off through static dischargers so it cannot build to a level that interferes with radio reception or produces a discharge. Refuelling is done with the aircraft, the bowser and the nozzle bonded together, because charge separated in flowing fuel is otherwise free to jump as a spark. Batteries pass current through their electrolyte as moving ions rather than as free electrons. And corrosion is fundamentally an ionic process: it needs a metal that gives up ions and an electrolyte to carry them, which is why trapped moisture and salt are so much more damaging than dry air. A gas in which a substantial fraction of the atoms has been ionised stops behaving like an ordinary gas altogether; that state is dealt with under states of matter below.

Electron Shells

Electrons orbit the nucleus in **energy levels** (shells), labelled K, L, M, N, etc. (or 1, 2, 3, 4...). Each shell holds a maximum number of electrons:

Shell	Number	Max Electrons	Formula
K	1	2	$2n^2$
L	2	8	
M	3	18	
N	4	32	

The outermost occupied shell contains the **valence electrons**, which determine the chemical behaviour of the element. Elements with a full outer shell (noble gases) are chemically stable and unreactive.

Sub-shells and Where 2n² Comes From

Each shell is itself divided into **sub-shells**, labelled s, p, d and f, and each type of sub-shell has a fixed capacity no matter which shell it belongs to: an s sub-shell holds 2 electrons, a p holds 6, a d holds 10 and an f holds 14. The shell maxima are simply those capacities added up, which is why the $2n^2$ formula works:

Shell	n	Sub-shells present	Capacities added	Shell maximum
K	1	1s	2	2
L	2	2s, 2p	2 + 6	8
M	3	3s, 3p, 3d	2 + 6 + 10	18
N	4	4s, 4p, 4d, 4f	2 + 6 + 10 + 14	32

Two exam answers fall straight out of that table. The **innermost shell holds a maximum of two electrons**, because it contains only a 1s sub-shell — and any s sub-shell, in any shell, holds exactly two. The **second shell holds a maximum of eight**, from one s and one p sub-shell.

Electrons fill the lowest available energy first, working outward. The shell closest to the nucleus is the lowest energy level, because that is where the electrostatic attraction to the positive nucleus is strongest; each shell further out is a higher energy level and its electrons are held more weakly. Filling is not strictly shell by shell, however: the 4s sub-shell lies slightly below 3d in energy, so it fills before the M shell has taken its full 18. That is why an outer shell in the elements an engineer actually handles never carries more than eight electrons, even though the M shell's theoretical maximum is 18 — once a shell reaches eight in its outer position, the next electron starts the shell beyond it. The 18 and 32 figures are reached only when those shells are buried underneath an occupied outer shell.

The spacing of the shells is worth stating carefully, because two things move in opposite directions. In the shell model the radius grows roughly as the square of the shell number — in proportion to 1, 4, 9 and 16 for the first four shells — so the gaps between successive shells are about 3, 5 and 7 in the same units. **The further from the nucleus the shells lie, the further apart from each other they get.** Their *energies*, however, converge: for hydrogen the four lowest levels sit at about -13.6, -3.4, -1.5 and -0.85 electron-volts, so the steps between them shrink from 10.2 to 1.9 to 0.7 eV as you move outward. An outer electron is therefore both physically further from its neighbours and energetically much easier to dislodge, which is precisely why chemistry and electrical conduction are governed by the outermost shell and never by the inner ones.

Give an electron a modest amount of energy — from heat, from a collision, or from absorbing light — and it can jump to a higher shell, leaving the atom in an **excited state**. It does not stay there; when it falls back it re-emits that energy as light of one definite wavelength, set by the size of the gap it fell across. That is why a sodium lamp is yellow, why a welding arc has its characteristic colour, and why every element produces its own line spectrum. Give the electron enough energy instead and it leaves the atom altogether, which is ionisation.

Electron Configurations of Some Familiar Elements

Writing out the electrons shell by shell turns the rules above into something you can use. Fill from the innermost shell outward, respect each shell's capacity, and stop when you have used up as many electrons as the element has protons:

Element	Z	Electrons per shell (K, L, M, N)	Outer shell	Consequence
Hydrogen	1	1	1	One electron short of a full K shell; forms one bond
Helium	2	2	2 (full)	Inert — the K shell needs only two
Nitrogen	7	2, 5	5	Completes its octet by sharing three
Oxygen	8	2, 6	6	Needs two; combining power 2
Neon	10	2, 8	8 (full)	Inert
Sodium	11	2, 8, 1	1	Loses one readily; a very reactive metal
Magnesium	12	2, 8, 2	2	Loses two; combining power 2
Aluminium	13	2, 8, 3	3	Three outer electrons; the principal airframe metal
Silicon	14	2, 8, 4	4	Shares four; the semiconductor case
Chlorine	17	2, 8, 7	7	One electron short of a full octet
Argon	18	2, 8, 8	8 (full)	Inert; the arrangement every other atom is working towards
Titanium	22	2, 8, 10, 2	2	M shell still filling beneath a settled outer shell
Chromium	24	2, 8, 13, 1	1	Variable valency — commonly 3 or 6
Iron	26	2, 8, 14, 2	2	Variable valency — commonly 2 or 3
Copper	29	2, 8, 18, 1	1	One loose outer electron above a filled M shell

Three things in that table repay a second look. First, nitrogen's 2, 5 is exactly the arrangement drawn in the atomic-structure diagram earlier in this section. Second, the noble gases are the only entries whose outer shell is full, and they are the only ones that do nothing chemically — the pattern is not a coincidence but the whole explanation. Third, the transition metals break the neat progression: titanium, chromium, iron and

copper all carry only one or two electrons in their outermost shell while a deeper shell — the M shell, capacity 18 — is the one whose filling is in progress across the row. Titanium, chromium and iron are caught part-way through it, holding 10, 13 and 14; copper is the element that has just completed it, at the full 18, leaving a single electron outside. Because those inner electrons are close in energy to the outer ones, they can also take part in bonding, which is why the transition metals show more than one valency — iron as ferrous or ferric, copper as cuprous or cupric, chromium as trivalent or hexavalent — and why they form so many alloys and coloured compounds.

Valency and the Octet Rule

A full outer shell is the arrangement every atom behaves as though it is trying to reach. For all the elements an engineer meets that means eight electrons in the outer shell — the **octet rule** — with helium the exception, needing only two because its shell is the K shell. An atom's **valency** is its combining power: the number of electrons it must gain, lose or share to get there.

The route it takes depends on how far it has to go. Up to three outer electrons and it is quicker to lose them, exposing the full shell underneath; five or more and it is quicker to gain or share the few that are missing. Four sits on the boundary and shares. So the combining power is the outer-electron count for atoms with three or fewer, and eight minus that count for atoms with five or more.

Outer-shell electrons	Term used for that count	Usual combining power	How the octet is completed	Examples
1	Univalent (monovalent)	1	Loses one	Hydrogen, lithium, sodium, potassium
2	Bivalent (divalent)	2	Loses two	Magnesium, calcium, zinc
3	Trivalent	3	Loses three	Aluminium, boron
4	Tetravalent	4	Shares four	Carbon, silicon
5	Pentavalent	3 (5 in some compounds)	Gains or shares three	Nitrogen, phosphorus
6	Hexavalent	2	Gains or shares two	Oxygen, sulphur
7	Heptavalent	1	Gains or shares one	Fluorine, chlorine
8 (2 for helium)	Inert, zero valency	0	Nothing to do — already full	Helium, neon, argon

✈ **Exam tip:** two questions on this topic can look as though they contradict each other. An atom with five electrons in its outer shell is called pentavalent, while oxygen — which has six — is called divalent, or equivalently bivalent, the two words meaning exactly the same thing. Both are right, because the first names the outer-electron *count* and the second names the *combining power*. Read the stem: if it describes an atom as having a stated number of electrons in its outer shell, name that count; if it asks for the valency of a named element, give the combining power. Oxygen has six outer electrons and a combining power of two, which is why its valency is 2.

Atoms whose outer shell is already full have nothing to gain, lose or share, so their valency is zero and they are practically inert — that is the whole of the chemistry of the noble gases, and the reason helium, neon and argon are the standard examples. Inertness is not the same as being a good conductor: an atom with a full outer shell has no loose electrons at all, so a full outer shell means unreactive, not conductive.

Valency is also the fastest way to predict a formula. Combine two elements so that the total combining power on each side balances — in practice, swap the valencies and use them as subscripts:

Worked Examples

Hydrogen (1) with oxygen (2): two hydrogens are needed for one oxygen, giving $\mathrm{H_2O}$.

Carbon (4) with oxygen (2): one carbon takes two oxygens, giving $\mathrm{CO_2}$.

Aluminium (3) with oxygen (2): swap the valencies to get two aluminiums and three oxygens, $\mathrm{Al_2O_3}$ — the protective oxide named in the compounds table further on. The same arithmetic with iron in its trivalent form gives $\mathrm{Fe_2O_3}$.

Sodium (1) with chlorine (1): one of each, giving NaCl.

A shortage of a valence electron is always a shortage in the *outer* shell, never an inner one, because the inner shells are full and their electrons are far too tightly held to take part. That is the single idea behind every question that asks where a missing or extra bonding electron sits.

Valence Electrons and Electrical Behaviour

The same outer-shell count that predicts chemistry also predicts whether a material will carry current, which is why this paragraph of the syllabus underpins the electrical modules that follow.

- **One to three outer electrons — conductor.** The outer electrons are so loosely held that they detach and drift through the lattice, and an applied voltage sets them moving as a current. Copper is the default wiring material on the strength of its single loose outer electron; silver is marginally better and gold is used where a contact must not corrode.
- **Four outer electrons — semiconductor.** Silicon and germanium sit exactly on the boundary. Their electrons are locked into shared bonds at low temperature but can be freed by heat, light or the deliberate

addition of impurities.

- **Five to eight outer electrons — insulator.** The outer electrons are tightly bound into completed or nearly completed shells and cannot move, so no current flows. PTFE, polyimide, rubber, glass and ceramics all work on this principle: an insulator is defined by having its electrons tightly bound, not by having none.

Temperature moves the two classes in opposite directions, and the direction is a favourite question. In a **metal** the free electrons are already there, so heating only makes the lattice vibrate more violently and scatter them — resistance *rises* with temperature. In a **semiconductor** heating frees additional carriers that were not previously available, and that effect dominates — resistance *falls* with temperature. Aluminium is worth a note of its own here: with three valence electrons it conducts about 60 % as well as copper for a given cross-section, but it is roughly a third of the density, so for a given weight it carries more current. That is why heavy feeder cables and busbars are frequently aluminium while general airframe wiring is copper.

Molecules

A **molecule** is the smallest particle of a substance that retains the chemical properties of that substance. It consists of two or more atoms bonded together. Molecules of elements contain atoms of the same type (e.g. $\mathrm{O_2}$, $\mathrm{N_2}$), while molecules of compounds contain different types of atoms (e.g. $\mathrm{H_2O}$, $\mathrm{CO_2}$).

Monatomic, Diatomic and Polyatomic Molecules

How many atoms a molecule contains is set by how those atoms complete their outer shells, and the count has a name:

- **Monatomic** — a single atom is the whole molecule. Only the noble gases behave this way, precisely because they have no bonding to do: a cylinder of argon holds single argon atoms, never $\mathrm{Ar_2}$.
- **Diatomic** — two atoms. Hydrogen, nitrogen, oxygen, fluorine and chlorine all exist as $\mathrm{H_2}$, $\mathrm{N_2}$, $\mathrm{O_2}$, $\mathrm{F_2}$ and $\mathrm{Cl_2}$, which is why free oxygen is written $\mathrm{O_2}$ and not O.
- **Polyatomic** — three or more. Ozone is $\mathrm{O_3}$; water and carbon dioxide are polyatomic compounds; the hydrocarbon chains in kerosene run to a dozen or more carbons apiece.

A molecule of a single kind of atom is still an **element**, however many atoms it holds — $\mathrm{O_2}$ is elemental oxygen, not a compound. Only when unlike atoms are bonded together does the substance become a compound. Nitrogen makes the practical case: its two atoms are held by a triple bond, one of the strongest bonds between like atoms, and that is why nitrogen is so nearly unreactive at normal temperatures. It is that inertness — not simply the fact that it is a gas — that makes dry nitrogen the choice for tyre inflation, oleo charging and fuel-tank inerting, since it cannot support combustion and carries no moisture to freeze.

Reading a formula is a matter of distinguishing two kinds of number. The **subscript** counts atoms within one molecule; the **coefficient** in front counts whole molecules. So $\mathrm{H_2}$ is one molecule containing two hydrogen atoms, while $2\mathrm{H_2}$ is two such molecules containing four hydrogen atoms in total. Putting hydrogen and oxygen together to make water is written:

Formation of Water

$$2\mathrm{H_2} + \mathrm{O_2} \rightarrow 2\mathrm{H_2O}$$

Count both sides: four hydrogen atoms and two oxygen atoms before, four hydrogen atoms and two oxygen atoms after. Nothing has been created or destroyed, only rearranged — which is the conservation rule from the start of this section written in chemical shorthand. Adding the relative atomic masses the same way gives the relative molecular mass: water is $2 \times 1 + 16 = 18$, and carbon dioxide is $12 + 2 \times 16 = 44$.

Brownian Motion

Everything above treats molecules as real objects in constant motion. The direct evidence for that is **Brownian motion**: view smoke particles in air, or a fine powder suspended in a liquid, under a microscope and each visible speck jitters about along a random, endlessly changing path. Nothing is stirring the sample, and the specks are far too large to be moving of their own accord.

The explanation is that the visible particle is being struck from every side by molecules that are themselves invisible. At any instant the bombardment does not balance exactly, and the particle is knocked in whichever direction happens to be receiving fewer or slower impacts; a moment later the imbalance is elsewhere. Brownian motion therefore describes *the motion of molecules*, not their bonding or their breakdown, and it establishes two things at once: that matter is made of discrete particles small enough to be individually unnoticeable, and that those particles are in permanent random motion.

The temperature of a substance is a measure of the average kinetic energy of exactly that motion. Raise the temperature and the molecules move faster, so the jitter becomes more vigorous; cool the sample and it slows. That single link between temperature and molecular speed is what the rest of this note rests on — it explains why gases exert pressure (molecules striking the container walls), why smells and gas leaks spread through still air without any draught to carry them (diffusion), why a liquid evaporates faster when warm, and ultimately why heating a solid far enough will melt it.

Chemical Compounds

A **compound** is a substance formed when two or more elements chemically combine in a fixed ratio. The properties of a compound are usually very different from those of its constituent elements.

Compounds, Mixtures, Solutions and Alloys

The word that does the work in the definition above is **chemically**. Put two elements in the same container and you have a mixture; make them bond and you have a compound. The distinction is not academic — it decides how the substance behaves, how you separate it and what its properties will be.

Feature	Compound	Mixture
How the components are held	Chemically bonded — electrons transferred or shared	Not bonded — simply intermingled
Proportions	Fixed and definite (water is always two hydrogens to one oxygen)	Any proportion at all
Separation	Only by chemical means (electrolysis, heating, reaction)	By physical means (filtering, distilling, settling, magnetism)
Properties	Quite unlike those of the constituent elements	Each component keeps its own properties
Energy change on forming	Heat is given out or taken in	Essentially none
Melting and boiling	A single definite melting and boiling point	A range, set by the components present

Sodium chloride is the standard demonstration of how completely properties change. Sodium alone is a soft metal that reacts violently with water; chlorine alone is a poisonous green gas. Bonded, they make ordinary table salt. Nothing of the parent elements' behaviour survives the bond, because the outer-shell arrangement that produced that behaviour no longer exists.

Water is a compound and not a mixture, however casually the word "mixture" gets used, because its hydrogen and oxygen are chemically bonded in a fixed two-to-one ratio and can be separated only by electrolysis or another chemical process. Air, by contrast, is a mixture: its nitrogen, oxygen, argon and carbon dioxide are simply intermingled, each keeping its own properties, and they can be separated physically by cooling the air until the components liquefy at their different temperatures. That is exactly how commercial oxygen and nitrogen are produced.

Two intermediate cases matter in a hangar. A **solution** is a mixture in which one substance is dispersed through another right down to molecular scale — salt in water, or the small quantity of water that dissolves in

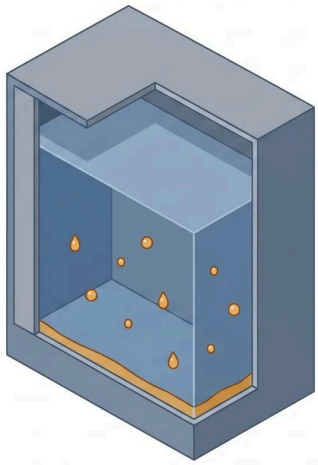
kerosene. It is uniform throughout and it is still a mixture, because no chemical bond has formed between solute and solvent. An **alloy** is a solid solution of one or more added elements in a host metal: the atoms of the added element occupy sites in the host lattice, distorting it, which is why alloying makes a metal stronger and harder than the pure element while lowering its electrical conductivity. Aluminium alloyed with copper, magnesium and zinc gives the airframe alloys; iron alloyed with carbon and chromium gives the steels. In neither case has a compound been formed, which is why an alloy's composition can be varied continuously to tune its properties instead of being locked to one fixed ratio.

Solutions, Solubility and Saturation

In any solution the substance that dissolves is the **solute** and the substance it dissolves in is the **solvent**. A solution that will accept no more solute at a given temperature is **saturated**, and how much it can hold before it reaches that point is the solute's **solubility**. Two temperature effects run in opposite directions, and both matter in a hangar.

- **A solid dissolved in a liquid is generally more soluble hot.** Heat helps break the solid apart and helps the solvent molecules surround its particles, so warming a solution lets it hold more — and cooling a saturated solution forces the excess back out as crystals.
- **A gas dissolved in a liquid is less soluble hot.** A dissolved gas molecule only needs enough kinetic energy to escape the surface, and heating supplies it. Warm a liquid and its dissolved gases come out of solution; this is why bubbles appear in water long before it boils.

Pressure works on gases the other way again: the more gas is pressed against the surface, the more dissolves, so raising the pressure increases the amount held in solution and dropping the pressure releases it. Between them these two rules explain a set of otherwise unconnected observations. Air dissolved in hydraulic fluid or engine oil under system pressure comes out as bubbles wherever the pressure falls — at a pump inlet, across a restrictor, or in a return line — producing foaming, a spongy response and, where the bubbles collapse again, cavitation erosion. Oxygen dissolved in water accelerates corrosion. And fuel always carries a small quantity of dissolved water, which comes out of solution as free water as the fuel cools.



Fuel always carries some dissolved water, which comes out as free water when the fuel cools.

A solution stays a mixture throughout: nothing is chemically bonded, the proportions are variable up to the saturation limit, and the solute can be recovered unchanged by physical means — evaporating the solvent, distilling it off, or simply letting the temperature change and collecting what separates out.

Types of Chemical Bonds

Bond Type	Mechanism	Example	Properties
Ionic	Electrons transferred from one atom to another, creating oppositely charged ions that attract	NaCl (sodium chloride)	High melting point, conducts electricity when dissolved, crystalline solid
Covalent	Atoms share one or more pairs of electrons	H_2O, CO_2	Can be gas, liquid, or solid; generally lower melting points
Metallic	Positive metal ions surrounded by a "sea" of delocalised electrons	Aluminium, copper	Good conductors of heat and electricity, malleable, ductile

How a Bond Forms: the Drive to a Full Outer Shell

All three bonding mechanisms above are the same idea carried out three different ways: each atom ends up with the electron arrangement of a noble gas. What differs is who ends up owning the electrons.

Ionic bonding — electrons are transferred. Sodium carries one electron in its outer shell and chlorine has seven. Sodium hands its single outer electron over, exposing the full shell beneath it and leaving a Na^+ ion; chlorine accepts it and completes its own octet as a Cl^- ion. The two are now oppositely charged, and the bond *is* the electrostatic attraction between them — there is no sharing involved. Because that attraction acts

equally in all directions, ionic solids build up into a regular crystal lattice with a very large number of such attractions to break, which is why their melting points are high. In the solid the ions are locked in place and cannot carry current; dissolve or melt the compound and the ions become free to move, and it conducts. Ionic bonding is characteristically a metal joining a non-metal.

Covalent bonding — electrons are shared. Two non-metals both need to gain electrons, and neither can, so they pool a pair instead: the shared pair counts towards the outer shell of both atoms at once. Two hydrogen atoms share one pair to give a single bond in $\mathrm{H_2}$. Two oxygen atoms, each two electrons short, share two pairs — a double bond. Two nitrogen atoms, each three short, share three pairs — the triple bond that makes nitrogen so unreactive. Oxygen forms two single bonds to two separate hydrogens in $\mathrm{H_2O}$; carbon, with four outer electrons, forms four bonds and gives the whole of organic chemistry, including the hydrocarbon chains of kerosene. Non-metals bonding with other non-metals form covalent bonds, and that is the pairing to look for in a question.

Metallic bonding — electrons are pooled across the whole solid. Metal atoms have few outer electrons and nothing to gain them from, so they release them into a shared cloud that belongs to the lattice as a whole rather than to any one atom. The positive ions left behind are held together by their attraction to that cloud.

The difference in what happens under load follows directly. Push on an ionic crystal hard enough to slide one plane of ions over the next and you bring like charges into contact; the crystal splits, which is why ionic solids are brittle. Push on a metal and the planes of positive ions slide while the electron cloud simply flows with them, so the metal deforms instead of fracturing. That is the mechanism behind every forming, bending and riveting operation carried out on an aircraft.

✈ **Aviation context:** Metallic bonding explains why metals are good electrical conductors (used for wiring) and are malleable (can be formed into aircraft skin panels). The "sea of electrons" allows current flow and mechanical deformation without fracturing.

Polar Molecules and the Forces Between Molecules

A bond *inside* a molecule and a force *between* whole molecules are two different things, and mixing them up is the usual reason a boiling point looks inexplicable. Breaking the covalent bonds inside water would give hydrogen and oxygen gas; boiling water only has to separate whole $\mathrm{H_2O}$ molecules from each other, and that takes far less energy. It is the weaker, between-molecule forces that set melting and boiling points, surface tension and viscosity.

In a covalent bond between unlike atoms the shared pair is not held equally. Oxygen attracts electrons more strongly than hydrogen does, so in a water molecule the shared electrons sit closer to the oxygen, leaving that end slightly negative and the hydrogen ends slightly positive. A molecule with such a permanent charge

separation is **polar**. Carbon dioxide is not polar even though its individual bonds are, because the molecule is symmetrical and the two pulls cancel; hydrocarbons such as kerosene are essentially non-polar for the same reason.

Three consequences follow, and all three are met daily:

- **Hydrogen bonding.** The positive hydrogen end of one water molecule is strongly attracted to the negative oxygen end of the next. This is much stronger than an ordinary intermolecular attraction, and it is why water boils at 100 °C when comparable molecules of similar size boil far below room temperature. It is also why water has an unusually high surface tension and an unusually high capacity to absorb heat.
- **Van der Waals forces.** Even non-polar molecules attract each other weakly, through the momentary uneven distribution of their own electrons. The forces are feeble but they add up along a long molecule, which is why heavier hydrocarbons are liquids and the lightest ones are gases.
- **Like dissolves like.** Polar solvents dissolve polar and ionic substances; non-polar solvents dissolve non-polar ones. Water dissolves salt and glycol; it does not dissolve kerosene, which is why water separates out and collects at the bottom of a fuel tank instead of staying mixed.

Two named cases of intermolecular attraction come up repeatedly in questions, and the only thing to keep straight is whether the molecules concerned are alike:

Force	Acts between	What it produces	Everyday example
Cohesion	Molecules of the same substance	Holds the substance together; surface tension; a droplet pulling itself into a bead	A drop of water holding its shape on a waxed surface
Adhesion	Molecules of different substances, at the surface where they touch	Makes one substance cling to or wet another	Water wetting glass; paint, adhesive and penetrant clinging to metal

Which of the two wins decides what a liquid does on a surface. Where adhesion to the solid beats cohesion within the liquid, the liquid spreads and wets the surface; where cohesion wins, it pulls away and beads instead. Every cleaning, painting and bonding operation depends on getting adhesion to win, which is why degreasing beforehand is not optional — a contaminated surface is one the liquid is being asked to adhere to and cannot.

Simple Molecular and Giant Covalent Structures

Covalent substances split into two families whose bulk properties could hardly be more different, and the difference is structural rather than chemical. In a **simple molecular** substance the covalent bonds tie a small fixed number of atoms into a discrete molecule, and only weak intermolecular forces hold one molecule to the

next. Melting or boiling need only overcome those weak forces, so the melting points are low and many such substances are gases or liquids at room temperature. Carbon dioxide, water and the hydrocarbons in fuel all belong here, and that is the family the general statement about lower melting points describes.

In a **giant covalent** structure there are no separate molecules at all. The covalent bonding continues in three dimensions throughout the whole crystal, so melting it means breaking strong covalent bonds by the million. These substances have some of the highest melting points known:

Structure	Arrangement	Behaviour	Aviation use
Diamond (carbon)	Each carbon covalently bonded to four others in a rigid three-dimensional network	Hardest natural material; an electrical insulator, since every electron is committed to a bond	Cutting, grinding and honing tools; hardness testing
Graphite (carbon)	Flat sheets of carbon bonded to three neighbours each, the sheets held together only weakly	Sheets slide over one another, so it is soft and slippery; the fourth electron is free, so it conducts	Dry lubricant; brushes and brush-type contacts; carbon fibre precursor
Silicon carbide, alumina, quartz	Giant networks of silicon, aluminium, carbon and oxygen	Extremely hard, high melting point, chemically stable, brittle	Abrasives and blast media; ceramic coatings; insulators

Carbon is the instructive case because diamond and graphite are the same element, differently arranged. One is the hardest substance in common use and an insulator; the other is soft enough to write with and conducts electricity. Nothing about the atoms differs — only the way the bonds are laid out — which is as clear a demonstration as there is that structure, not just composition, determines properties.

Chemical Reactions: Synthesis, Analysis and Catalysts

A chemical change rearranges the bonds between atoms and produces a substance with new properties; a physical change does not. Melting, dissolving and evaporating are physical — the substance is recoverable unchanged. Burning, corroding and curing are chemical.

Two words name the two directions, and each is the exact reverse of the other, so the pair is best learned together:

- **Synthesis** — building a compound up by chemically bonding two or more elements together, which is also how a compound is produced from commonly available raw materials. Sodium plus chlorine to give salt, or hydrogen plus oxygen to give water, are syntheses.
- **Analysis** — the reverse: splitting a chemical compound into simpler substances. Electrolysing water back to hydrogen and oxygen is an analysis.

A **catalyst** is a substance that *speeds a reaction up* without being consumed by it and without appearing in the final products. It does not start a reaction that could not otherwise happen, it does not stop one, and it does not reverse the direction — it lowers the energy barrier so that more of the collisions between reacting molecules succeed. Because it is not used up, a very small quantity keeps working indefinitely until it is poisoned or physically degraded. The everyday example is a catalytic converter, in which a thin layer of platinum-group metal converts exhaust gases without itself being consumed; the same principle appears in two-part adhesives and sealants, where an accelerator brings the cure time down without becoming part of the cured polymer.

Energy always moves in a reaction, in one direction or the other. An **exothermic** reaction releases heat, because the bonds formed are stronger than those broken: combustion, curing resin and corroding metal all warm their surroundings. An **endothermic** reaction absorbs heat from its surroundings. Combustion is the one an engineer must be able to write down:

Complete Combustion of a Hydrocarbon

$$\text{hydrocarbon} + \text{O}_2 \rightarrow \text{CO}_2 + \text{H}_2\text{O} + \text{heat}$$

Burn kerosene with sufficient oxygen and the products are carbon dioxide and water vapour. Starve it of oxygen and combustion is incomplete, producing carbon monoxide and soot instead — which is why carbon monoxide is the hazard associated with a leaking exhaust or a faulty combustion heater, and why visible smoke indicates a fuelling or combustion problem rather than merely a dirty engine. Combustion is a specific case of **oxidation**, which means any reaction in which a substance combines with oxygen or, more generally, loses electrons; **reduction** is the opposite. Corrosion is oxidation running slowly at ambient temperature, and it is the same chemistry as a fire, only spread over years.

Common Aviation-Related Compounds

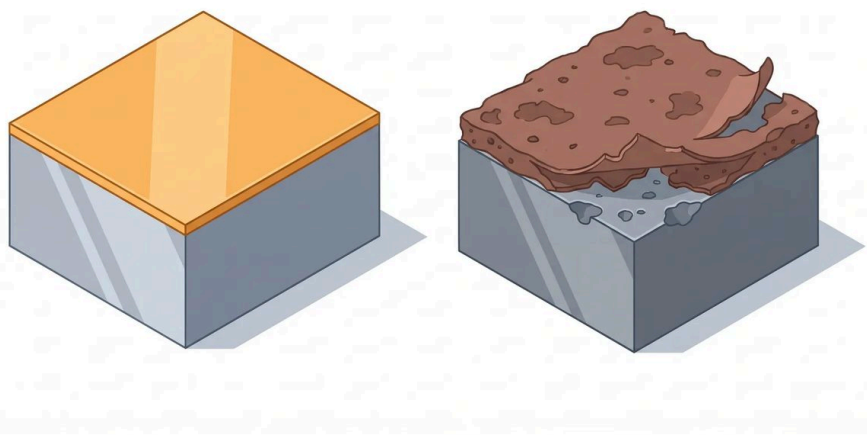
Compound	Formula	Aviation Relevance
Water	H_2O	Cooling, cleaning, weather effects
Carbon dioxide	CO_2	Fire extinguishing agent, combustion product
Aluminium oxide	Al_2O_3	Protective corrosion layer on aluminium
Iron oxide (rust)	Fe_2O_3	Corrosion product on steel — must be prevented
Kerosene (Jet A-1)	Hydrocarbon mix	Primary aviation turbine fuel

Oxides, Corrosion and Protection

The compounds table above puts aluminium oxide and iron oxide side by side, and the contrast between them is one of the most useful pieces of chemistry an airframe engineer has. Both are oxides. Both form spontaneously when the metal meets oxygen. One protects the metal underneath and the other destroys it, and the reason is structural.

Aluminium oxide forms almost instantly on bare aluminium and builds a film only a few molecules thick that is dense, hard, tightly adherent and impervious to oxygen. Once it has formed, it seals the surface and the reaction stops itself. That self-limiting film is the whole reason a metal as reactive as aluminium can be used as an airframe material at all — and it is why a freshly filed or abraded aluminium surface must be protected promptly, because the film that reforms on it is not the same as the controlled one that was removed.

Iron oxide forms a film that is porous, flaky and of greater volume than the metal it came from. It neither adheres nor seals; it lifts, exposing fresh metal underneath, and the process continues until the section is consumed. That is rust, and it is the reason steel components must be protected by a coating rather than allowed to build their own.



Aluminium oxide seals the metal in a thin, tight film; iron oxide flakes and lifts, exposing fresh metal.

The processes that follow from this are ordinary maintenance work:

- **Anodising** thickens the natural aluminium oxide film electrolytically into a much deeper, harder layer, which also holds paint well. Chromic and sulphuric acid processes are the usual routes.
- **Chemical conversion coatings** replace the natural film with a chemically formed one that resists corrosion and provides a paint key. Traditional chromate coatings used chromium in its hexavalent form,

which is why their use is now tightly controlled on health grounds and trivalent chromium alternatives have replaced them in many applications.

- **Cladding** bonds a thin layer of nearly pure aluminium to each face of a high-strength alloy sheet. The pure surface builds a better oxide film than the alloy would, so the strong core is protected by the weaker skin.
- **Titanium** forms an oxide film of the same self-sealing kind as aluminium's, but tougher and stable to much higher temperatures, which is much of the reason it is chosen for hot engine and exhaust-area structure.

Corrosion accelerates sharply when two dissimilar metals are in electrical contact in the presence of an electrolyte such as salt-laden moisture. The more reactive of the two gives up ions preferentially and corrodes, protecting the less reactive one — the reason dissimilar-metal joints are separated by a barrier, sealant or a compatible primer rather than being bolted directly together. The principle is deliberately exploited in a sacrificial anode, where a deliberately reactive metal is fitted to be consumed first.

Carbon dioxide appears in the same table for two unrelated reasons, and both are worth being clear about. It is a *product* of complete combustion, and it is also an extinguishing *agent*: it is denser than air and chemically stable, so it displaces the oxygen a fire needs without itself burning. Stored in a cylinder it is a liquid under its own vapour pressure; when it is discharged, the sudden expansion cools it so violently that some of it freezes into solid carbon dioxide, which is why a discharge produces visible white snow and why a bare-hand grip on the horn risks a cold burn.

Hydrocarbons and Aviation Fuel

A **hydrocarbon** is a compound of carbon and hydrogen only, and the whole family exists because carbon is tetravalent: with four outer electrons to share, a carbon atom can bond to four neighbours, including other carbon atoms, so the atoms can be strung into chains and rings of any length. Each carbon in the middle of a simple chain uses two of its bonds to link to its neighbours and the other two to carry hydrogen, while the two carbons at the ends of the chain carry three hydrogens each.

Chain length is what sets the physical properties, and it does so through the intermolecular forces rather than the bonds inside the molecule. A long molecule offers more surface for its neighbours to attract, so the attraction between molecules grows with chain length — and with it the boiling point, the density and the viscosity. That single relationship separates the products of a refinery: the shortest chains are gases at room temperature, the next are petrol and then kerosene, and the longest are lubricating oils, greases and waxes.

Aviation turbine fuel is described in the table above as a hydrocarbon mix, and the word *mix* is exact. It is not a compound with a formula but a mixture of many different hydrocarbons, blended to a specification. Because it is a mixture it has no single boiling point: it distils over a range, the lighter fractions evaporating first. Two properties of that mixture are worth carrying:

- The **flash point** is the lowest temperature at which a liquid gives off enough vapour to form a mixture with air that will ignite if a flame is applied. It is a property of the vapour above the liquid, not of the liquid itself, and it is the number that governs how a fuel must be handled. Jet A-1 has a specification minimum around 38 °C, so at ordinary ambient temperatures its tank vapour is normally too lean to ignite. Aviation gasoline has a flash point far below any temperature it will meet in service, so its vapour is ignitable essentially all the time — which is why the two fuels demand quite different handling precautions even though both burn identically once alight.
- The **freezing point** of a kerosene fuel is the temperature at which wax crystals that have formed as it cools disappear again on warming. It is a specification limit of the mixture, and it is a fuel-management concern on long high-altitude sectors.

Because it is a mixture and not a compound, fuel can also carry things it has not bonded with: dissolved water, dissolved air, and additives put in deliberately — anti-oxidants, static-dissipator additives to reduce the charge separated by fuel flowing through pipes and filters, and icing inhibitors. None of these is chemically part of the fuel, which is precisely why they can be blended in at controlled concentrations and why their effect can be lost by dilution or settling.

Polymers: Very Large Molecules

A **polymer** is a compound whose molecules are built by linking a small repeating unit, the **monomer**, into a chain thousands of units long. Nothing new happens chemically — the bonds along the chain are ordinary covalent bonds — but the sheer size of the molecule changes the bulk behaviour completely, because the weak attractions between neighbouring chains, feeble at any one point, add up over an enormous length.

How the chains are connected to each other divides polymers into three families that an engineer handles quite differently:

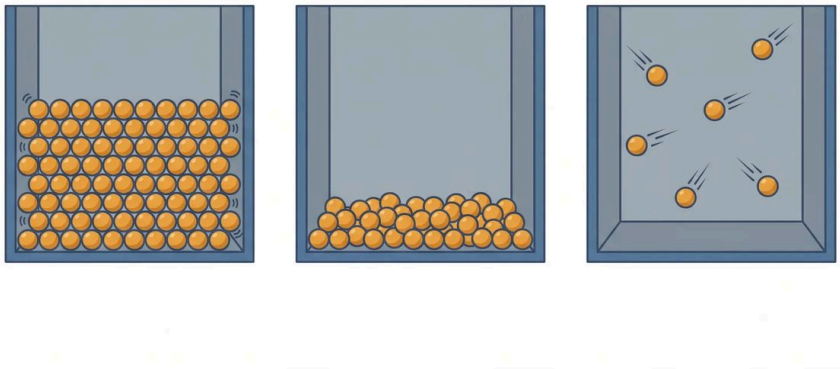
Family	Chain arrangement	Behaviour	Typical aircraft use
Thermoplastic	Separate chains held to each other only by weak intermolecular attraction	Softens when heated and hardens again on cooling, repeatedly; can be formed, welded and recycled	Acrylic and polycarbonate transparencies, cable insulation, ducting, PTFE hose liners
Thermoset	Chains chemically cross-linked into one continuous network during cure	Cannot be re-softened — heating it enough degrades it instead; cure is irreversible	Epoxy and phenolic matrices in composites, structural adhesives, potting compounds
Elastomer	Long coiled chains with a few widely spaced cross-links	Stretches greatly under load and returns to shape when released	Seals, O-rings, hoses, vibration mounts, tyre compounds

The cross-linking is the whole distinction, and it explains the storage rules as well as the mechanical ones. A thermoset resin is supplied as separate components, or frozen as a pre-impregnated fabric, precisely because the cross-linking reaction begins as soon as the components meet and warm up; once it has run, the network is a single molecule spanning the part and there is no way back. That is also why a cured composite is repaired by bonding on new material rather than by melting the old, and why cure temperature, cure time and out-time are controlled so carefully.

Polymers age in service in ways metals do not. Ultraviolet light, ozone, heat and contact with solvents or fluids all attack the chains or the cross-links, so an elastomer hardens, cracks and takes a permanent set, and a transparency crazes. That is the reasoning behind cure-date and shelf-life limits on rubber items, fluid-compatibility restrictions on seals, and the protection of transparencies from ultraviolet.

States of Matter

Matter exists in three principal states. The state depends on the balance between the **kinetic energy** of the particles (which tends to separate them) and the **intermolecular forces** (which tend to hold them together).



Solid, liquid and gas: packed and ordered, close but disordered, and widely spaced and moving freely.

Property	Solid	Liquid	Gas
Shape	Fixed	Takes shape of container	Fills entire container
Volume	Fixed	Fixed	Variable (fills container)
Particle arrangement	Regular, closely packed	Close but disordered	Widely spaced, random
Particle motion	Vibrate in fixed positions	Slide past each other	Move freely at high speed
Intermolecular forces	Very strong	Moderate	Very weak

Property	Solid	Liquid	Gas
Compressibility	Almost incompressible	Almost incompressible	Easily compressed
Density	High	High	Low

Why a Substance Is Solid, Liquid or Gas at a Given Temperature

The balance described above is not a metaphor — it is a straight competition between two measurable quantities, and knowing which one is winning tells you the state without having to memorise it.

One side of that competition can be changed and the other cannot. The particles' kinetic energy rises and falls with temperature. The intermolecular forces are fixed by what the substance is: how polar its molecules are, whether they can hydrogen-bond, how large they are, and whether the bonding continues through the whole solid rather than stopping at the edge of a molecule. Raising the temperature does not weaken those forces at all; it simply gives the particles more energy to fight them with. So:

- **Forces winning comfortably** — particles are locked into fixed positions and can only vibrate about them. The substance is a **solid** and holds its own shape.
- **Roughly matched** — particles have enough energy to break out of fixed positions and slide past each other, but not enough to escape one another's attraction altogether. The substance is a **liquid**: it keeps its volume but takes the shape of whatever holds it.
- **Kinetic energy winning** — particles escape the attraction entirely and move independently at high speed. The substance is a **gas**, with neither a shape nor a volume of its own; it expands until it fills the container.

This is why a substance's melting point is a direct readout of the strength of the forces holding its particles together. Compare, at atmospheric pressure:

Substance	What holds the particles	Melts at	Boils at
Nitrogen	Very weak forces between small non-polar molecules	About –210 °C	About –196 °C
Water	Hydrogen bonding between polar molecules	0 °C	100 °C
Aluminium	Metallic bonding	About 660 °C	Around 2,500 °C
Copper	Stronger metallic bonding	About 1,085 °C	Around 2,560 °C
Iron	Stronger metallic bonding again	About 1,538 °C	Around 2,860 °C
Titanium	Strong metallic bonding	About 1,668 °C	Around 3,290 °C
Tungsten	Exceptionally strong metallic bonding	About 3,420 °C	Around 5,550 °C

Nothing in that table is arbitrary: the order of the melting points is the order of the bond strengths. It also explains a fact worth carrying into engine work — tungsten is chosen for filaments and electrodes because it stays solid where every other common metal has vaporised, and aluminium alloys are restricted to cool structure because their melting point is low and their strength falls away well before it.

One qualification matters in a workshop. A pure substance melts at a single sharp temperature, because every bond in it is the same. An alloy or any other mixture does not: it begins to melt at one temperature and is not fully liquid until a higher one, so it passes through a pasty range in between. Welding and brazing depend on that range, and heat-treatment temperatures are chosen with reference to it.

Whether a solid is **crystalline** or **amorphous** matters for the same reason. In a crystalline solid — every metal, and most minerals — the particles occupy a regular repeating lattice, and there is one temperature at which the whole lattice comes apart. In an amorphous solid such as glass or a cured thermoset resin the particles are frozen in a disordered arrangement, and there is no single melting point at all: the material softens progressively over a range. That is why glass can be worked over a wide range of temperature rather than at one, and why a composite structure has a maximum service temperature quoted rather than a melting point.

Inside a Solid: Lattice, Grains and Why Metals Bend

Calling a solid's arrangement "regular and closely packed" is accurate but hides the detail that decides how a metal behaves under a hammer. The atoms of a metal sit in a repeating three-dimensional lattice, and which lattice it is has real consequences. Aluminium, copper, nickel and austenitic stainless steel adopt a face-centred cubic arrangement, which offers many planes along which atoms can slide, so those metals are notably ductile and stay ductile when cold. Ordinary steel at room temperature is body-centred cubic, with fewer such planes, and titanium and magnesium are close-packed hexagonal, with fewer still — which is a large part of why magnesium is difficult to form cold and why titanium is usually worked hot.

A real component is not one crystal. As metal solidifies, crystals start growing at many points at once and meet each other, so the finished part is a mosaic of small crystals called **grains**, each with its lattice pointing in a different direction, joined at grain boundaries. Deformation happens by planes of atoms slipping over one another within a grain, and a grain boundary obstructs that slip. Finer grains therefore mean more boundaries and a stronger, harder metal; this is why grain size is controlled during heat treatment and why overheating a component, which lets the grains grow, is so damaging.

Slip does not require a whole plane of atoms to move at once — it happens by the movement of line defects in the lattice, so the force needed is far less than a perfect crystal would demand. Everything that makes a metal harder works by making those defects harder to move: alloying (foreign atoms distort the lattice), work hardening (the defects tangle with one another), and heat treatment (fine particles are precipitated in their path). Annealing reverses it, letting the lattice re-order and restoring ductility. This is the mechanism behind the

statement earlier in this note that metals are malleable and ductile, and it explains why a metal that has been repeatedly flexed becomes hard and eventually cracks.

Solids are described as "almost incompressible" for a straightforward reason: the atoms are already as close as their mutual repulsion allows, so squeezing them further needs enormous pressure. What a solid does do under load is deform elastically — strain in proportion to stress, recovering when the load is removed — and that behaviour is the subject of the statics section rather than this one.

Inside a Liquid: Surface Tension, Capillarity and Viscosity

A liquid keeps its volume because its molecules are still within each other's attraction, and takes its container's shape because they are free to slide. Three consequences of that half-way condition are directly useful in maintenance.

Surface tension. A molecule in the body of a liquid is pulled equally in every direction by its neighbours. A molecule at the surface has neighbours only below and to the sides, so the net pull is inward. The surface behaves as though it were under tension, pulling itself to the smallest possible area — which is why a free droplet is spherical, why a small quantity of mercury beads up, and why a clean needle can be floated on water. Surface tension is a cohesive effect, and it falls as temperature rises, because the molecules are moving faster and holding each other less effectively.

Capillarity. Put a narrow tube into a liquid that wets it and the liquid climbs, because adhesion to the tube wall beats cohesion within the liquid and surface tension drags the bulk up behind the wetted edge. The narrower the gap, the higher it climbs. Put the same tube into a liquid that does not wet it — mercury in glass — and the level is pushed *down* instead. Capillary action is what draws dye penetrant into a surface-breaking crack during a penetrant inspection, and what draws it back out into the developer afterwards, so the whole method depends on the crack being clean enough for the penetrant to wet it. It is also why a hairline crack or an unsealed lap joint will pull water in and hold it there, and why capillary gaps are a classic corrosion site.



Capillary action draws penetrant back out of a crack into the white developer, showing the crack's line.

Viscosity. Viscosity is a liquid's internal resistance to flow, and it comes from the same intermolecular attraction plus the mechanical interlocking of molecules sliding past one another. The temperature dependence goes in opposite directions for the two fluid states, and the reason is worth understanding rather than memorising:

- In a **liquid**, heating gives the molecules enough energy to slip past their neighbours more easily, so viscosity *falls* as temperature rises. Cold oil is thick and hot oil is thin, which is why oil pressure is high and flow poor on a cold start, and why a viscosity index describing how little an oil thins with heat is a specified property of an aircraft lubricant.
- In a **gas**, there is almost no attraction to overcome; resistance comes from molecules crossing between fast and slow layers and transferring momentum. Heating makes them cross more often, so gas viscosity *rises* as temperature rises — the opposite of a liquid.

Liquids are described as almost incompressible on the same grounds as solids: the molecules are already nearly touching. "Almost" is the honest word, though. At the pressures a hydraulic system works at — commonly 3,000 psi, about 207 bar — the fluid does compress measurably, by of the order of one per cent of its volume. That is small enough for the hydraulic system to work at all, but large enough to matter: it is part of the sponginess felt in a system with a poorly matched accumulator, and it is one reason a system must be bled properly, since trapped air compresses by orders of magnitude more than the fluid does.

Inside a Gas: Spacing, Diffusion and the Air Itself

In a gas the molecules are perhaps ten diameters apart on average, moving in straight lines at hundreds of metres per second until they collide. Almost all of the volume is empty space, and that single fact accounts for the properties listed for a gas at the head of this section: it has no shape and no fixed volume because nothing holds its molecules together, it exerts pressure because those molecules hammer continuously on the

container walls, its density is low because so few molecules occupy a given volume, and it compresses easily because there is space to compress into.

Diffusion follows from the same random motion. Release a gas at one point in a still room and it spreads throughout, without any draught to carry it, simply because its molecules travel in random directions and collisions gradually distribute them evenly. This is why a fuel or hydraulic-fluid leak can be smelled some distance from its source, and why a leaking oxygen system enriches a whole compartment rather than a local pocket — a serious fire risk, because an oxygen-enriched atmosphere makes materials that are normally difficult to ignite burn fiercely.

It is worth separating two words that everyday usage runs together — while keeping hold of the fact that the separation is between two kinds of gas, not between a gas and something else: a vapour *is* a gas, and the three principal states of matter remain solid, liquid and gas. A **gas** is above its critical temperature and cannot be liquefied by pressure alone. A **vapour** is the gaseous form of something that is a liquid at ordinary temperature, and it will condense if compressed enough. Water vapour in the air is a vapour; the nitrogen alongside it is a gas.

Air itself is a mixture of gases, and its composition is examined directly:

Component	Approximate proportion by volume	Notes
Nitrogen	78 % (about four fifths)	Chemically inert at normal temperatures; the bulk of the atmosphere
Oxygen	21 % (about one fifth)	Supports combustion and respiration; the component that matters at altitude
Argon	About 0.9 %	A noble gas; the third most abundant component
Carbon dioxide and trace gases	Under 0.1 %	Includes water vapour, which varies widely with conditions

The working approximation is one fifth oxygen to four fifths nitrogen. Each gas in a mixture contributes its own share of the total pressure in proportion to how much of the mixture it forms, so at sea level oxygen accounts for roughly a fifth of the atmospheric pressure. That proportion does not change with altitude — air remains about 21 % oxygen far above the cruising levels of any transport aircraft — but the total pressure falls, so the oxygen's share of it falls with it. That, and not any change in composition, is why supplemental oxygen and cabin pressurisation are needed.

One practical point follows directly from a gas being compressible: it stores a great deal of energy in its own expansion, so a failed pneumatic component, a burst tyre or a dropped high-pressure cylinder releases that energy explosively — which is why a nitrogen charging trolley is fitted with a regulator and why a wheel

assembly is deflated before it is split. A hydraulic system, holding a nearly incompressible fluid, stores far less energy for the same pressure and volume, so a hydraulic leak is a hazard chiefly through its jet and its fluid, not through stored expansion.

> ✈ **Aviation context:** Hydraulic systems rely on liquids being virtually incompressible — a force applied at one point is transmitted faithfully through the fluid. Pneumatic systems use gases (compressed air or nitrogen) and exploit their compressibility. The airframe itself is a solid structure designed to resist deformation.

How State Changes Density

The density entries in this section's opening table — high for solids and liquids, low for gases — follow directly from particle spacing, and the size of the difference is worth having a feel for. In a solid or a liquid the particles are touching, while at the gas spacing described above the same number of particles occupies of the order of a thousand times the volume. Liquid water at 1,000 kg/m³ becomes steam at atmospheric pressure with a density of only about 0.6 kg/m³, and air at sea level is about 1.225 kg/m³ — roughly eight hundred times lighter than the water it sits above.

Between solid and liquid the change is slight, because the particles do not move appreciably further apart — only their arrangement changes. That is why a solid usually sinks in its own melt, and why ice floating on liquid water stands out as the exception.

Material	Approximate density (kg/m³)	Note
Air at sea level	1.2	Falls with altitude as pressure falls
Aviation kerosene	About 800	Varies with temperature and between batches
Ice at 0 °C	917	Less dense than its own liquid — the anomaly
Water	1,000	The reference against which relative density is quoted
Carbon fibre composite	About 1,600	Roughly 40 % lighter than aluminium alloy
Magnesium alloy	About 1,800	The lightest structural metal in aircraft use
Aluminium alloy	About 2,800	The principal airframe metal
Titanium alloy	About 4,500	Denser than aluminium but far stronger and heat resistant
Steel	About 7,800	Landing gear, fasteners, engine components

Those figures are the reason composite and aluminium structure dominate an airframe while steel and titanium are confined to the places where their strength or temperature capability is genuinely needed. Density

— mass per unit volume — is developed further, together with the way materials are compared against water and its role in fluid pressure and buoyancy, in the fluids sections of this module.

Plasma: the Fourth State

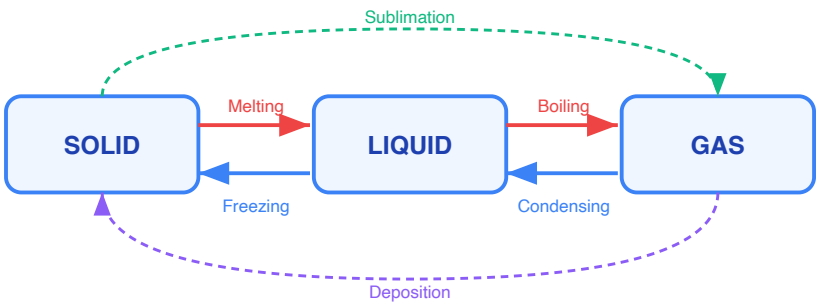
Solid, liquid and gas are the three states a substance passes through as it is heated, and they are the three the syllabus concentrates on. Keep heating a gas, though, and something further happens: the collisions become violent enough to knock electrons off the atoms themselves. What remains is a mixture of free electrons and positive ions, still gaseous in the sense that it has no fixed shape or volume, but electrically quite unlike a gas. That is **plasma**, and it is normally counted as the fourth state of matter — which is why a question offering a choice of four states is not a trick.

A plasma differs from a gas in three ways that matter. Because it contains free charge carriers it *conducts electricity*, where an ordinary gas is an insulator. Because those carriers are charged it responds to electric and magnetic fields. And because recombining ions release their energy as light, a plasma glows.

Engineers meet it more often than the exotic name suggests: in a lightning channel and in the corona around a highly charged surface; in the arc of an arc welder and the much hotter arc of a plasma cutter; in the ionised gas used in plasma-spray coating to deposit ceramic thermal barriers on turbine components; and in every fluorescent tube, neon sign and old-style discharge lamp. On the largest scale, plasma is the most common state of matter in the universe, since stars are made of it — but on Earth it needs to be produced deliberately.

Changes Between States

When heat energy is added to or removed from a substance, it may change state. These changes are **physical** (not chemical) — the substance itself is unchanged, only the arrangement and energy of its particles differs.



Change	Direction	Energy
Melting (fusion)	Solid → Liquid	Heat absorbed (endothermic)
Boiling (vaporisation)	Liquid → Gas	Heat absorbed (endothermic)
Freezing (solidification)	Liquid → Solid	Heat released (exothermic)
Condensation	Gas → Liquid	Heat released (exothermic)
Sublimation	Solid → Gas (directly)	Heat absorbed
Deposition	Gas → Solid (directly)	Heat released

What Happens to the Particles at Each Change

Each of the six changes is one event described at particle level, and picturing that event is what makes the energy direction obvious rather than something to be memorised.

- **Melting.** The particles are already touching; heating them only has to shake them free of their fixed lattice positions so they can slide. They stay in contact, so the volume barely changes — for most substances it increases by a few per cent, because the disordered liquid packs slightly less efficiently than the ordered crystal.
- **Boiling.** Far more is being asked: every particle has to be pulled clear of its neighbours' attraction altogether and given room to move independently. That is why the volume change is enormous. One kilogram of water occupies about a litre; as steam at 100 °C and atmospheric pressure that same kilogram occupies roughly 1,670 litres — an expansion of more than 1,600 times — and it is why a small quantity of trapped water flashing to steam can do structural damage, and why an autoclave or a steam line is treated with such respect.
- **Sublimation.** A particle at the surface of a solid escapes straight into the gas phase without ever being part of a liquid. It happens because the particles in a solid do not all carry the same energy: a few in the tail of the distribution have enough to break away, even far below the melting point. A material capable of doing this is a **sublimate**.
- **Deposition.** The exact reverse — gas molecules losing energy at a cold surface and locking straight into a solid lattice without a liquid stage. Frost forming on a cold wing overnight is deposition, not freezing: the water was vapour in the air, never liquid on the metal.
- **Freezing and condensation** simply run melting and boiling backwards, and release exactly the energy those changes absorbed.

The energy direction follows from that picture with no memory work at all. Any change that pulls particles further apart — melting, boiling, sublimation — must supply the energy to overcome the attraction, so heat is *absorbed* and the change is endothermic. Any change that lets particles fall back together — freezing, condensation, deposition — releases that same energy, so heat is *given out* and the change is exothermic. If you can decide which way the particles are moving, you can always decide the sign.

Because these are physical rather than chemical changes, they are all reversible and none of them alters the substance itself. Water is water whether it is ice, liquid or steam, and a compound like water can therefore exist in all three states — which is why boiling a kettle produces steam and not hydrogen and oxygen. Only a chemical change would break the molecule apart.

Evaporation Is Not Boiling

Both turn a liquid into a gas, but they are not the same process and the differences are worth holding separately.

	Evaporation	Boiling
Where it happens	Only at the free surface	Throughout the bulk of the liquid, with bubbles forming inside it
Temperature	Any temperature at all	Only at the boiling point for the prevailing pressure
Rate	Slow; increases with temperature, surface area and air movement	Rapid, and limited only by how fast heat can be supplied
Effect on the liquid left behind	Cools it	Temperature stays constant while it boils

The mechanism of evaporation is worth stating properly, because the cooling effect and the humidity behaviour that follow both fall straight out of it. Molecules in a liquid do not all move at the same speed; there is a spread. A molecule that happens to be at the surface, moving upward, and carrying more than the average energy can break free of its neighbours' attraction and escape as gas. It does not need the liquid to be hot — a puddle dries at 5 °C.

Two consequences follow. First, evaporation is selective: it removes the *fastest* molecules, so the average energy of those left behind falls and the liquid cools itself. That is why solvent or fuel on the skin feels cold, why a wetted surface reads below air temperature, and why an aircraft standing in rain or with fuel spilled on it can chill locally below the surrounding air. Second, evaporation continues until the air in contact with the liquid can hold no more — the air is then **saturated**, and molecules are returning to the liquid as fast as they leave.

How much vapour air can hold rises steeply with its temperature. **Relative humidity** is the amount of water vapour present expressed as a percentage of the maximum the air could hold at that temperature. Cool a sample of air without changing its water content and the maximum falls until it equals what is actually there: at that temperature, the **dew point**, the air is saturated and any further cooling forces water out as condensation. If the dew point is below freezing, the water is deposited directly as frost instead, and that temperature is called the frost point.

This is the single most useful idea in this section for practical work. Condensation and frost do not appear because the air is wet; they appear because a *surface* is at or below the dew point of the air touching it. Warm humid air meeting cold aircraft structure — a fuselage descending from cruise, a fuel tank cold-soaked after a long flight, a windscreen on a humid morning — produces water on the metal whatever the weather is doing. That is one of the ways water accumulates in a fuel tank sump, and it is why the sump is drained before flight rather than after.

Pressure Changes the Melting and Boiling Points

The melting and boiling points quoted for a substance are only valid at a stated pressure, and boiling is the more sensitive of the two by far. A liquid boils when its molecules can form vapour bubbles inside it, and they can only do that when the vapour's own pressure has risen to match the pressure pushing down on the surface. Raise the surrounding pressure and the liquid must be hotter before that happens; lower it and boiling starts sooner.

- At the standard sea-level pressure of 1,013 hPa (about 14.7 psi) water boils at 100 °C. That figure is a consequence of the pressure, not a property of water on its own.
- At a cabin altitude of 8,000 ft the pressure is around 75 kPa and water boils at roughly 92 °C. On an unpressurised aircraft at high altitude it is lower still.
- Raise the pressure instead and the boiling point rises. That is the principle of a pressurised cooling system, which allows a coolant to run above 100 °C without boiling, and of an autoclave, which cures composite structure at a temperature that would boil off the volatiles at atmospheric pressure.

Melting points shift far less with pressure, and water is the notorious exception to the usual direction. For almost every substance the solid is denser than the liquid, so squeezing it favours the solid and raising the pressure *raises* the melting point. Water's solid is *less* dense than its liquid, so pressure favours the liquid and raising the pressure *lowers* the melting point of ice. It is a small effect — it takes a great deal of pressure to move the melting point by a degree — but it is the reason ice under high local contact pressure can develop a film of water, and it is a direct consequence of the expansion-on-freezing dealt with below.

⚠ **Reduced pressure works both ways:** a fluid that is comfortably below its boiling point at sea level may not be at altitude or on the suction side of a pump, where the local pressure is far lower than the system's nominal figure. Fuel vapour forming in a line is one form of vapour lock; the same mechanism drives cavitation, where vapour bubbles form in a low-pressure region of a pump and then collapse violently as the pressure recovers, eroding the metal.

The Triple Point and the Critical Point

Plot pressure against temperature for a substance and the three states occupy three regions, separated by lines along which two states can coexist. The melting point line, for instance, is the set of conditions at which solid and liquid are in equilibrium — which is the precise definition of a melting point: not the temperature at which a solid vanishes, but the temperature at which its solid and liquid phases can exist together indefinitely, neither one growing at the other's expense.

Two points on that diagram have names. The triple point is asked about directly; the critical point earns its place for a different reason, because it is what divides a gas from a vapour:

- The **triple point** is the single combination of temperature and pressure at which all three states — solid, liquid and vapour — coexist in equilibrium. For water it is 0.01 °C, so for exam purposes approximately 0 °C, at a pressure of only about 611 Pa (roughly 6 mbar, well under one hundredth of atmospheric). Do not confuse it with the dew point, which is a property of a particular sample of air and its moisture content, not a fixed property of the substance.
- The **critical point** is the temperature above which a substance cannot be liquefied by pressure alone, however hard it is squeezed. For water it is about 374 °C at roughly 22 MPa. Above its critical temperature a substance is a gas in the strict sense; below it, the gaseous form is properly a vapour.

The diagram also explains why some substances sublime in ordinary conditions and others never do. If a substance's triple-point pressure is *below* atmospheric pressure, then at atmospheric pressure heating the solid takes it into the liquid region first, and it melts — which is what water does, its triple point being far below one atmosphere. If the triple-point pressure is *above* atmospheric, there is no liquid region to pass through at atmospheric pressure at all, and the solid goes straight to gas.

Carbon dioxide is the standard demonstration: its triple point is at about 5.1 atmospheres, so at ordinary atmospheric pressure liquid carbon dioxide cannot exist and solid carbon dioxide sublimes directly to gas at about –78.5 °C. That is why it is called dry ice, why it leaves no puddle, and why it is used to keep cargo cold without wetting it. Inside a pressurised extinguisher cylinder, where the pressure is far above 5.1 atmospheres, the same substance sits quite happily as a liquid.

Fixed Points of Water at a Glance

Melting and freezing point at 1 atmosphere: 0 °C. Boiling point at 1 atmosphere: 100 °C. Triple point: 0.01 °C at about 611 Pa. Critical point: about 374 °C at about 22 MPa. Maximum density: about 4 °C. Absolute zero, for reference, is –273.15 °C.

Why Water Expands When It Freezes

Almost every substance contracts as it solidifies, because an ordered lattice packs its particles more tightly than a disordered liquid. Water does the opposite, and the reason is the hydrogen bonding described earlier. As water cools towards freezing, its molecules line up so that each hydrogen bond can point directly at a neighbouring oxygen, and that arrangement is an open hexagonal framework with a hole at the centre of every ring. The ordered structure is *less* densely packed than the jumbled liquid.

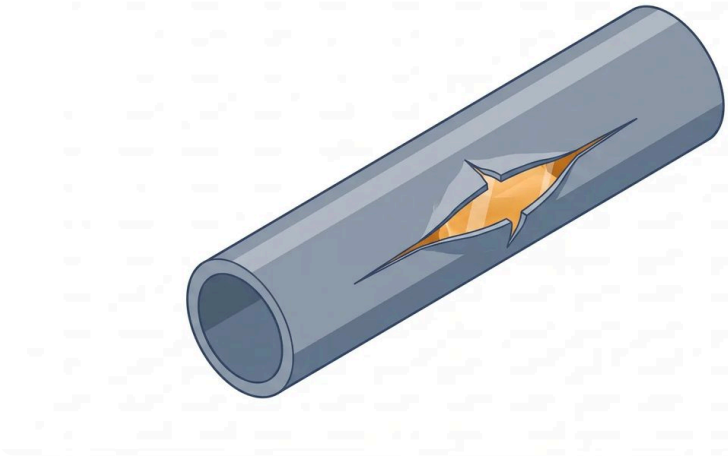
The numbers make the consequences concrete. Liquid water has a density close to 1,000 kg/m³, while ice at 0 °C is about 917 kg/m³. Freezing a given mass of water therefore increases its volume by about 9 %. Run it the other way and the same arithmetic says that melting a block of ice leaves the water occupying about 8 % *less* space than the ice did: melted ice takes up less room, not more.

Worked Example

Take 1 kg of ice. Its volume is $1 \div 917 = 0.00109$ m³, or 1.09 litres. Melt it and the same kilogram of water occupies $1 \div 1000 = 0.00100$ m³, or 1.00 litre. The mass has not changed — nothing was lost — but the volume has fallen by 0.09 litre, about 8 % of the original.

Being less dense, ice floats: it displaces its own weight of water, and in fresh water roughly 8 % of its bulk stands proud of the surface. Water is also unusual in reaching its maximum density at about 4 °C rather than at its freezing point, so the coldest water in a still pond rises to the top and freezes there, leaving the depths liquid.

In maintenance the expansion is a hazard rather than a curiosity, because it is powerful and it happens in confined spaces. Water trapped in a blind hole, a lap joint, a honeycomb core or a control-surface hinge line expands as it freezes with enough force to spread a joint, unseat a fastener, split a water line or crush a honeycomb cell — and the damage is only discovered later, when the ice has melted and left no evidence. It is also why potable-water and waste systems in unheated areas are drained or heated, and why water found in a fuel sump matters as much for what it can do to structure as for what it does to the fuel.

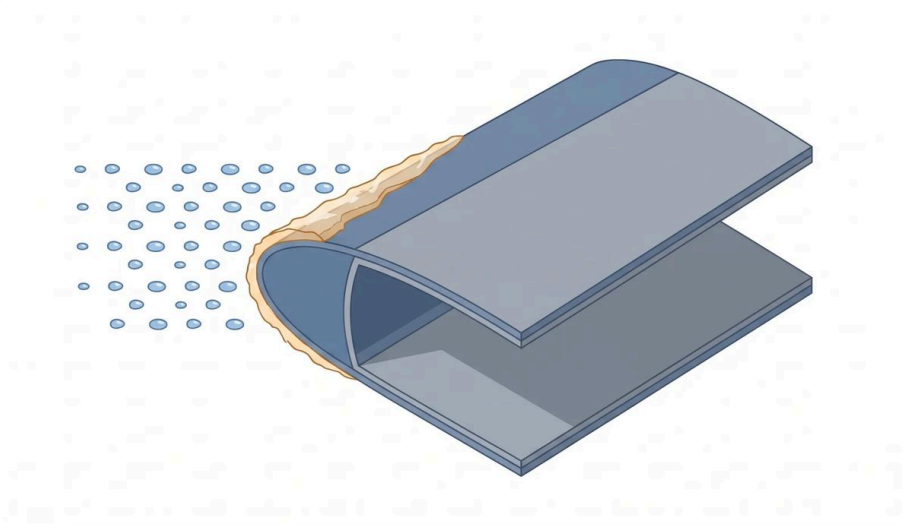


Water trapped in a line expands about 9 % as it freezes, with enough force to split the pipe wall.

Supercooling and the Effect of a Dissolved Solute

A liquid does not always freeze the moment it reaches its freezing point. To start freezing it needs a **nucleus** — a speck of dust, a scratch on the container, an existing ice crystal — on which the first ordered lattice can build. Clean water in a smooth container, or a small airborne droplet with nothing to nucleate on, can be cooled well below 0 °C and remain liquid. That condition is called **supercooling**, and it is unstable: disturb the liquid or give it a nucleus and it freezes almost instantly, its temperature jumping back up to the freezing point as the latent heat it had been holding is released.

This is not a laboratory curiosity for an aircraft engineer — it is the whole basis of in-flight icing. Cloud droplets are small, clean and suspended, so they routinely exist as liquid water at temperatures far below freezing. The aircraft arriving through them is the nucleus: on impact the droplet freezes onto the structure at once. No supercooling, no airframe icing: ice crystals that are already frozen largely bounce off a cold surface rather than adhering to it.



Supercooled cloud droplets stay liquid below freezing until they strike the wing, then freeze on at once.

Dissolving anything in water moves both of its fixed points, and the directions are opposite:

- The **freezing point falls**. Solute particles get in the way of water molecules trying to lock into the ice lattice, so a lower temperature is needed before freezing can proceed. Salt water freezes below 0 °C, and a glycol and water mixture stays liquid far below it — the more concentrated the mixture, the lower the freezing point.
- The **boiling point rises**. Solute particles at the surface reduce the rate at which water molecules escape, so a higher temperature is needed to reach boiling. This is why an engine coolant mixture both resists freezing in winter and boils at a higher temperature than plain water.

Both effects depend on how much is dissolved rather than on what it is, so they are undone by dilution: add more water to the mixture and its freezing point climbs back towards 0 °C. That is the physics a ground de-icing operation is working with, and it is taken up again at the end of this section.

Latent Heat

During a change of state, the temperature remains constant even though heat is being added or removed. The energy is used to break (or form) intermolecular bonds rather than to change the temperature. This energy is called **latent heat**.

⚠ **Key concept:** During melting or boiling, the temperature stays constant until the change of state is complete. A graph of temperature vs. time shows flat "plateaus" at the melting and boiling points.

The two latent heats have names and standard values. The **latent heat of fusion** is the energy needed to melt one kilogram of a solid already at its melting point; for water it is about 334 kJ/kg. The **latent heat of vaporisation** is the energy needed to boil one kilogram of a liquid already at its boiling point; for water it is about 2,260 kJ/kg. Freezing and condensing release exactly those same quantities back again — the direction reverses, the magnitude does not.

Setting them beside the energy needed simply to warm the water shows how large they are. Raising 1 kg of water through the whole 100 °C from freezing to boiling takes about 419 kJ. Boiling that same kilogram away, at constant temperature, takes about 2,260 kJ — more than five times as much energy for no rise in temperature whatever.

Worked Example

How much energy is needed to turn 2 kg of ice at 0 °C into 2 kg of water at 0 °C?

Only the change of state is involved, so only the latent heat of fusion applies: $2 \times 334 = 668$ kJ. The thermometer reads 0 °C at the start and 0 °C at the end, and all 668 kJ has gone into breaking the lattice apart.

The same figure explains why removing ice from an airframe is expensive in energy: every kilogram of it needs about 334 kJ to melt, on top of whatever is needed to warm it to 0 °C first and whatever the airflow carries away in the meantime.

The reverse direction is the reason steam is so dangerous. Steam at 100 °C condensing on skin delivers the whole 2,260 kJ/kg of latent heat at the point of contact before the resulting water has cooled by a single degree, which is why a steam burn is far worse than a scald from boiling water at the same temperature. The same effect works usefully elsewhere: a **vapour-cycle** refrigeration system moves heat by evaporating a refrigerant where heat is to be absorbed and condensing it where heat is to be rejected, and the latent heat is what does the carrying. An **air-cycle** machine, which is what most turbine-powered aircraft use instead, is the instructive contrast: its working fluid is the air itself, passing straight through the machine and into the cabin rather than round a sealed loop, and it is cooled by the shaft work taken out of it as it expands through a turbine. Nothing changes state to carry the load there, so the heat is carried as *sensible* heat — a temperature change rather than a change of state — and a much larger mass flow is needed for the same cooling duty. Any moisture that condenses in such a system is water being wrung out of the air stream for removal, not a refrigerant condensing to reject the load.

The relative size of the two latent heats also sets the shape of the temperature-against-time graph. Supply heat at a steady rate to a block of ice and the trace rises, flattens at 0 °C while the ice melts, rises again through the liquid range, then flattens a second time at 100 °C while the water boils. The second plateau is far the longer of the two — about seven times as long for water, since 2,260 kJ/kg has to be delivered at the boiling point against 334 kJ/kg at the melting point. Cooling the same substance reproduces the identical trace in reverse, with heat coming out at each plateau instead of going in.

Changes of State in Aircraft Operation

✈ **Aviation context:** Ice formation on aircraft surfaces is a critical hazard — water changes state from liquid to solid. De-icing systems provide heat energy to reverse this. Fuel can also contain dissolved water that freezes at altitude, which is why fuel-icing inhibitors (FSII) are used.

The mechanisms behind that are worth separating, because they call for different remedies. **Airframe icing in flight** is supercooling put into practice: small droplets freeze on impact almost immediately and give opaque rime ice, while larger ones spread across the surface before freezing and give clear ice, which is heavier, harder and adheres more strongly. **Ground frost** is a different process altogether: it is deposition, water vapour turning straight to solid on structure that is below the frost point, and it forms on a clear night with no precipitation at all. A wing that has been cold-soaked by cold fuel can grow frost on its upper surface in the morning even when the air temperature is above freezing.

The counter-measures divide the same way. An **anti-icing** system is run before ice can form and prevents accretion; a **de-icing** system is allowed to accrete a layer and then removes it. Thermal systems, whether they bleed hot air or use electrical heating elements, supply the latent heat needed to melt the ice at the surface. Ground de-icing and anti-icing fluids work on a different principle again: they are glycol-based, so they depress the freezing point of the water film rather than adding heat to melt it, and the holdover time they provide is the period before dilution undoes that protection.

Sublimation and deposition appear in the same operation. Frost on a cold-soaked upper surface can sublime away in dry air at altitude without ever melting; conversely, moisture in an oxygen or pneumatic system can deposit as ice at a restriction and block it, which is why aviator's breathing oxygen is specified dry and why moisture separators are fitted where compressed air is used.

Fuel deserves its own note because two separate things happen to it in the cold. Jet A-1 is specified to a maximum freezing point of about –47 °C, below which wax crystals are present in the fuel and can block filters; that is a property of the fuel, and it is managed by fuel specification and by monitoring fuel temperature. Quite separately, the small quantity of water that fuel always carries in solution comes out as the fuel cools and freezes into ice crystals at around 0 °C, long before the fuel is anywhere near its own freezing point. Ice crystals are the reason for water drains, filter icing indications and inhibitor additives. Confusing the two leads to the wrong troubleshooting entirely.

Finally, the same physics runs in the useful direction throughout an aircraft. Water drains work because water is denser than fuel and separates to the lowest point. Carbon dioxide and halogenated extinguishing agents are stored as liquids and discharge as gases, the change of state cooling the discharge sharply. Air-conditioning water separators condense moisture out of cabin air by cooling it below its dew point. Cabin

condensation forms on cold structure behind trim panels during descent into warm humid air, and it drips back out later — a nuisance, and a corrosion risk in exactly the places that are hardest to inspect.

Module 3 — Electrical Fundamentals

3.1 — Electron Theory

Electron theory explains the fundamental nature of electricity by examining atomic structure and the behaviour of electrons. This section lays the groundwork for understanding how electrical current flows and why different materials conduct electricity to different degrees — essential knowledge for every aircraft maintenance engineer working with electrical systems.

Structure of Matter

All matter is composed of **atoms**, the smallest particles of an element that retain its chemical identity. Each atom consists of a central **nucleus** containing positively charged **protons** and (except in ordinary hydrogen) electrically neutral **neutrons**, surrounded by orbiting negatively charged **electrons**.

Particle	Location	Charge	Relative Mass
Proton	Nucleus	Positive (+1)	1
Neutron	Nucleus	Neutral (0)	1
Electron	Shells (orbits)	Negative (−1)	≈ 1/1836

In a neutral atom the number of electrons equals the number of protons, so the overall charge is zero. The **atomic number** (Z) equals the number of protons and defines the element. The **mass number** (A) is the total of protons plus neutrons.

Electron Shells and Energy Levels

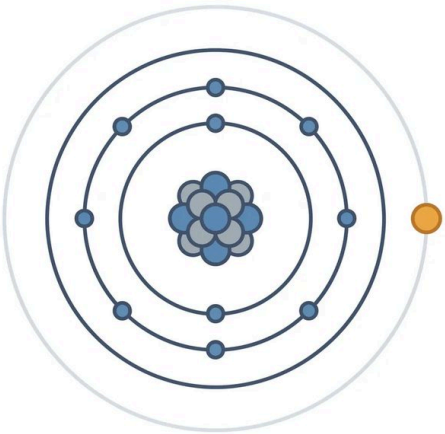
Electrons occupy discrete energy levels called **shells**, labelled K, L, M, N outward from the nucleus. Each shell has a maximum capacity:

Shell	Number (n)	Max Electrons
K	1	2
L	2	8
M	3	18
N	4	32

Maximum Electrons per Shell

$$\text{Max electrons} = 2n^2$$

The outermost shell is called the **valence shell**, and the electrons in it are **valence electrons**. These determine the electrical and chemical properties of the element.



A nucleus of protons and neutrons inside electron shells, with the outer valence electron highlighted.

Charge, Mass and Scale

The charge on a proton and the charge on an electron are exactly equal in size and opposite in sign. That size is the **elementary charge**, and no smaller quantity of free charge exists anywhere in nature — charge comes only in whole multiples of it. A proton carries $+1.602 \times 10^{-19}$ C; an electron carries -1.602×10^{-19} C. The minus sign is part of the value and should never be dropped: it is the reason adding electrons to a body makes it more negative rather than more positive.

Elementary Charge

$$e = 1.602 \times 10^{-19}\ \text{C} \qquad \text{so} \qquad 1\ \text{C} = \frac{1}{1.602 \times 10^{-19}} \approx 6.24 \times 10^{18}\ \text{electrons}$$

The masses are nothing like as balanced as the charges. A proton has a mass of about 1.673×10^{-27} kg and a neutron is marginally heavier still, while an electron is only about 9.11×10^{-31} kg — some 1,836 times lighter. Two consequences follow directly. Almost the whole mass of an atom is concentrated in its nucleus, so the mass number counts nucleons and ignores electrons altogether. And because electrons weigh so little, an

atom that gains or loses a few of them changes its electrical charge dramatically while its mass barely alters at all.

The nucleus is also extraordinarily small compared with the atom that surrounds it: of the order of 10^{-15} m across against roughly 10^{-10} m for the whole atom, so its diameter is around one hundred-thousandth of the atom's. An atom is therefore mostly empty space, and what one atom presents to its neighbours is not its nucleus but the outer edge of its electron cloud. This is why every chemical property and every electrical property of a material is settled by the outermost electrons, and why nothing an aircraft engineer does to a wire ever touches its nuclei.

What holds the structure together is the electrostatic attraction between the positive nucleus and the negative electrons. That force obeys an inverse-square law, so an electron twice as far out is attracted only a quarter as strongly; on top of that, the inner shells of electrons partly screen the nuclear charge from the outer ones. Both effects point the same way: **the further out a shell lies, the more weakly its electrons are held**. Everything in this note about conductors, insulators and ionisation is a consequence of that one statement.

Neutrons, Isotopes and the Hydrogen Exception

Because the mass number counts protons plus neutrons and the atomic number counts protons alone, the number of neutrons in any nucleus is simply the difference between them. A copper-63 atom has $63 - 29 = 34$ neutrons alongside its 29 protons, and if it is electrically neutral it also has 29 orbiting electrons. An aluminium-27 atom has $27 - 13 = 14$ neutrons and 13 electrons.

Atoms of the same element can carry different numbers of neutrons; these variants are called **isotopes**. They share an atomic number, so they are chemically and electrically identical — conduction depends entirely on electrons, and isotopes differ only in nuclear mass. The one case worth memorising is the lightest atom of all: ordinary hydrogen consists of a single proton with a single electron orbiting it and **no neutron at all**. This is why the correct statement is that neutrons are found in the nucleus of *most* atoms rather than all of them.

> ⚠ **Two exam traps:** an atom is the smallest part of an **element** that retains that element's characteristics — it is *not* "the smallest particle of matter", because protons, neutrons and electrons are all smaller than an atom. And a neutron sits in the nucleus of **most** atoms, not all: hydrogen-1 has none.

How the Shells Actually Fill

The $2n^2$ rule gives the *maximum capacity* of a shell, not the order in which electrons are added. In a neutral atom in its normal, lowest-energy state the outermost shell never holds more than eight electrons, no matter

how large its capacity. So filling proceeds K, then L, then M up to eight — and at that point the next electron starts the N shell rather than continuing to fill M toward its capacity of 18.

Potassium shows this clearly. With 19 electrons it fills as 2, 8, 8, 1: the M shell stops at eight and the nineteenth electron opens the N shell. Potassium therefore has a single, remote, weakly held valence electron and behaves as a strongly conducting metal, even though its M shell holds only eight of the eighteen electrons it could take. Copper, with 29 electrons, fills as 2, 8, 18, 1 — here the M shell does reach 18, and the lone N-shell electron is the one valence electron that makes copper the conductor of choice for aircraft wiring.

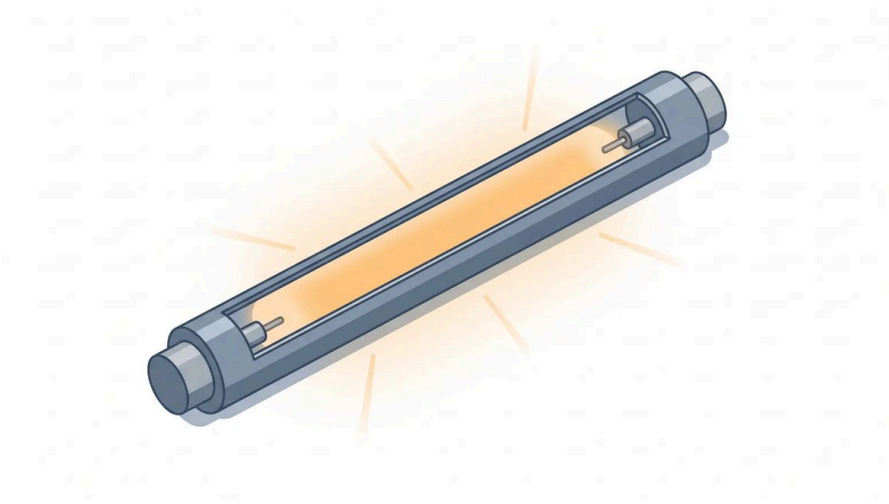
Element	Atomic number (Z)	Electrons per shell (K, L, M, N)	Valence electrons	Bulk electrical behaviour
Hydrogen	1	1	1	Insulating gas; bonds as $\mathrm{H_2}$
Helium	2	2	2 (K shell full)	Inert insulating gas; forms no bonds
Oxygen	8	2, 6	6	Insulating gas; bonds as $\mathrm{O_2}$
Aluminium	13	2, 8, 3	3	Conductor (power feeders, busbars)
Silicon	14	2, 8, 4	4	Semiconductor
Chlorine	17	2, 8, 7	7	Non-conductor; readily gains one electron
Copper	29	2, 8, 18, 1	1	Excellent conductor
Germanium	32	2, 8, 18, 4	4	Semiconductor

Read down the valence column and much of what follows in this note is already visible: among the solid elements, one to three valence electrons gives a metal, four gives a semiconductor, and a full or nearly full outer shell — most often seven or eight valence electrons — gives an insulator. Aluminium is worth noting particularly, because it is a genuine conductor carrying three valence electrons — the boundary is not at one or two. The two gases at the head of the table are a useful reminder that the rule describes how electrons are bound in the bulk material rather than in the isolated atom: hydrogen's single electron is not free to wander, because in $\mathrm{H_2}$ it is tied up in a shared pair, and helium's shell is complete already.

Energy Levels, Excitation and Ionisation

Calling the shells "energy levels" is more than a label. An electron may occupy an allowed level but never the space between two of them, so it cannot drift outward gradually; it must absorb a specific quantum of energy that exactly matches the step and jump. An electron pushed to a higher shell this way is said to be **excited**, and it falls back after a short time, releasing the same energy again — often as light, which is exactly how a

gas-discharge lamp works. Supply more energy still and the electron leaves the atom's influence altogether. That is **ionisation**, and it is the point at which an electron becomes available to carry current.



A gas-discharge lamp: excited electrons in the gas fall back to their shells and release the energy as light.

Two other familiar light sources are worth placing against that one. A light-emitting diode works on the same principle one step removed: the electron drops across the forbidden gap of a semiconductor rather than between the shells of an isolated atom, and the size of that gap fixes the colour. An incandescent filament does *not* work this way at all. It emits because it is a solid heated until it glows, and a hot solid radiates a continuous thermal spectrum — which is why a filament reddens as it is dimmed and why its light contains no spectral lines, while a gas-discharge lamp shows the sharp lines of its own atoms' energy steps. That fits what the band model later in this note says about solids: pack atoms together into a metal and their outer levels merge into continuous bands, leaving no discrete shell steps to jump between.

Energies this small are quoted in **electron-volts** rather than joules. One electron-volt is the energy gained by one electron moving through a potential difference of one volt, which numerically is the elementary charge expressed in joules.

The Electron-Volt

$$1\ \mathrm{eV} = 1.602 \times 10^{-19}\ \mathrm{J}$$

Keep this unit in mind: the band-gap figures quoted later for semiconductors and insulators are in electron-volts, and they only become meaningful once they can be compared with the energy actually available inside a material at working temperature.

Molecules, Compounds and Chemical Bonding

Isolated atoms are rare. Almost every material an engineer meets is atoms joined to other atoms, and the *way* they are joined decides where the charge sits and whether any of it is free to move. Bonding is therefore not a chemistry digression — it is the direct explanation of why copper conducts, why PVC does not, and why dry salt insulates while salt water does not.

Elements, Compounds and Mixtures

- **Element** — a substance built from one kind of atom only, identified by its atomic number. Copper, aluminium, silicon and oxygen are elements.
- **Compound** — two or more *different* elements chemically bonded in a fixed proportion. The compound's properties bear no relation to those of its ingredients: sodium is a metal that reacts violently with water and chlorine is a poisonous gas, yet sodium chloride is common salt.
- **Mixture** — substances present together but not chemically bonded, in any proportion, and separable by physical means. Air is a mixture of gases; brass is a mixture (an alloy) of copper and zinc.

A **molecule** is the smallest particle of a substance that can exist on its own and still show that substance's chemical properties. It is built from two or more atoms bonded together, which makes it larger than an atom and vastly larger than an electron. Where those atoms are all the same kind the result is a molecule of an element — oxygen travels as $\mathrm{O_2}$, nitrogen as $\mathrm{N_2}$, hydrogen as $\mathrm{H_2}$. Where they are of different kinds the result is a molecule of a compound, such as water $\mathrm{H_2O}$ or carbon dioxide $\mathrm{CO_2}$. The unit built from two or more *different* types of atom is therefore a molecule of a compound, and that distinction is a favourite of examiners.

⚠ **Read the noun — in the stem or in the options:** the smallest part of an **element** that keeps its characteristics is an **atom**; the smallest part of a **compound or substance** that can exist independently and keep its properties is a **molecule**. Both statements are true at the same time, and which of them is wanted is settled by whether the question pairs the definition with "element" or with "compound / substance". Sometimes that noun is in the stem, as in a question asking for the smallest particle a *substance* can be split into; just as often the stem is bare — "An atom is", "What is a molecule?" — and the discriminating noun sits in the options instead, so read those with the same care.

One refinement is worth knowing because it explains the ionic case below. Discrete molecules genuinely exist only where atoms are bonded covalently. Ionic compounds and metals are continuous lattices with no individual molecules in them, so their smallest repeating unit is properly called a formula unit. In Part-66 examinations, however, the expected answer for the smallest part of a compound or substance remains the molecule.

Why Atoms Bond at All

Bonding is driven entirely by the valence shell. A full outer shell — two electrons for the K shell, eight for any shell beyond it — is the stable, low-energy arrangement, which is why the noble gases such as helium neither bond nor conduct. Every other atom reaches that arrangement in one of three ways: by giving valence electrons away, by taking them, or by sharing them. Which of the three routes it takes is exactly what determines the material's electrical character.

Covalent Bonding – Electrons Shared

Two atoms that are both short of a full shell can each contribute electrons to a shared pair which then counts toward both shells at once. An oxygen atom has six valence electrons and needs two more; two oxygen atoms therefore share two pairs between them, so each atom sees an effective eight. That double bond is what holds an oxygen molecule together, and note precisely what is shared: **electrons**. Protons and neutrons are locked in their nuclei by the nuclear force and are never exchanged in any chemical or electrical process.

The consequence for charge distribution is that nothing is transferred and nothing is set loose. The shared pairs sit in fixed positions between the two nuclei, every electron is accounted for, and there are no spare carriers. A wholly covalent solid is therefore an insulator until something breaks a bond. Diamond is the extreme illustration: every carbon atom shares four pairs with four neighbours, the whole lattice is tied up, and the energy step needed to release an electron is above five electron-volts. It is also the standing exception to the valence-counting rule set out earlier: a solid element with four valence electrons that is an insulator rather than a semiconductor, because counting valence electrons only predicts the class, while the size of the gap decides it. Silicon uses the same four-neighbour arrangement, but its energy step is smaller by a factor of about five — and that single difference is the whole distinction between an insulator and a semiconductor.

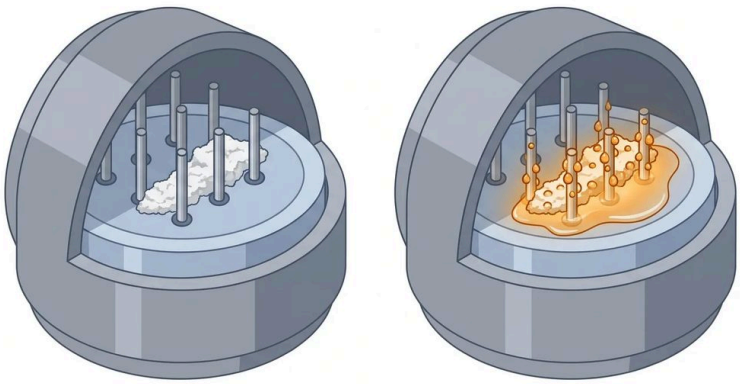
Ionic Bonding – Electrons Transferred

Where one atom has very few valence electrons and the other is only just short of a full shell, sharing is not the cheapest route — outright transfer is. Sodium fills as 2, 8, 1; shedding its single outer electron leaves the stable 2, 8 arrangement behind. Chlorine fills as 2, 8, 7 and completes itself by accepting exactly one. So sodium hands its valence electron to chlorine, and neither particle is neutral any more: the sodium is left with 11 protons and 10 electrons, the chlorine with 17 protons and 18 electrons. The bond is nothing more than the electrostatic attraction between those two opposite charges.

Here the charge really is separated, and it is locked into a rigid three-dimensional lattice of alternating positive and negative sites. There are no free electrons anywhere in that lattice, and the ions themselves cannot move, so **a dry ionic solid is an insulator**. Melt it, or dissolve it in water, and the lattice releases its ions; now the

charged particles themselves are mobile and the material conducts readily. This is conduction by *ion transport* rather than by electron transport, and it is the second of the two ways charge moves through matter.

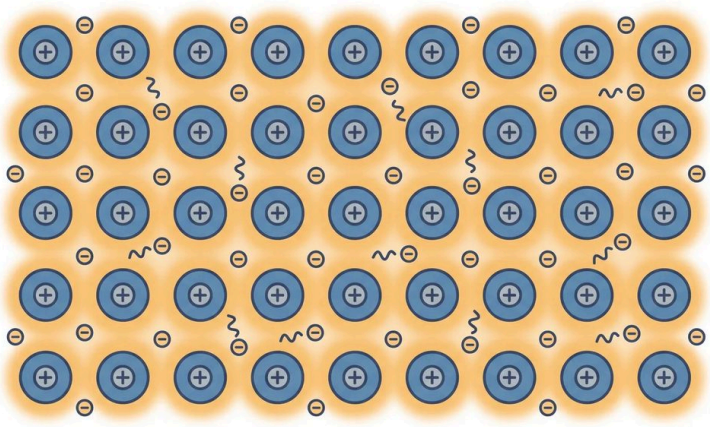
✈ **Aviation context:** this is why a dry salt deposit on a connector is electrically harmless while the same deposit wetted by condensation or a coastal atmosphere becomes a conducting bridge across the pins. It is why battery electrolyte conducts. And it is why an electrolytic corrosion cell needs moisture to work at all: without a liquid to carry the ions there is no circuit, so bonding, sealing and drainage are corrosion-control measures every bit as much as electrical ones.



The same salt deposit on a connector: dry it is harmless, wetted by condensation it bridges the pins.

Metallic Bonding – Electrons Delocalised

Metals take a third route. Their valence electrons are not paired off with one particular neighbour but surrendered to the lattice as a whole, so what remains is a regular array of positive ion cores immersed in a cloud of electrons that belong to no single atom. The bond *is* the mutual attraction between those fixed positive cores and that mobile negative cloud, which is why it holds the metal together perfectly well even though the electrons themselves are wandering.



A metal: fixed positive ion cores in a cloud of delocalised electrons that belong to no single atom.

Three practical consequences follow, and all three matter on an aircraft. First, the charge carriers are already free before any voltage is applied, so conduction needs no bond to be broken and works at any temperature. Second, the bond has no preferred direction, so planes of ions can slide past one another without it being destroyed — that is why copper can be drawn into wire, bent around a former and crimped without fracturing, while an ionic or covalent solid of the same size would shatter. Third, the same mobile electrons that carry current also carry heat, so good electrical conductors are also good thermal conductors, and a heavy feeder can act as a heat path into the equipment at either end of it.

Polar Molecules and Dipoles

When a covalent bond joins two *different* elements, the shared pair is rarely shared fairly. The atom with the stronger pull on electrons keeps the pair closer to itself, so although the molecule remains neutral overall, one end of it carries a small negative charge and the other an equal small positive charge. A molecule with this permanent internal charge separation is a **polar molecule**, and the separated pair of charges is an **electric dipole**. Water is the standard example: its bent shape prevents the two bond dipoles from cancelling, leaving the oxygen end negative and the hydrogen end positive.

Three electrical consequences follow. Polar molecules can pull an ionic lattice apart, each ion being surrounded by water molecules turning their opposite ends toward it — that is precisely how water dissolves salt and manufactures an electrolyte out of two insulators. Polar insulating materials also absorb moisture more readily than non-polar ones, so their insulation resistance falls when damp and recovers when dried. And although an applied electric field cannot free electrons inside an insulator at any strength short of breakdown, it *can* rotate the dipoles and distort the electron clouds, storing energy in that displacement; the effect is called polarisation, and it is why placing a dielectric between two plates lets them hold more charge at the same voltage. Under alternating voltage those dipoles have to reverse every cycle, and the internal friction of doing so appears as heat in the insulation, growing as the frequency rises.

The fluoropolymers used for aircraft wire insulation are chosen partly on this basis. Their molecules are highly symmetric, so the individual bond dipoles cancel and the net molecular dipole is close to zero. The results are a low dielectric constant, very low dielectric loss and little moisture uptake — exactly what is wanted where wiring runs alongside sensitive signal lines, coaxial feeders and high-frequency equipment.

Where the Charge Actually Sits

The syllabus asks for the structure and distribution of electrical charge in atoms, molecules, ions and compounds. Collecting the cases above gives the complete answer, and the final column is the one that decides whether the material can carry a current.

Structure	Positive charge	Negative charge	Net charge	Mobile carriers
Neutral atom	Protons, concentrated in the tiny nucleus	Electrons, spread through the shells	Zero	None until an electron is freed
Molecule of an element, e.g. O_2	Nuclei, symmetrically placed	Shared pairs, symmetrically placed	Zero	None
Polar molecule of a compound, e.g. H_2O	Hydrogen end slightly positive	Oxygen end slightly negative	Zero overall, but separated inside the molecule	None; the dipole can only rotate
Ion	Protons, unchanged in the nucleus	Electrons, now too few or too many to balance them	Positive if electrons were lost, negative if they were gained	The whole ion, wherever it is free to move
Ionic compound, solid	Positive ion sites in the lattice	Negative ion sites in the lattice	Zero	None — the lattice is rigid
Ionic compound, molten or dissolved	Free positive ions	Free negative ions	Zero	Both, drifting in opposite directions
Metal	Fixed positive ion cores	Delocalised electron cloud	Zero	Electrons, throughout the material

Ions and Ionisation

An atom is electrically neutral only while its electron count exactly matches its proton count. Disturb that balance by adding or removing one or more electrons and the atom is left with a net charge. A charged atom is called an **ion**, and the process of creating one is **ionisation**.

- **Positive ion (cation)** — the atom has *lost* one or more electrons. Protons now outnumber electrons, so the surplus positive charge of the nucleus is no longer cancelled and the net charge is positive.
- **Negative ion (anion)** — the atom has *gained* one or more electrons. Electrons now outnumber protons and the net charge is negative.

Note carefully what does *not* change. The protons stay locked in the nucleus throughout; ordinary ionisation moves electrons and nothing else. An ionised copper atom is still copper, because its atomic number is untouched — only its charge has altered. Removing a proton would change the element itself, but that is a nuclear process and has nothing to do with electricity.

Worked Example

An aluminium atom (Z = 13, filling 2, 8, 3) gives up all three of its valence electrons. What charge does it carry, and is it still aluminium?

It now has 13 protons and 10 electrons, a deficit of three electrons:

$$Q = 3 \times 1.602 \times 10^{-19} = +4.806 \times 10^{-19}\ \text{C}$$

It is a positive ion carrying three elementary charges. Its 13 protons are untouched, so it is still an aluminium atom — an ionised one. A chlorine atom that accepts a single electron becomes a negative ion carrying -1.602×10^{-19} C by exactly the same reasoning.

⚠ **Get the direction right: losing** electrons produces a **positive** ion; **gaining** electrons produces a **negative** ion. The intuitive trap is to link "gaining something" with "becoming positive". What is being gained is negative charge, so gaining makes the atom more negative, not less.

Ionisation Energy

The **ionisation energy** of an atom is the energy needed to strip away its most loosely held electron. That electron is always in the outermost shell, where the nuclear attraction has been weakened both by distance and by the screening effect of the inner shells, so it is by far the cheapest one to remove. Across any row of the periodic table the ionisation energy rises as the valence shell fills: an atom with one or two valence electrons parts with them easily, while one with a full or nearly full outer shell holds on tightly and will not release an electron under ordinary conditions at all.

That ordering is the same ordering as electrical conductivity, and it is not a coincidence — both are measuring how firmly the outer electrons are bound. "Few valence electrons", "low ionisation energy" and "good conductor" are three descriptions of a single physical property. Watch the wording carefully, though, because one phrase carries two meanings. A deficiency of electrons *in the outer shell* is exactly this property, and it describes a material of low resistance. A deficiency of electrons in the atom as a whole is something quite

different: it means the atom has been ionised and now carries a positive charge, and resistance is a property of a material rather than of a single ion.

How Atoms Are Ionised in Practice

Anything that can deliver the ionisation energy to an outer electron will free it. Five mechanisms account for essentially everything an engineer meets.

Mechanism	What happens	Where it is met on an aircraft
Heat	Thermal agitation of the lattice or the gas shakes electrons free	Thermionic emission from a heated cathode; rising leakage through hot-zone insulation
Electric field	The field itself pulls electrons out of their bonds	Dielectric breakdown, arcing across a contaminated connector, corona on high-voltage leads
Radiation	Photons or energetic particles knock electrons out of atoms	Cosmic and solar particles ionising air and depositing charge in avionic semiconductors at altitude; ultraviolet ageing of exposed insulation
Friction or contact separation	Electrons are transferred from one surface to the other as they part	Precipitation static in flight, refuelling hazards, electrostatic discharge damage to line-replaceable units
Chemical action	A reaction drives electrons onto one electrode and removes them from the other	Every battery cell

In every one of these, charge is never created — it is only separated. If one body ends up with a surplus of electrons, some other body has an exactly equal deficit, and the two together still total zero. That is the conservation of charge, and it is why a static problem on an aircraft is always a question of where the charge went rather than where it came from.

Conduction by Ions

In a metal only electrons move; the positive ion cores are pinned in the lattice and go nowhere. In a liquid or a gas the picture changes completely, because whole ions are free to travel, and carriers of both signs move at the same time in opposite directions. Positive ions migrate toward the **cathode** and are called cations; negative ions migrate toward the **anode** and are called anions — that migration is precisely what the two names mean. In an electrolysis cell driven by an external supply, the cathode is the electrode connected to the negative side and the anode the one connected to the positive side, so there the cations travel toward the negative electrode and the anions toward the positive electrode. Because the two sets of carriers have opposite charges as well as opposite directions of travel, both movements carry charge the same way round the circuit and the contributions add rather than cancel.

A battery is the everyday case. The external circuit carries electrons through the aircraft wiring while the electrolyte carries ions between the plates, and only the two together make a complete circuit. The same ion movement drives electroplating, anodising and electrolytic corrosion.

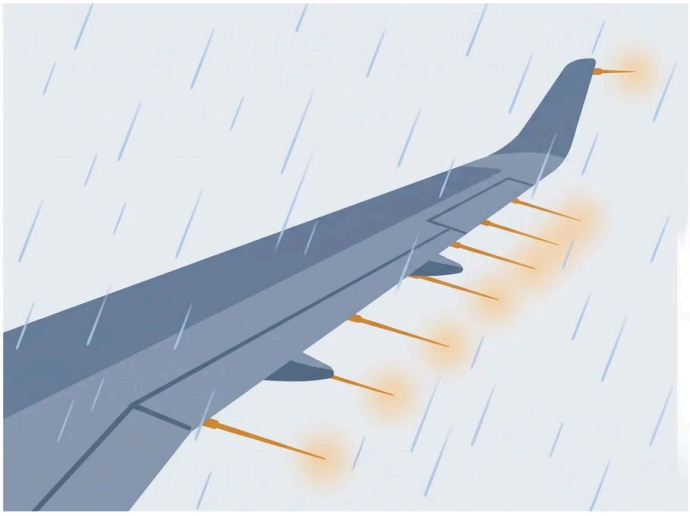
✈ **Aviation context:** lead-acid and nickel-cadmium cells depend entirely on ion transport through their electrolyte, so anything that hinders that transport reduces the current the battery can deliver — low temperature, low electrolyte level and low specific gravity all show up as poor cranking or a failed capacity check. The same three ingredients that make a cell work also make a corrosion cell work: two dissimilar metals, an electrolyte, and a conducting path between them.

Ionisation of Air

Dry air is an excellent insulator, but only up to a threshold — and it is a comparatively low one, because air breaks down at roughly 3 kV per millimetre, well below the figure for any of the solid insulating materials used on an aircraft. Raise the electric field high enough and it strips electrons from air molecules directly; each freed electron is then accelerated by that same field, collides with further molecules and frees more electrons still. The process feeds itself, which is why breakdown is sudden rather than gradual: the air converts from insulator to conductor almost instantly and the result is an arc.

Several familiar requirements follow from this. Creepage and clearance distances exist to keep the field below the threshold. Contamination and moisture lower the threshold, so a dirty or damp connector arcs at a voltage a clean dry one holds comfortably — which is why cleanliness is an electrical requirement and not merely a cosmetic one. Sharp points concentrate the field and start ionising first, producing corona. And an airframe flying through precipitation accumulates charge which will eventually leave from its extremities.

Static dischargers, the small wicks fitted at trailing edges and wing tips, exist to control that departure. Each provides a deliberately high-resistance path terminating in fine points where the field is intense, so the accumulated charge bleeds away continuously and quietly as a controlled corona. Without them the charge builds until it jumps as a spark across the structure, and the broadband noise from those sparks floods the radio and navigation receivers.



Static dischargers at the trailing edge and wing tip bleed built-up charge away as a controlled corona.

Valence and Free Electrons

When a valence electron gains enough energy, it can break free from its parent atom and move through the material. Such electrons are called **free electrons**. The ease with which this happens determines whether a material is a conductor, semiconductor, or insulator.

The **energy band model** explains this:

- **Valence band** — the energy range occupied by valence electrons bound to atoms.
- **Conduction band** — the energy range where electrons are free to move and carry current.
- **Band gap (forbidden gap)** — the energy difference between the valence and conduction bands.

Why Bands Exist at All

A single isolated atom has sharp, well-separated energy levels — the shells of the previous sections. Pack atoms together into a solid at a density of the order of 10^{28} per cubic metre and every atom's outer levels are disturbed by all of its neighbours. No two electrons in the assembly may occupy the same state, so each original level is forced to split into an enormous number of very slightly different ones. Crowded that closely, they stop behaving as separate levels and act instead as a continuous *band* of allowed energies. Between two such bands lies a span of energies that no electron in the solid is permitted to have, and that span is the forbidden gap.

Why the Gap Decides Everything

An electron can only contribute to a current if there is an empty state close by for it to move into. This is the deeper reason behind the familiar three-way division of materials, and it goes further than counting valence electrons:

- In a metal the valence and conduction bands overlap, so empty states sit immediately above the occupied ones. Even a vanishingly small applied field produces a net drift, no bond has to be broken first, and conduction happens at any temperature.
- In a semiconductor or an insulator the valence band is completely full. A full band can carry no current at all, however strong the field, because there is simply nowhere for an electron to move to. Current becomes possible only when an electron is lifted right across the gap into the empty conduction band above it.

So the question that sorts materials is not how large the gap is in isolation, but how large it is *compared with the energy actually available*. At room temperature the characteristic thermal energy in a material is only about 0.026 eV.

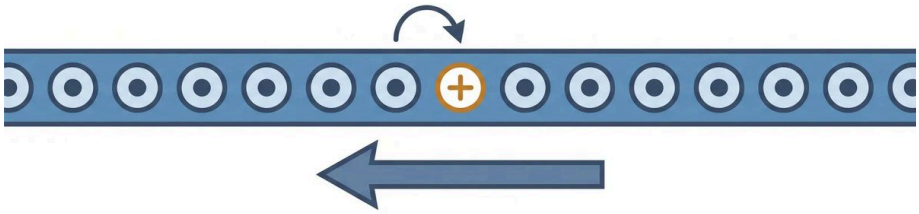
Thermal Energy at Room Temperature

$$kT \approx 0.026\ \text{eV} \quad \text{at } T = 300\ \text{K}$$

Set that against the gaps. Silicon's is about 1.1 eV, roughly forty times the thermal energy available, so only the small fraction of electrons in the high-energy tail of the distribution ever manage the jump — but the fraction is not zero, which is why a semiconductor conducts a little. An insulator's gap is above 5 eV, roughly two hundred times the thermal energy available, and the fraction able to cross is smaller again by many orders of magnitude — effectively none. One number, compared against one other number, sorts every material into its class.

Electrons and Holes

When a bond does break, **two** carriers appear rather than one: the electron that escaped, and the vacancy it left behind. That vacancy, called a **hole**, is not merely an absence. A valence electron in a neighbouring bond can hop sideways into it, which fills the original vacancy but opens a new one where the neighbour came from. The vacancy therefore travels through the lattice in the opposite direction to the hopping electrons, and for circuit purposes it behaves exactly like a mobile particle carrying one positive elementary charge.



A neighbouring electron hops into the hole, so the vacancy travels the opposite way to the hopping electrons.

Under an applied field the free electrons and the holes drift in opposite directions, but because their charges are also opposite, both contribute current the same way round the circuit. Holes drift along the electric field, in the same sense as conventional current; free electrons drift against it.

How Many Free Electrons Are There?

The numbers are worth carrying, because they show that the difference between a conductor and a semiconductor is one of quantity on a staggering scale rather than one of kind. A monovalent metal such as copper contributes about one free electron per atom, which works out at 8.5×10^{28} per cubic metre; a trivalent metal such as aluminium contributes three, giving roughly 1.8×10^{29} per cubic metre, about twice copper's figure. Either way the count is of the order of 10^{28} to 10^{29} free electrons per cubic metre, and that population is present whether or not any voltage is applied. Pure silicon at room temperature manages of the order of 10^{16} per cubic metre. Between the two lies a factor of some twelve to thirteen orders of magnitude. It is that difference in the *number* of carriers, rather than any difference in how freely an individual carrier moves, that accounts for the enormous gap between copper's resistivity and silicon's. Within the metals themselves the balance is different: having more carriers does not by itself make the better conductor, and aluminium's resistivity is still higher than copper's, because conductivity depends on how far a carrier drifts between collisions as well as on how many carriers there are, and aluminium's electrons are scattered more heavily. Note 3.7, Resistance/Resistor, works that comparison through.

It also explains why doping is so effective. A donor or acceptor atom introduces an energy level lying only a few hundredths of an electron-volt from the nearest band, which is comparable with the thermal energy already available at room temperature. Its carrier therefore does not have to cross the full gap and is released essentially every time, so an impurity of a few atoms in a million can multiply the carrier population by orders of magnitude and set the material's conductivity deliberately and precisely.

How Fast Do Free Electrons Actually Move?

Two different speeds are involved here and confusing them is a classic error. Free electrons in a metal are already moving very fast at random, at speeds of the order of 10^6 m/s, in every direction at once — and that motion transports no charge anywhere, because for every electron heading one way there is another heading back. Applying a field superimposes on that chaos a small average velocity along the field, called the **drift velocity**. The drift is what constitutes the current, and it is astonishingly slow.

Drift Velocity

$$v_d = \frac{I}{nAe}$$

where n is the free electrons per cubic metre, A the cross-sectional area and e the elementary charge.

Worked Example

A copper conductor of 1 mm² cross-section carries 1 A. Copper has about 8.5×10^{28} free electrons per cubic metre. How fast do the electrons drift?

$A = 1\ \text{mm}^2 = 1 \times 10^{-6}\ \text{m}^2$

$$v_d = \frac{1}{(8.5 \times 10^{28})(1 \times 10^{-6})(1.602 \times 10^{-19})} = \frac{1}{1.36 \times 10^{4}} \approx 7.3 \times 10^{-5}\ \text{m/s}$$

That is about 0.073 mm per second — roughly a quarter of a metre in an hour. At this current, an electron entering one end of a long harness run would still be well short of the far end many hours later.

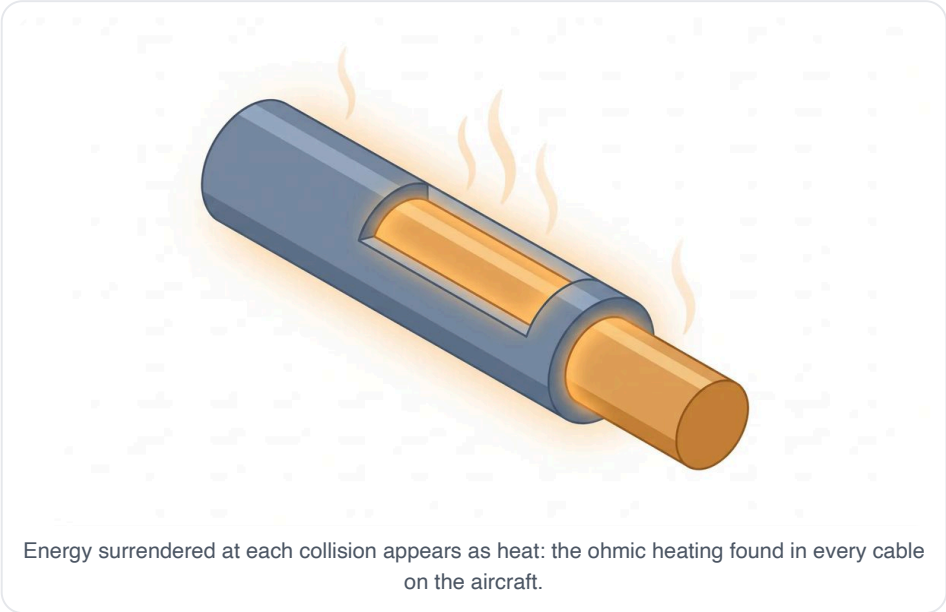
Yet a lamp lights the instant the switch closes. The resolution is that what travels along the conductor at close to the speed of light is not the electrons but the electric field that pushes them. Every free electron along the entire length of the wire begins to drift almost simultaneously, and the electron that lights the lamp was already sitting in the filament before the switch was touched. Two practical points follow: a fault is felt everywhere in a circuit effectively at once, and the question "how fast does electricity travel?" is a question about field propagation, not about electron speed.

Which Way the Free Electrons Go

The direction follows from nothing more than repulsion and attraction. The negative terminal of a source is a region carrying a surplus of electrons, and like charges repel, so it pushes free electrons in the conductor away from itself. The positive terminal has an electron deficit, and unlike charges attract, so it pulls them toward itself. Work the concrete case: an electron sitting midway along a copper wire strapped across a battery feels a push from the crowded negative plate and a pull from the depleted positive plate, and both act in the same sense. **Free electrons in the external circuit therefore drift away from the negative terminal and toward the positive one.** The two naming conventions built on this physical fact are set out in the final section of this note.

What Slows Them Down

A free electron does not accelerate indefinitely under the field. It picks up energy, then collides with a vibrating lattice ion, an impurity atom or a defect in the crystal, and surrenders the energy it gained as heat. Repeated many times over, that gain-and-collide cycle produces a steady average drift rather than continuous acceleration, and the energy handed over at each collision is the ohmic heating found in every cable, joint and winding on the aircraft.



Energy surrendered at each collision appears as heat: the ohmic heating found in every cable on the aircraft.

Resistance, at this level of explanation, simply *is* the rate at which drifting electrons are scattered. Stating it that way makes three otherwise separate facts obvious at once: resistance depends on temperature, because heat controls how violently the lattice vibrates; it depends on purity, because foreign atoms are extra scattering centres, which is why alloys have higher resistivity than the pure metals they are made from; and it depends on mechanical condition, because work-hardening, corrosion products and a poorly made crimp all add scattering exactly where the current is most concentrated.

Effect of Temperature on Conduction

Heating a material changes two things at once, and the two push in opposite directions. Which of them dominates is what decides whether a material's resistance goes up or down when it gets hot, and it is a different answer for each of the three classes of material.

- **Scattering.** A hotter lattice vibrates more violently, so a drifting electron collides more often and travels a shorter distance between collisions. Less drift results for the same applied field. Acting alone, this effect **raises** resistance.

- **Carrier generation.** A hotter material also has more thermal energy on hand, so more electrons can be lifted across the forbidden gap. More carriers means more current for the same applied field. Acting alone, this effect **lowers** resistance.

Metals: Scattering Wins

In a metal the carrier population is effectively already at its ceiling. Every atom has already surrendered its valence electron or electrons before any heat is applied, the bands overlap so nothing has to be lifted across a gap, and warming the metal cannot produce any meaningful number of additional carriers. Only the scattering effect is left in play, so a metal's resistance **rises** as it gets hotter. Metals are said to have a positive temperature coefficient of resistance.

The magnitude is easy to underestimate. Copper's temperature coefficient is roughly 0.0039 per degree Celsius referred to 20 °C, so a winding measuring 10 ohms at 20 °C rises to $10\,(1 + 0.0039 \times 50) = 11.95$ ohms at 70 °C — nearly 20% more resistance from the same piece of wire. That is why a generator field or a motor winding checked cold reads noticeably lower than the same winding measured immediately after a run, and why a resistance figure quoted without a temperature is of limited use. Note 3.7, Resistance/Resistor, carries the temperature coefficient further, including the trap that the coefficient itself depends on the reference temperature it is quoted against.

There is a feedback loop hidden in this. A cable already warm from its load current has a higher resistance, so at the same current it dissipates more power, so it warms further, so its resistance rises again. The loop settles at a higher temperature than a naive calculation predicts, and it is one of the reasons wiring current ratings are de-rated for high ambient temperature and again for large bundles where the heat generated in the core of the bundle has no easy escape path.

Semiconductors: Carrier Generation Wins

In a semiconductor the carrier population at room temperature is minute, and it grows very steeply indeed with temperature, because the number of electrons able to cross the gap depends exponentially on it. That gain completely overwhelms the extra scattering, so a semiconductor's resistance **falls** as it gets hotter: a negative temperature coefficient.

This behaviour is exploited deliberately. A negative-temperature-coefficient thermistor is a semiconductor whose resistance drops predictably as it warms, which makes it useful both for temperature sensing and for limiting the inrush current to equipment at switch-on. A positive-temperature-coefficient device behaves the opposite way, its resistance climbing sharply once it self-heats past a threshold, which makes it a resettable overcurrent protector.

Aviation context: the same effect is a hazard as well as a tool. Leakage through a silicon junction climbs steeply with temperature — a common rule of thumb is that it roughly doubles for every 10 °C rise. That leakage dissipates power in the junction, which raises its temperature, which increases the leakage again. Unchecked, the loop ends in thermal runaway, and it is precisely why power semiconductors are heat-sinked and why manufacturers publish de-rating curves against case temperature.

Insulators: The Same Mechanism, From a Much Larger Gap

An insulator is not a fundamentally different kind of material in this respect — it behaves like a semiconductor with a very large gap. The same thermal generation operates, so its resistance falls as it gets hotter, for exactly the same reason a semiconductor's does. Because the dependence is exponential rather than linear, the drop per degree is far steeper than the near-linear rise a metal shows: the maintenance rule of thumb is that insulation resistance roughly **halves for every 10 °C rise**, which is about seven per cent per degree Celsius, against the 0.39 per cent per degree by which copper's resistance climbs. Exponential changes compound as well, so between a cold-soaked airframe and the same insulation sitting at engine-bay temperature the accumulated reduction in insulation resistance can be very large indeed.

Two practical consequences follow directly from that, and both are routine maintenance concerns rather than theoretical curiosities.

- **Insulation resistance readings only mean something alongside the conditions they were taken in.** The same harness measured hot reads lower than it does cold, so comparing an as-found hot reading against a cold baseline can suggest a defect that is not there — or, worse, hide a real one behind a favourable cold measurement. Where a trend is being followed over time, readings are corrected to a common reference temperature before they are compared.
- **The loop can run away here too.** Leakage current flowing through warm insulation dissipates power inside the insulation itself, which raises its temperature, which lowers its resistance, which increases the leakage. Contamination, absorbed moisture or mechanical damage concentrates the process at one spot, and the resulting hot track carbonises into a permanently conducting path.

This is why insulation failures so often appear sudden while having been developing for a long time. The final event is the electric field tearing electrons free from a material that has already been quietly degraded by heat, damp and contamination — the breakdown described at the end of the next section is the last step of that sequence, not the whole of it.



A scorched spot found where contamination and damp sat on a harness: the end of a long, slow degradation.

⚠ **Record the conditions with the reading:** an insulation resistance value taken on a hot harness, or in humid conditions, is not automatically evidence of a defect — and a good value taken cold and dry is not automatically a clean bill of health. Note the temperature and the humidity with every measurement, because without them the number cannot be compared with anything.

Conductors, Semiconductors, and Insulators

Property	Conductor	Semiconductor	Insulator
Valence electrons	1–3	4	5–8 (or full shell)
Band gap	None (bands overlap)	Small (≈ 0.7–1.1 eV)	Large (> 5 eV)
Conductivity	Very high	Moderate (variable)	Extremely low
Free electrons	Abundant at room temp	Few at room temp; increases with temp	Almost none
Temp effect on resistance	Increases with temp	Decreases with temp	Decreases markedly (roughly halves per 10 °C rise)
Examples	Copper, aluminium, silver, gold	Silicon, germanium	Glass, rubber, PVC, mica, ceramic

✈ **Aviation context:** Aircraft wiring predominantly uses **copper** (excellent conductor) and **aluminium** (lighter, used in power feeders). Wire insulation is typically **PTFE** or **ETFE** fluoropolymer, composite fluoropolymer/polyimide constructions, or **silicone rubber** in high-temperature zones; pure polyimide (**Kapton**) is still found on older aircraft but is no longer specified for new installations because of its susceptibility to arc tracking.

Conductors in Detail

Good conductors have atoms with only 1 to 3 valence electrons loosely bound to the nucleus. In metals, these electrons form a "sea" of free electrons that can drift through the material when a voltage is applied. Copper has one valence electron and is the most commonly used conductor in aviation.

Semiconductors in Detail

Silicon and germanium have 4 valence electrons and form covalent bonds with neighbouring atoms. At absolute zero they behave as insulators, but at room temperature some bonds break, releasing electron-hole pairs. Their conductivity can be dramatically altered by **doping** — adding tiny amounts of impurity atoms:

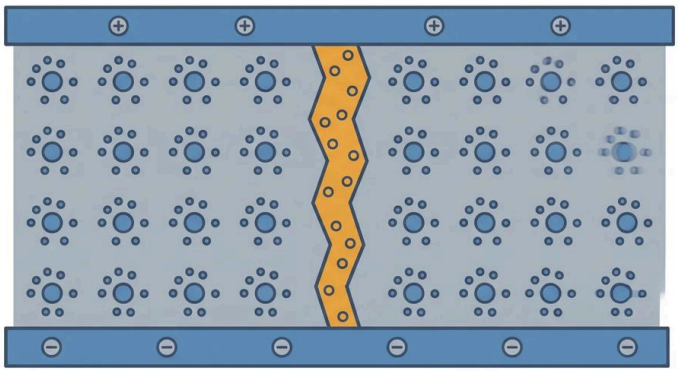
- **N-type** — doped with a pentavalent element (e.g. phosphorus), providing extra free electrons.
- **P-type** — doped with a trivalent element (e.g. boron), creating "holes" (positive charge carriers).

Aviation context: Semiconductor devices (diodes, transistors, integrated circuits) are the building blocks of every electronic system on a modern aircraft — from flight computers and engine controllers to cockpit displays and communication radios.

Insulators in Detail

Insulators have a full or nearly full valence shell, so electrons are tightly bound and the band gap is very large. Virtually no free electrons are available for conduction. Insulators are essential for preventing unwanted current flow and protecting personnel from electric shock.

Critical note: No insulator is perfect. Under extreme voltage, any insulator can experience **breakdown** — the electric field tears electrons free and the material suddenly conducts. This is why insulation must be rated well above the operating voltage and regularly tested.



Breakdown: under extreme voltage the field tears electrons free and the insulator suddenly conducts.

Conventional Flow vs. Electron Flow

There are two conventions for describing current direction:

Convention	Direction	Used By
Conventional flow	Positive (+) to negative (–)	Circuit diagrams, engineering standards, EASA exams
Electron flow	Negative (–) to positive (+)	Physics, understanding actual particle movement

In reality, electrons (the actual charge carriers in metals) flow from negative to positive. However, before the electron was discovered, current was defined as flowing from positive to negative. This **conventional current** direction is used in virtually all circuit analysis and is the standard in EASA Part 66 examinations.

⚠ **Exam tip:** Unless a question specifically asks about electron flow, always use **conventional current** direction (positive to negative) in your answers and circuit analysis.

Module 4 — Electronic Fundamentals

4.1.1 — Diodes

Diodes are the simplest semiconductor devices and form the building blocks of all modern electronics. This section covers the physics of semiconductor materials, how PN junctions work, and the wide variety of diode types used in aircraft electronic systems — from simple rectifiers to Zener voltage regulators, LEDs for cockpit lighting, and protective varistors.

Semiconductor Materials and Electron Configuration

All materials can be classified by their electrical conductivity into three groups:

Category	Resistivity	Examples	Valence Electrons
Conductors	Very low	Copper, aluminium, silver	1–3 (loosely held)
Semiconductors	Between conductors and insulators	Silicon (Si), germanium (Ge)	4
Insulators	Very high	Glass, rubber, ceramic	7–8 (tightly held)

Silicon is the most widely used semiconductor in modern electronics. Each silicon atom has 4 valence electrons in its outer shell and forms **covalent bonds** with 4 neighbouring atoms, creating a crystal lattice structure. In its pure (intrinsic) state, silicon is a poor conductor because almost all electrons are locked in bonds.

The key property that makes semiconductors useful is that their conductivity can be precisely controlled by adding tiny amounts of impurity atoms — a process called **doping**.

Energy Bands

In any solid, electrons occupy energy levels grouped into bands:

- **Valence band** — the highest energy band that is fully occupied at absolute zero. Electrons here are bound to atoms.
- **Conduction band** — the next higher band. Electrons here are free to move and carry current.
- **Band gap** — the energy gap between the valence and conduction bands. In semiconductors, this gap is small enough (about 1.1 eV for silicon) that thermal energy or an applied voltage can push electrons across it.

In conductors, the valence and conduction bands overlap (no gap). In insulators, the band gap is very large (5+ eV). In semiconductors, the small band gap gives them their unique, controllable behaviour.

Why the Valence Shell Decides Everything

Electrons orbit a nucleus in shells, and only the outermost occupied shell — the **valence shell** — takes part in conduction or chemical bonding. Electrons in the inner shells are held so tightly by the nucleus that no voltage an engineer will ever apply to a circuit can dislodge them. The whole of the table above is therefore a statement about one shell.

A valence shell is most stable when it holds eight electrons. That single fact explains all three rows:

- **Conductors carry one to three valence electrons.** Those few electrons are a long way from the stable count of eight and are weakly bound, so in a solid metal they detach from their parent atoms altogether and form a shared "sea" of free electrons drifting between fixed positive ion cores. Copper and silver contribute one each, aluminium three — which is why the range in the table is 1 to 3 and not a single number.
- **Insulators carry seven or eight.** Their shells are complete or nearly complete, so there is nothing loosely held to give up and nothing useful to gain. Every electron is locked into a bond.
- **Semiconductors carry exactly four** — the awkward middle. Four is equally far from emptying the shell and from filling it, so a silicon atom neither donates nor accepts: it *shares*.

Covalent Bonding and What Breaks a Bond

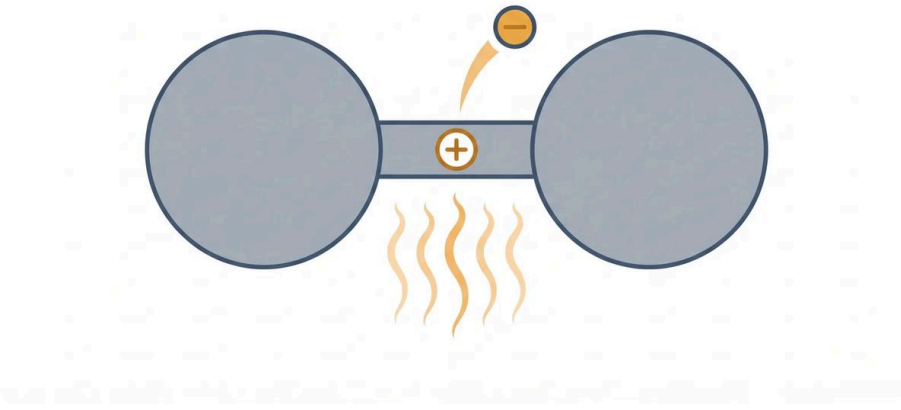
Each silicon atom shares one of its four valence electrons with each of four neighbours, and receives a share of one electron back from each. Every atom therefore sees eight electrons in its valence shell and every bond is a **shared pair**. The result is a rigid three-dimensional crystal in which no electron belongs to any single atom and none is free to wander. That is why a perfect silicon crystal at absolute zero is an insulator, not a conductor — the material does not become useful until something breaks a bond.

Three things can break one: **heat** (lattice vibration), **light** of sufficient photon energy, and a sufficiently strong **electric field**. All three appear later in this note as the operating principle of a real device — heat as the source of diode leakage current, light as the mechanism of the photodiode, and field strength as the mechanism of Zener breakdown.

Intrinsic Conduction and the Electron-Hole Pair

An **intrinsic** semiconductor is one that is chemically pure: no deliberate impurity has been added. When thermal energy breaks a covalent bond in such a crystal, two things are created at once and they are always

created together:



Heat breaks a bond and creates two carriers at once: a free electron and the hole it leaves behind.

- a **free electron**, which has escaped the bond and can move through the crystal, and
- the **hole** it left behind — the vacancy in the broken bond, which behaves as a mobile positive charge.

Because they are created in pairs, an intrinsic semiconductor contains **exactly as many free electrons as holes**. This is the defining electrical property of intrinsic material and it is the one an examiner most often asks for. Notice also that breaking a bond creates no net charge: the electron carries away a negative charge and leaves an equal positive charge behind. A block of pure silicon or pure germanium is therefore **electrically neutral** no matter how hot it is, and remains so — it has neither an excess nor a deficit of electrons overall.

The reverse process runs at the same time. A wandering free electron that meets a hole falls into it, re-forming the bond and destroying both carriers. This is **recombination**, and the average time a carrier survives before it happens is the **carrier lifetime**. At any fixed temperature the generation rate and the recombination rate settle into balance, and the carrier concentration is steady. Raise the temperature and generation outruns recombination until a new, higher balance is reached.

How a Hole Actually Moves

A hole is not a particle. Nothing physically travels when a hole moves — what happens is that a valence electron from a *neighbouring* bond hops sideways into the vacancy, filling it and leaving a new vacancy where it came from. Repeat that along a row of atoms and the vacancy appears to travel in the direction opposite to the hopping electrons. Since the electrons hop toward the positive terminal, the vacancy migrates toward the negative terminal, exactly as a positive charge would. Treating it as a positive carrier is therefore not a convenient fiction but a correct description of the observed current.

One practical consequence follows from the mechanism. A free electron in the conduction band travels through open space between the atoms; a hole advances only as fast as valence electrons can hop from bond to bond. Holes are therefore markedly **less mobile** than electrons in the same material — roughly a third as mobile in silicon. That difference is why N-type material of a given doping level has a lower resistivity than P-type material of the same doping level, and why the fastest semiconductor devices are built to conduct with electrons rather than holes.

Drift and Diffusion: the Two Ways Charge Moves

Almost every current in every device in this note is one of two mechanisms, and telling them apart is essential to understanding the PN junction later:

- **Drift** is motion caused by an **electric field**. Apply a voltage and free electrons drift toward the positive terminal, holes toward the negative. Drift needs a field; remove the field and drift stops.
- **Diffusion** is motion caused by a **concentration difference**, and needs no field at all. Carriers spread out from where they are crowded toward where they are scarce, purely by random thermal motion, in the same way a drop of dye spreads through still water.

Diffusion is what forms the depletion region when P-type and N-type material are first joined. Drift is what opposes it and stops it. The equilibrium between the two is the barrier potential.

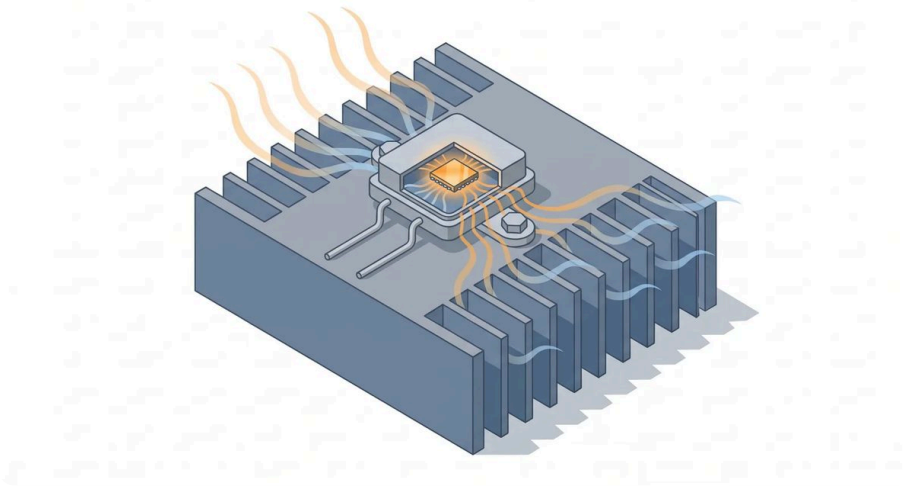
Temperature: Why a Semiconductor Behaves Backwards

A metal and a semiconductor respond to heat in *opposite* directions, and both mechanisms need naming because they compete inside real components:

- **A metal has a positive temperature coefficient.** Its free-electron population is fixed by its chemistry and does not grow with heat. Heating only makes the ion cores vibrate more violently, which scatters the drifting electrons more often. Resistance therefore **rises** with temperature.
- **A semiconductor has a negative temperature coefficient.** The same increase in scattering occurs, but it is swamped by a far larger effect: heat breaks additional covalent bonds and multiplies the number of carriers available. The carrier population wins by a wide margin, so resistance **falls** with temperature.

Work one case through to fix the direction. A germanium diode at 25°C might pass a leakage current of a few microamperes when reverse biased. Raise the junction to 75°C and the leakage climbs by roughly a factor of thirty, because that current is carried entirely by thermally generated pairs. The junction has become a better conductor by being made hotter — the opposite of what a copper wire does.

⚠ **Thermal runaway:** the negative coefficient is self-reinforcing and this is why semiconductors fail suddenly rather than gradually. A hot junction leaks more current; the extra current dissipates more power in the junction; the extra power makes it hotter still. Once the heat generated exceeds the heat the package can remove, the cycle accelerates to destruction in seconds. Adequate heatsinking and honest current derating are not refinements — they are what keeps the loop stable.



A heatsink carries junction heat away, keeping the thermal runaway loop of a semiconductor stable.

Silicon Compared with Germanium

Both elements have four valence electrons and both form the same diamond-cubic lattice, so they behave in the same way; they differ in how tightly the bonds are held. Germanium's smaller band gap makes it easier to break a bond, which is simultaneously its one advantage and the reason it was displaced:

Property	Silicon (Si)	Germanium (Ge)	Consequence
Band gap	About 1.1 eV	About 0.7 eV	Ge bonds break more readily at any given temperature
Junction barrier potential	0.6–0.7 V	0.2–0.3 V	Ge conducts on a smaller signal
Reverse leakage current	Very small (nanoamperes)	Far larger (microamperes)	Si blocks much more cleanly when off
Usable junction temperature	Substantially higher	Limited — leakage dominates while still well below Si's limit	Si survives hot equipment bays; Ge does not
Oxide	Forms a stable, insulating oxide	No comparable stable oxide	Si can be made into integrated circuits; Ge cannot easily

Property	Silicon (Si)	Germanium (Ge)	Consequence
Typical modern role	Rectifiers, all mainstream devices	Small-signal detection where 0.3 V matters	Ge survives in niches, not in power circuits

The comparison is worth holding on to because two exam answers fall straight out of it. A germanium diode is chosen when a signal is too small to overcome a silicon junction — radio detection and demodulation, and low-level rectification. A silicon diode is chosen everywhere else, and above all wherever the equipment gets hot.

Worked Example — How Little Impurity Doping Needs

Silicon contains roughly 5×10^{22} atoms per cubic centimetre. At room temperature, thermal generation alone gives an intrinsic carrier concentration of about 1.5×10^{10} free electrons per cubic centimetre (and the same number of holes). Suppose one silicon atom in every ten million is replaced by a donor impurity.

Donor atoms added: $5 \times 10^{22} \div 10^{7} = 5 \times 10^{15}$ per cubic centimetre.

Each donates one free electron, so the electron concentration becomes about 5×10^{15} per cubic centimetre.

Ratio to the intrinsic value: $(5 \times 10^{15}) \div (1.5 \times 10^{10}) \approx 3.3 \times 10^{5}$.

An impurity of one part in ten million multiplies the free-electron population by more than three hundred thousand times. Two things follow. First, the crystal purity required before doping is extreme — a stray contaminant at the same one-in-ten-million level would swamp the intended doping completely. Second, doping is a fine control, not a bulk additive: the impurity displaces so few atoms that the crystal structure and the mechanical properties are unchanged.

Exam vocabulary: intrinsic means pure — nothing added and nothing taken away. **Extrinsic** means deliberately doped. A question that offers "a pure material", "one with added elements" and "one with removed elements" is testing that one word. Note also that removing elements is not a category at all: you cannot dope a semiconductor by subtraction.

Putting Numbers on the Three Categories

The table at the head of this section describes the three groups qualitatively; **resistivity** is the quantity that makes the description measurable. Resistivity is the resistance of a standard cube of the material, quoted in ohm-metres, and it is a property of the substance alone — independent of the size or shape of the piece in front of you. Its usefulness here is the sheer span it covers:

- **Conductors** sit around 10^{-8} ohm-metres. Copper, silver and aluminium are all within a factor of two of each other.
- **Insulators** reach 10^{12} ohm-metres and beyond.

- **Semiconductors** occupy the enormous gap between, and — this is the point — a single semiconductor can be placed almost anywhere in that gap by the choice of doping.

Twenty orders of magnitude separate the extremes, which is why the middle category cannot be defined by any single number and is defined instead by its position between the other two. **Conductivity** is simply the reciprocal of resistivity and is used when it is more natural to talk about how well something conducts than how much it resists; the two carry exactly the same information.

Compound Semiconductors

Silicon and germanium are elements, but many of the devices later in this note are built from **compounds** — combinations of elements from either side of group 4 that average out to four valence electrons per atom. Gallium has three and arsenic has five, so gallium arsenide behaves as a semiconductor in the same way silicon does. Compounds are used because two properties can be engineered that an element does not offer:

- **The band gap can be chosen.** Mixing the proportions of a ternary compound moves the gap continuously, and since a light-emitting junction's photon energy equals its band gap, this is how an LED's colour is selected at the design stage.
- **The recombination can be made to produce light at all.** In silicon, an electron falling from the conduction band to the valence band gives up its energy as lattice vibration — heat — rather than as a photon. Compounds such as gallium arsenide and indium gallium nitride release it as light instead, which is the whole reason there is no such thing as a silicon LED.

A separate group of compounds is chosen for the opposite reason: a band gap much *wider* than silicon's. Silicon carbide and gallium nitride break bonds far less readily, so they leak less, withstand a higher electric field before breaking down, and remain usable at junction temperatures that would destroy a silicon device. They are the materials behind the high-temperature, high-voltage devices appearing in modern power conversion.

Two Components That Are Nothing But the Temperature Effect

The negative temperature coefficient described above is usually something to be designed around. Two components exist to exploit it directly, and both are found in aircraft equipment:

- **The NTC thermistor** is a semiconductor whose resistance falls steeply and predictably as it warms. Used as a temperature sensor it gives a much larger change per degree than a metallic sensor does, at the cost of a curved rather than a straight response. Used as an **inrush limiter** it is placed in series with a supply: cold at switch-on, its resistance is high and it holds down the surge that charges the filter capacitors; once current has been flowing for a second or two its own dissipation warms it, its resistance falls, and it stops

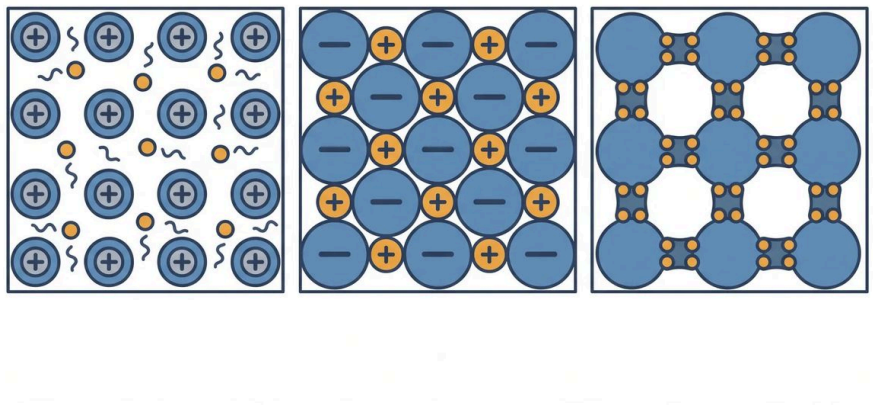
wasting power. One component solves the surge problem described in the rectifier section, with no control circuit at all.

- **The PTC thermistor** is engineered the other way, so that its resistance rises sharply above a threshold temperature. In series with a load it acts as a self-resetting fuse: an overload heats it, its resistance climbs, the current falls to a trickle, and it recovers once the fault is cleared and it has cooled.

A third material property in the same family is **photoconductivity** — light, like heat, breaks bonds and creates carriers, so illuminating a semiconductor lowers its resistance. That is the basis of the light-dependent resistor, and it is why a bare semiconductor junction has to be packaged in an opaque case unless light response is wanted.

Three Ways Atoms Bond, and Why Only One of Them Gives a Semiconductor

The covalent bond described above is one of three ways the valence electrons of a solid can be arranged, and comparing all three explains the resistivity table rather than merely restating it:



Metallic, ionic and covalent bonds: a free electron sea, fixed ions, and shared electron pairs.

Bond	What the valence electrons do	Electrical behaviour	Example
Metallic	Leave their parent atoms entirely and form a shared sea moving freely between fixed positive ion cores	Conductor. Carriers are permanently available, so conduction needs no energy input at all	Copper, aluminium, silver
Ionic	Transfer completely from one atom to another, leaving oppositely charged ions held together by attraction	Insulator when solid, because every electron is bound to a specific ion and the ions themselves cannot move through the lattice	Common salt, many ceramics

Bond	What the valence electrons do	Electrical behaviour	Example
Covalent	Are **shared** between neighbouring atoms in pairs, belonging fully to neither	**Semiconductor.** Electrons are bound, but only moderately, so a modest energy input frees them	Silicon, germanium, diamond

The middle column is the whole explanation. A metal has carriers whether you want them or not, an ionic solid has none that can move, and a covalent crystal sits between with carriers that *can* be released on demand. Being controllable is what makes the third category useful, and it is the direct consequence of sharing rather than donating or keeping.

Note that covalent bonding alone is not sufficient: diamond is covalently bonded and is an excellent insulator, because carbon's bonds are so strong that its band gap is enormous. The useful semiconductors are the covalent crystals whose bond strength happens to fall in a narrow band — strong enough that they are not conductors, weak enough that ordinary thermal and electrical energies can reach across.

One further condition applies. All of this assumes a **single crystal**, in which the lattice runs unbroken from one face of the wafer to the other. A polycrystalline or amorphous sample of the same element has boundaries and disordered regions where the bonding pattern is interrupted, and those act as traps and recombination sites that ruin carrier lifetime. Devices are therefore built on single-crystal wafers grown for the purpose, and any mechanical damage that cracks the crystal — including excessive lead bending or thermal shock — degrades the device permanently rather than cosmetically.

P-Type and N-Type Materials

N-Type Semiconductor

When a **pentavalent** (5 valence electrons) impurity such as phosphorus (P), arsenic (As), or antimony (Sb) is added to silicon, four of its electrons bond with surrounding silicon atoms. The fifth electron is free to move and conduct current.

- The impurity atom is called a **donor** (it donates a free electron)
- **Majority carriers**: electrons (negative)
- **Minority carriers**: holes
- The material is electrically **neutral** overall (the donor atom has an equal positive charge in its nucleus)

P-Type Semiconductor

When a **trivalent** (3 valence electrons) impurity such as boron (B), gallium (Ga), or indium (In) is added to silicon, it can only form three bonds. The missing bond creates a **hole** — an absence of an electron that behaves as a positive charge carrier.

- The impurity atom is called an **acceptor** (it accepts an electron to fill the hole)
- **Majority carriers**: holes (positive)
- **Minority carriers**: electrons
- The material is electrically **neutral** overall

Remember: N-type has extra **n**egative carriers (electrons). P-type has extra **p**ositive carriers (holes). But both materials are electrically neutral — doping does not add charge, it adds charge *carriers*.

Choosing a Dopant: It Is Only the Group Number

Which type a doped crystal becomes is decided by one number — how many valence electrons the impurity atom brings to a lattice site that expects four. There is nothing else to remember, and the exam tests it in both directions: given the valency, name the type; given the type, name the valency.

Impurity valency	Name	Common elements	Effect at the lattice site	Resulting material
5 (pentavalent)	**Donor**	Phosphorus, arsenic, antimony	Four electrons bond; the fifth is surplus and becomes free	**N-type**
3 (trivalent)	**Acceptor**	Boron, gallium, indium, aluminium	Only three bonds can be made; one bond is left one electron short	**P-type**
4 (tetravalent)	None — it is the host	Silicon, germanium, carbon	Bonds perfectly; adds no carrier of either sign	Intrinsic (undoped)

Note that a tetravalent impurity does nothing useful: it fits the lattice too well. Doping works precisely because the impurity *does not fit* — it either brings one electron too many or one too few.

Why the Carrier Is Nearly Free: Ionisation Energy

A donor's fifth electron is not held by a covalent bond, because there is no fifth bond for it to occupy. It is retained only by the weak electrostatic attraction of its own nucleus, and the energy needed to release it

altogether is a small fraction of the energy needed to break a covalent bond — a few hundredths of an electron-volt against silicon's 1.1 eV. In band terms the donor sits on a narrow energy level just *below* the conduction band, and an acceptor level sits just *above* the valence band.

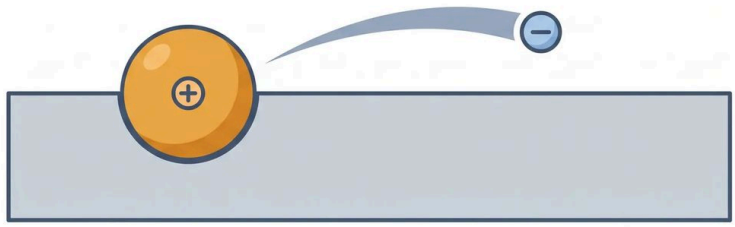
The practical consequence is important and is easy to miss: at ordinary temperatures the ambient thermal energy is already far more than enough, so **essentially every dopant atom is ionised all of the time**. The majority-carrier concentration is therefore set by the doping recipe and is effectively constant with temperature. That stability is exactly what the designer is buying, and it is what distinguishes doped material from the intrinsic crystal whose carrier count roughly doubles for each 10°C rise in temperature.

Neutral, But Not Symmetrical: Fixed Ions and Mobile Carriers

The statement that doped material is electrically neutral is exact, and understanding *why* is what makes the PN junction comprehensible rather than something to be memorised. Take an N-type crystal and account for the charge one donor atom at a time:

- Before ionisation the donor atom is neutral — five valence electrons balancing five protons.
- Once its fifth electron is released, the atom is left with one more proton than it has electrons, so it becomes a **positive ion**.
- That ion is **locked into the crystal lattice and cannot move**. It is not a charge carrier.
- The released electron is mobile and carries an equal and opposite negative charge.

Add the two together and the block is neutral. But the two charges are of a completely different nature: one is **fixed**, the other is **mobile**. Everything the junction does depends on being able to sweep the mobile charge out of a region and leave the fixed charge behind, uncovered. The same argument runs for P-type, with the signs exchanged: an acceptor that has captured an electron to complete its fourth bond becomes a **fixed negative ion**, and the hole it created is the mobile positive carrier.



A donor ion stays locked in the crystal while the electron it released is free to move.

Hold on to this, because it is the single most reversed fact in the subject: the immobile ion left behind in each region carries the *opposite* sign to that region's mobile majority carrier. N-type has mobile **negative** electrons and fixed **positive** donor ions. P-type has mobile **positive** holes and fixed **negative** acceptor ions. When the mobile carriers are later swept away from the junction, the charge that remains exposed on each side is therefore the opposite of what that side is named after.

Majority and Minority Carriers in Numbers

"Majority" and "minority" are not close-run descriptions — the two populations differ by many orders of magnitude, and they have different origins:

- **Majority carriers come from the doping.** Their number is fixed by manufacture and barely changes with temperature.
- **Minority carriers come from thermal generation only** — broken covalent bonds, exactly as in the intrinsic crystal. Their number therefore climbs steeply as the material gets hotter.

There is also a quiet second effect that catches people out. Doping a crystal N-type does not merely add electrons; it *reduces* the hole population below its intrinsic value. With so many more free electrons wandering the crystal, any hole that appears is found and filled far sooner, so the recombination rate for holes rises and the equilibrium hole concentration falls. The product of the two concentrations stays constant at a given temperature — raise one and the other must fall.

Worked Example — How Lopsided the Two Populations Are

Take the silicon from the previous section: intrinsic concentration about 1.5×10^{10} carriers of each kind per cubic centimetre, doped to give 5×10^{15} free electrons per cubic centimetre. The product of the electron and hole concentrations must stay at the intrinsic value squared.

Intrinsic product: $(1.5 \times 10^{10})^2 = 2.25 \times 10^{20}$

Hole concentration after doping: $(2.25 \times 10^{20}) \div (5 \times 10^{15}) = 4.5 \times 10^{4}$ per cubic centimetre

Ratio of majority to minority: $(5 \times 10^{15}) \div (4.5 \times 10^{4}) \approx 1.1 \times 10^{11}$

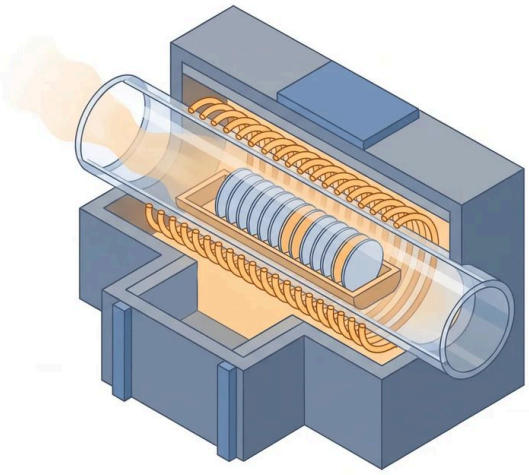
Electrons outnumber holes by about a hundred thousand million to one, and the hole population has fallen to roughly a three-hundred-thousandth of its value in the undoped crystal. This is why reverse leakage current — which is carried by minority carriers alone — is measured in nanoamperes while forward current is measured in amperes.

Light and Heavy Doping

Doping is not a single setting. The concentration is chosen for the job, and three of the devices later in this note exist only because of an unusual choice:

- **Heavy doping** (often written N+ or P+) gives low resistivity and, at a junction, a **narrow** depletion region — the same amount of exposed charge is reached over a shorter distance because there are more ions per unit volume. Narrow depletion regions break down at low reverse voltage. This is deliberately exploited in the Zener diode, and taken to an extreme in the tunnel diode.
- **Light doping** gives higher resistivity and a **wide** depletion region, hence a high reverse-voltage rating. High-voltage rectifiers are built this way, accepting the higher forward resistance that comes with it.
- **No doping at all in a deliberate layer** — an intrinsic region sandwiched between P and N — gives the PIN diode its very wide depletion region, low capacitance and the resistance-control behaviour used in radio-frequency switching.

A junction is normally created not by joining two separate pieces of material but by diffusing or implanting the opposite dopant into part of an already-doped wafer, so that the transition from P to N occurs inside one continuous crystal. This matters electrically: a mechanical contact between a lump of P-type and a lump of N-type would not work, because the crystal lattice must be unbroken across the junction for the carriers to move between the two regions.



Diffusion: wafers are heated in a dopant atmosphere, and the dopant migrates into their surface.

⚠ **A doped block on its own is only a resistor.** N-type and P-type material each conduct in either direction and have no preferred polarity whatever. Nothing rectifies, nothing switches and nothing regulates until a *junction* exists between the two. Every property in the rest of this note belongs to the junction, not to the materials.

Carrier Lifetime, and How It Is Manipulated

A minority carrier injected into the opposite region does not survive indefinitely: it recombines, and the average time it lasts is the **carrier lifetime**. The average distance it travels before that happens is the **diffusion length**. Both are properties of the material and of its defects, and both have direct consequences for how a finished device performs.

A long lifetime means a large quantity of injected charge is in transit at any moment. That charge is stored, and when a conducting junction is suddenly reversed it must all be removed before the junction can block — which is precisely the reverse recovery delay that limits how fast a rectifier can work. Shorten the lifetime and there is less charge to remove, so the device recovers faster.

Manufacturers do exactly that, deliberately. Introducing a controlled quantity of a heavy metal such as gold creates **recombination centres** — energy levels in the middle of the band gap that act as stepping stones, letting an electron and a hole meet and annihilate far more readily than they otherwise would. The result is the fast-recovery rectifier. The trade is real and worth naming: the same recombination centres also increase the reverse leakage current and raise the forward voltage drop slightly, so a fast-recovery diode is a poorer blocker and a slightly lossier conductor than the ordinary rectifier it replaces. Speed is bought, not given.

Why the Leads Are Not Two Extra Diodes

A junction between a metal and a semiconductor rectifies — that is exactly what a Schottky diode is. Every diode has metal leads bonded to semiconductor at each end, so a fair question is why those two contacts do not add two unwanted diodes in series with the one that was wanted.

The answer is doping. The semiconductor immediately beneath each contact is doped extremely heavily, which makes the depletion region there so thin that carriers pass straight through it in *both* directions rather than having to climb over it. A contact made this way conducts equally well either way and is called an **ohmic contact** — it behaves as a plain resistance of a fraction of an ohm. A Schottky diode is made by deliberately omitting that heavy doping, so the same physical arrangement of metal on semiconductor becomes rectifying instead.

The same reasoning explains the extreme case in the other direction. Dope a semiconductor heavily enough and the dopant levels merge into the conduction or valence band and the material begins to behave like a metal; it is then described as **degenerate**. Both sides of a tunnel diode are doped to that point, which is why its depletion region is thin enough for carriers to tunnel through at a fraction of a volt.

How a Junction Is Actually Manufactured

Method	What happens	Characteristics of the result
Grown junction	The dopant in the melt is changed part-way through pulling the crystal, so the transition is built into the boule as it forms	The original technique; large junction area, low frequency capability, now largely historical
Alloyed junction	A pellet of trivalent metal is melted onto an N-type wafer and allowed to recrystallise, forming a P region where it penetrated	Simple and robust, good current handling, poor control of junction depth
Diffused junction	The wafer is heated in an atmosphere containing the dopant, which migrates a controlled distance into the surface	Accurate, repeatable depth; the basis of the planar process and hence of all integrated circuits
Epitaxial layer	A new crystal layer of chosen doping is grown on the wafer surface, continuing the existing lattice	Allows a lightly doped layer for voltage rating over a heavily doped substrate for low resistance — the standard structure of a modern power rectifier
Ion implantation	Dopant ions are accelerated and fired into the surface, then the damage is annealed out	The finest control of all — dose and depth are set electrically rather than chemically

The epitaxial row explains a structure that otherwise looks contradictory. A rectifier wants a *lightly* doped region to give it a wide depletion region and a high reverse voltage rating, but light doping also means high

resistivity and therefore a large forward drop. Growing a thin lightly doped epitaxial layer on top of a thick heavily doped substrate resolves it: the thin layer supports the reverse voltage, and the thick substrate underneath carries the current with almost no resistance.

Doping Stops Working at Both Temperature Extremes

The statement that the majority-carrier concentration is fixed by the doping and independent of temperature is true across the normal working range and fails at both ends of it. Both failures are worth understanding, because the hot one sets the junction temperature limit that appears on every data sheet.

At high temperature the material reverts to intrinsic behaviour. Thermally generated carriers multiply exponentially with temperature while the dopant-supplied carriers stay constant, so there must be a temperature at which the thermal population catches up with the doping and then overtakes it. Beyond that point the crystal contains roughly equal numbers of electrons and holes regardless of what was added to it, and it has lost its P or N character altogether. A junction between two regions that are both effectively intrinsic is no junction at all: the barrier disappears, the device stops rectifying, and it conducts in both directions.

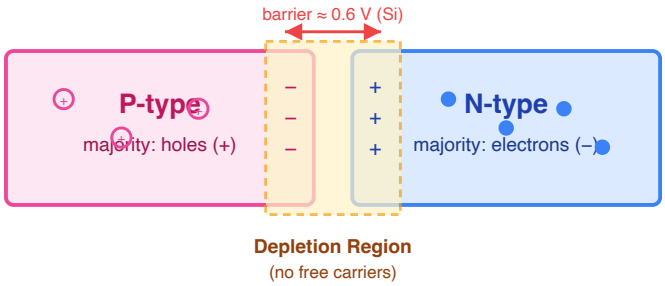
Work the comparison through and the silicon-versus-germanium table above stops being a list to memorise. Germanium's smaller band gap means its thermal carrier population is far larger at every temperature, so it reaches the crossover point at a much lower temperature than silicon does. That single fact is why germanium devices are restricted to modest temperatures while silicon devices work in hot equipment bays — not because germanium is fragile, but because it runs out of doping first.

⚠ **This is the real meaning of a maximum junction temperature.** Exceeding it does not simply risk melting something; it takes the semiconductor toward the condition where it is no longer a semiconductor in the useful sense. And because leakage current rises steeply on the way there, and leakage current itself generates heat, the approach to that limit is self-accelerating — which is the thermal runaway described earlier. The junction temperature limit is where a device stops being a device, and it is reached at a case temperature well below any figure that sounds dangerous to the touch.

At very low temperature the opposite happens, and it is included for completeness rather than for the flight line. If the crystal is cooled far enough, the thermal energy available falls below even the small amount needed to release a dopant's surplus electron, so the dopants stop ionising and the majority-carrier population collapses — an effect called carrier freeze-out. This belongs to cryogenic work; at the temperatures an airframe actually cold-soaks to, silicon dopants remain fully ionised and the effect is not encountered. What *is* encountered at low temperature is the ordinary junction behaviour running the other way: the barrier potential rises as the junction gets colder, at the same 2 mV per degree, so a cold-soaked circuit needs slightly more voltage to start conducting than the same circuit warm.

The PN Junction

When P-type and N-type materials are joined, a **PN junction** is formed — this is the heart of every diode. At the junction, free electrons from the N-side diffuse across and fill holes on the P-side. This creates a thin region with no free carriers called the **depletion region** (or depletion zone).



The exposed fixed ions on each side of the junction create a small **potential barrier** (about 0.6 V for silicon, 0.3 V for germanium). This barrier opposes further diffusion and establishes equilibrium. No current flows through an unbiased PN junction.

How the Barrier Builds Itself

The junction is not assembled with a barrier already in place; the barrier is a consequence of the joining and it builds in a few sequential steps, each caused by the one before it. Following the sequence is the only reliable way to get the polarities the right way round.

- **1 — A concentration cliff exists.** One side has an enormous surplus of free electrons, the other an enormous surplus of holes. Diffusion needs no field, only a difference, and the difference here is around eleven orders of magnitude.
- **2 — Carriers diffuse across.** Electrons cross from N into P; holes cross from P into N. Each one that crosses arrives in a region where it is now a **minority carrier**, surrounded by the opposite kind, and so is quickly lost to recombination near the junction.
- **3 — Fixed ions are uncovered.** An electron that leaves the N-side takes its negative charge with it and abandons a **fixed positive donor ion**; a hole that leaves the P-side leaves a **fixed negative acceptor ion**. Neither ion can follow. A layer of exposed, immobile charge therefore appears on each side of the junction — positive on the N side, negative on the P side.
- **4 — That charge pair is an electric field.** Two sheets of opposite fixed charge facing each other across a thin gap produce a field pointing from the positive sheet to the negative one, that is, from the N side toward the P side.

- **5 — The field shuts the process down.** A hole trying to diffuse from P into N now has to move *against* that field, and an electron trying to diffuse from N into P likewise. Every further carrier that crosses strengthens the field that opposes the next one, so diffusion slows and then stops.

The self-limiting nature of step 5 is why the barrier settles at a definite, repeatable value instead of growing without limit. It is also why the barrier is a property of the material rather than of the manufacturing: silicon always arrives at roughly 0.6 to 0.7 V and germanium at roughly 0.2 to 0.3 V, because those are the voltages at which each material's field is strong enough to hold back its own diffusion.

Two Currents That Cancel

Saying that the barrier "stops" diffusion is a simplification worth unpicking, because the exam tests the precise version. The junction at rest is not still — it is in dynamic balance, with two opposed currents of equal size crossing it at every instant:

Current	Which carriers	Driven by	Direction across the junction	Effect of the barrier on it
Diffusion current	Majority carriers — electrons from N, holes from P	The concentration difference	Majority carriers moving into the opposite region	**Opposed.** They must climb the barrier
Drift current	Minority carriers — the few electrons in P and holes in N	The barrier field itself	Minority carriers swept back the other way	**Assisted.** They fall down the barrier

These two are equal in size and opposite in direction, so the *net* current is zero — which is precisely why an unbiased junction passes nothing to an external circuit even though charge is crossing it continuously. The distinction is a favourite examination point: the correct reason for zero current is not that no charges are crossing, but that **equal and opposite currents are crossing**.

Read the right-hand column again, because it contains the whole of diode behaviour in one line. **The junction barrier opposes majority carriers in both regions and assists minority carriers in both regions.** Forward bias lowers the barrier and so releases the majority-carrier flood; reverse bias raises it and leaves only the minority-carrier trickle. Nothing else has to be remembered about bias.

The Barrier Is Not a Battery

A newly qualified engineer will sooner or later put a voltmeter across an unconnected diode expecting to read 0.6 V, and will read zero. The barrier potential exists **inside** the crystal, across the depletion region only, and it

cannot appear at the terminals.

The reason is worth stating properly rather than as a rule. Two further contact potentials exist where the metal leads meet the semiconductor at each end, and around the complete loop these exactly cancel the junction's own potential. If they did not, a diode with its leads shorted together would drive a permanent circulating current from nothing at all, and would grow hotter or colder for ever. The barrier is a potential *step* that charge must climb, not a source that can deliver energy.

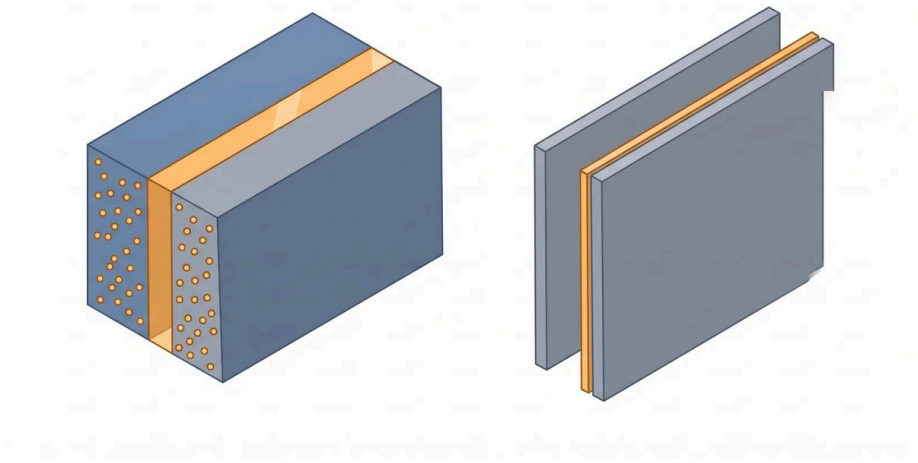
What the 0.6 V on a meter really is: when a multimeter shows about 0.6 V in diode-test mode it is not measuring the barrier directly. The meter is pushing its own small test current through the junction and displaying the forward voltage that current produces. That is why the reading only appears with the leads one way round, and why the meter must be in diode-test mode rather than simply on a resistance range.

Depletion Width, and Why the Junction Is a Capacitor

The depletion region is thin — typically a fraction of a micrometre — and its width is not fixed. Two things set it:

- **Doping concentration.** The region must expose enough fixed ionic charge to support the barrier voltage. Heavily doped material packs more ions into each cubic micrometre, so the required charge is reached over a **shorter** distance and the depletion region is **narrow**. Lightly doped material needs a **wider** region for the same charge.
- **Doping symmetry.** The exposed charge on each side must be equal and opposite in total, so if one side is doped a hundred times more heavily than the other, the depletion region extends a hundred times further into the **lightly** doped side. Almost all of a practical junction's depletion region lies in whichever side is more lightly doped.

Now look at the structure that produces: a layer with no free carriers — an insulator — sandwiched between two regions that are full of free carriers and therefore conduct. That is a parallel-plate capacitor, with the depletion region as its dielectric and the neutral P and N material as its plates. Every PN junction has capacitance whether the designer wants it or not.



A PN junction is a parallel-plate capacitor: a carrier-free layer between two conducting regions.

Because plate separation and capacitance are inversely related, the direction follows immediately: **widening the depletion region reduces the junction capacitance, and narrowing it increases the capacitance.** Reverse bias widens the region, so increasing the reverse voltage *decreases* the capacitance. In an ordinary diode this is a nuisance that limits switching speed; in the varactor it is the entire purpose of the device.

The Barrier Falls as the Junction Gets Hotter

Heat breaks additional covalent bonds and supplies more carriers with enough energy to climb the barrier, so less applied voltage is needed to start conduction. The barrier potential of a silicon junction therefore **decreases** by roughly 2 mV for every degree Celsius rise. Over a realistic 100°C swing in an equipment bay that is a change of about 0.2 V, which is a third of the whole barrier — large enough to matter in any circuit that uses a junction drop as a reference, and the reason such circuits are built with a matched pair of junctions whose drifts cancel rather than with one.

The same mechanism runs the other way on the minority-carrier side: a hotter junction has more thermally generated minority carriers, so the drift current, and therefore the reverse leakage current, grows steeply with temperature. Two numbers describe the same underlying cause — more broken bonds.

Abrupt and Graded Junctions

How *sharply* the doping changes from P to N is a design variable in its own right, and it decides how the junction's capacitance responds to voltage:

- An **abrupt** junction changes from full P doping to full N doping over a very short distance. Its capacitance falls off roughly as the square root of the applied reverse voltage.

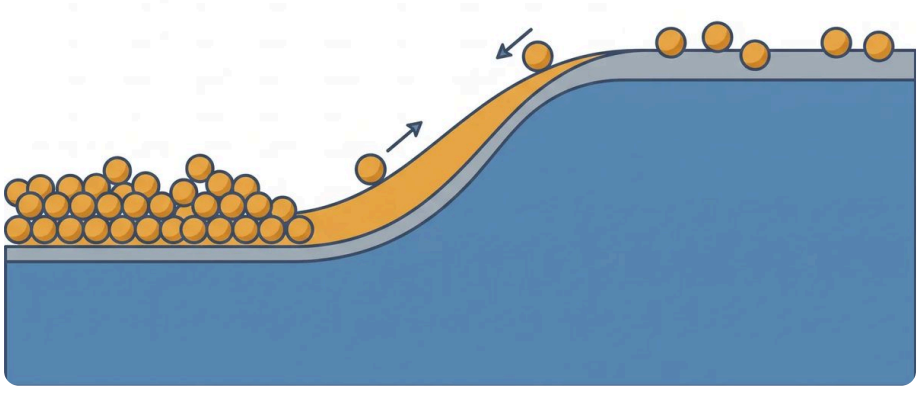
- A **graded** junction changes gradually over a greater distance. The capacitance falls off more slowly, roughly as the cube root.
- A **hyperabrupt** junction is manufactured with the doping deliberately increasing toward the junction, which makes the capacitance change much faster than either.

For an ordinary rectifier none of this matters. For a varactor it is the entire specification, because it determines how wide a frequency range one device can tune and how linear the relationship between control voltage and frequency turns out to be. Two varactors with the same nominal capacitance and different grading are not interchangeable.

The Junction Described in Energy Terms

The same junction can be drawn as an energy diagram rather than as a picture of charges, and this second view makes several later results obvious. Before contact the two materials' conduction bands sit at the same energy, but the N material's occupancy level — its Fermi level — sits higher than the P material's, because its surplus electrons fill states further up toward the conduction band. When the two are joined, charge redistributes until the occupancy levels line up across the junction, and the only way that can happen is for the bands on the N side to shift down relative to those on the P side. The bands therefore **bend** through the depletion region, and the height of the bend is the barrier potential expressed as energy instead of voltage.

Three things fall straight out of the picture. A majority carrier crossing the junction must climb the bend, which is why it needs energy and why the barrier obstructs it. A minority carrier arriving at the junction is at the top of the bend and slides down, which is why the barrier assists it. And applying forward bias reduces the bend while reverse bias increases it, which is the same statement as raising and lowering the barrier.



Majority carriers must climb the band bend; minority carriers arriving at the top slide down it.

The barrier's height is not arbitrary either: it depends on how heavily each side is doped and on the material's intrinsic carrier concentration. Heavier doping on either side raises it slightly. This is why silicon's figure is quoted as a range from about 0.6 to 0.7 V rather than as a single constant — the spread is real, and it is set at the factory by the doping recipe.

A Junction Has Two Different Capacitances

The parallel-plate capacitance described above is the **transition** or **depletion** capacitance, and it belongs to the reverse-biased junction, where there is a genuine carrier-free layer to act as a dielectric. Under forward bias that layer has largely collapsed, so it might seem that the capacitance disappears. It does not — it is replaced by a different and much larger effect.

A forward-biased junction is continuously injecting minority carriers into the material on each side, and at any instant a quantity of that injected charge is in transit. Changing the applied voltage changes how much charge is in transit, and a stored charge that varies with applied voltage is, by definition, a capacitance. This is the **diffusion** capacitance, and it is typically far larger than the transition capacitance of the same junction.

The distinction is worth carrying forward because the two dominate in opposite conditions and they cause two different practical limits:

- **Transition capacitance** dominates in reverse bias, decreases as the reverse voltage rises, and is the basis of the varactor. It is also what lets a high-frequency signal leak past a supposedly blocking junction.
- **Diffusion capacitance** dominates in forward bias, increases with forward current, and is the stored charge that must be swept out before a conducting junction can block. It is therefore the origin of reverse recovery time.

That last connection explains something about the Schottky diode that would otherwise look like a coincidence. Because a metal-semiconductor junction injects no minority carriers, it has no diffusion capacitance at all — and so it has no stored charge, and therefore no reverse recovery delay. One structural difference accounts for both.

Forward and Reverse Bias

Forward Bias

Connect the **positive** terminal of a battery to the P-side and the **negative** terminal to the N-side. This pushes holes toward the junction from the P-side and electrons toward the junction from the N-side, narrowing the

depletion region. Once the applied voltage exceeds the barrier potential (~0.6 V for Si), current flows freely through the junction.

Reverse Bias

Connect the **positive** terminal to the N-side and the **negative** terminal to the P-side. This pulls carriers away from the junction, widening the depletion region. Only a tiny **leakage current** (due to minority carriers) flows. The junction effectively blocks current.

If the reverse voltage exceeds the **breakdown voltage**, the junction conducts heavily in reverse — this is destructive in normal diodes but is exploited in Zener diodes.

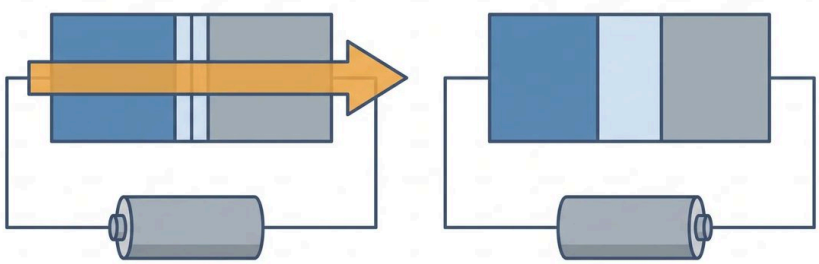
> **Key Rule**
> **Forward biased:** P → positive, N → negative → current flows (diode ON)
> **Reverse biased:** P → negative, N → positive → current blocked (diode OFF)

What the Applied Voltage Does to the Barrier

Both bias conditions act on the same thing — the height of the potential barrier — and once that is fixed in mind the two behaviours stop being separate facts to memorise:

- **Forward bias opposes the barrier field and so LOWERS the barrier.** The external source drives holes toward the junction from the P side and electrons toward it from the N side, and these mobile carriers move into the depletion region and neutralise part of the exposed fixed ionic charge. Less exposed charge means a smaller internal field, a lower barrier and a **narrower** depletion region.
- **Reverse bias reinforces the barrier field and so RAISES the barrier.** The external source pulls majority carriers *away* from the junction on both sides, uncovering still more fixed ions. More exposed charge means a stronger internal field, a higher barrier and a **wider** depletion region.

Note that the depletion width and the barrier height always move together, and both follow the applied voltage in the same sense. Forward: barrier down, depletion narrow, resistance low. Reverse: barrier up, depletion wide, resistance high. There is no condition in which one goes up and the other goes down.



Forward bias narrows the depletion region and lets current flow; reverse bias widens it and blocks current.

Reading a Bias Condition from Two Node Voltages

A question may give the voltage at each terminal instead of naming the bias condition, and expect the candidate to work it out. The arithmetic is then a subtraction, not a comparison against zero:

Worked Example — Is This Diode Conducting?

A silicon diode has its anode connected to a +4 V supply and its cathode connected to a +2 V supply. Both terminals are positive with respect to ground. Is it forward biased, and is it conducting?

Voltage across the diode, anode with respect to cathode: $4 - 2 = +2$ V.

The anode is positive with respect to the cathode, so the diode is **forward biased**.

Is 2 V enough? The silicon barrier is about 0.7 V, and $2 > 0.7$, so the diode is also **conducting**.

Both parts of the answer are needed. Forward biased and conducting are not the same statement: a diode with +0.3 V across it in the forward direction is forward biased and is *not* conducting, because the applied voltage has not yet overcome the barrier. Change the figures to anode +2 V and cathode +4 V and the diode sees –2 V, is reverse biased and blocks. Neither terminal ever needs to be negative with respect to ground for either conclusion.

Who Carries the Forward Current, and Which Way

Forward conduction is a **majority-carrier** process, and it happens on both sides of the junction at once. Electrons — the majority carriers of the N region — flow toward the junction and cross it; holes — the majority carriers of the P region — do the same in the opposite direction. Both contribute, which is why a question asking what carries the current in a forward-biased junction is correctly answered either as "majority carriers" or as "electrons and holes". The one answer that is wrong is "minority carriers": those are responsible for the reverse leakage current instead.

Direction needs stating in both conventions, because the exam uses both and they are opposite:

Convention	Direction inside the diode	Enters the device at	Relationship to the symbol
Conventional current	Anode → cathode	The anode	The same way the triangle points
Electron flow	Cathode → anode	The cathode	Against the triangle

In circuit terms the cathode is connected to the negative side of the supply, so electrons are pushed into the diode at the cathode end, travel through the N region, cross the junction and leave at the anode. Saying that **electrons leave the cathode** in a forward-biased diode describes the same motion as saying that conventional current enters at the anode — they are two descriptions of one event, not a contradiction.

Above the Knee, the Diode Stops Limiting Anything

Once the applied voltage exceeds the barrier, current rises **exponentially** with voltage: a further 60 mV multiplies the current by about ten. The practical consequence is that the forward drop is almost constant over a wide current range — but also that the diode's own resistance has become negligible, so it no longer restrains the current at all. Whatever the rest of the circuit allows through is what will flow.

⚠ **A forward-biased diode must never be connected directly across a voltage source.** With nothing to limit it, the current is set only by the source impedance and the wiring, and the junction is destroyed almost instantly. Every forward-conducting diode in a circuit has a current-limiting element somewhere in series with it — a resistor, an inductor, a transformer winding or the load itself.

Worked Example — Sizing the Series Resistor

A silicon diode is to be run at about 20 mA from a 28 V DC bus. Find the series resistor and check its power rating.

Voltage left for the resistor: $28 - 0.7 = 27.3$ V

Required resistance: $27.3 \div 0.020 = 1365\ \Omega$, so the nearest standard value upward is 1.5 kΩ

Actual current: $27.3 \div 1500 = 0.0182$ A, that is **18.2 mA**

Power in the resistor: $0.0182^2 \times 1500 = 0.497$ W, so a 1 W resistor is required

Power in the diode: $0.7 \times 0.0182 = 0.0127$ W, that is 12.7 mW — trivial

Almost all the power is burned in the resistor, not the diode, because the resistor carries thirty-nine times the voltage at the same current. Choosing the next standard resistor value *upward* is the safe direction: it reduces the current below the target rather than above it.

Reverse Bias: What the Leakage Current Actually Is

Under reverse bias the majority-carrier diffusion current is shut off completely — the barrier is now too high for any majority carrier to climb. What remains is the drift current, carried by the thermally generated **minority carriers** that happen to wander into the depletion region and are then swept across by the field. That is the reverse leakage current, and its behaviour follows directly from its origin:

- **It hardly changes with reverse voltage.** Every minority carrier that reaches the depletion region is already being swept across; raising the voltage cannot sweep them any harder. This is why it is also called the reverse **saturation** current, and why the reverse part of the characteristic curve is nearly flat.
- **It changes enormously with temperature**, roughly doubling for each 10°C rise, because temperature is what creates the carriers in the first place. A junction that leaks 10 nA at 25°C leaks in the order of a microampere at 90°C.
- **It is far larger in germanium than in silicon**, for the same reason germanium has the smaller barrier: its bonds are easier to break.

The Two Breakdown Mechanisms

Push the reverse voltage high enough and the current stops being a trickle and rises almost vertically. Two separate physical mechanisms can produce that, they dominate in different voltage ranges, and they respond to temperature in *opposite* directions:

	Zener breakdown	Avalanche breakdown
Mechanism	The field across a very thin depletion region becomes so intense that it pulls valence electrons directly out of their covalent bonds	A minority carrier is accelerated by the field until it has enough energy to knock an electron out of a lattice atom on impact; the freed carriers are accelerated in turn and knock out more, as a chain reaction
Needs	Heavy doping, giving a narrow depletion region and a very high field at modest voltage	Lighter doping, giving a wide depletion region in which a carrier has room to accelerate between collisions
Dominates	Below about 5 V	Above about 6 V
Temperature coefficient	**Negative** — heat weakens the bonds, so a slightly lower field suffices and the breakdown voltage falls	**Positive** — heat increases lattice vibration, so carriers collide sooner, gain less energy between collisions, and a higher field is needed; the breakdown voltage rises

Between roughly 5 V and 6 V both mechanisms contribute and their opposite temperature coefficients very nearly cancel, which is why the most temperature-stable reference diodes are made in that voltage range.

Note the naming trap: every diode sold as a "Zener" above about 6 V is actually working by avalanche, and the label has simply stuck.

Breakdown is not the same as damage. What destroys a junction is power, not the breakdown itself — reverse voltage multiplied by reverse current, concentrated in a very small volume. A Zener diode operates in continuous reverse breakdown for its whole service life and is perfectly happy, because a series resistor holds the current to a safe value. An ordinary rectifier taken past its peak inverse voltage has nothing limiting the current, so the same physics ends in a short-circuited junction within microseconds.

The Three Conditions Side by Side

Condition	Barrier height	Depletion region	Majority-carrier current	Minority-carrier current	Net current	Junction resistance
Unbiased	At its natural value	At its natural width	Diffusion, moderate	Drift, equal and opposite	**Zero** (the two cancel)	Not defined — no external voltage
Forward biased	**Lowered** by the applied voltage	**Narrowed**	Very large above the knee, rising exponentially	Unchanged and negligible by comparison	Large, anode to cathode	Very low
Reverse biased	**Raised** by the applied voltage	**Widened**	Suppressed to nothing	Unchanged — this is all that is left	Tiny leakage, cathode to anode	Very high

Finding the Operating Point: the Load Line

The exponential characteristic creates a problem that catches out anyone applying Ohm's law directly. The diode's current depends on its voltage in a curved way, and its voltage depends on how much current the rest of the circuit will deliver — each depends on the other. There are two constraints and they have to be satisfied simultaneously.

The graphical solution is the **load line**. The external circuit — a source and a resistor — obeys a straight-line relationship between the voltage it can leave across the diode and the current it will pass, and that line is drawn on the same axes as the diode's curve. The point where they cross satisfies both, and it is the operating point.

Worked Example — Drawing and Reading a Load Line

A silicon diode is fed from a 5 V source through a 100 Ω resistor. Find the two ends of the load line, then the operating point.

End 1 — diode open circuit. No current flows, so the resistor drops nothing and the whole 5 V appears across the diode. Point: 5 V, 0 mA.

End 2 — diode short circuit. The whole 5 V is across the resistor, so $5 \div 100 = 0.050$ A. Point: 0 V, 50 mA.

Join the two points with a straight line and read where it crosses the diode curve: at about 0.7 V, giving $(5 - 0.7) \div 100 = 0.043$ A, that is **43 mA**.

Now change the resistor to 1 kΩ and repeat. The short-circuit end moves to 5 mA, the line becomes much shallower, and it crosses the curve at a slightly lower voltage — around 0.64 V — for a current of about 4.4 mA. Note what changed and what did not: the current fell by a factor of ten while the diode voltage moved by only about 60 mV. That is the practical meaning of the near-vertical forward characteristic, and it is why treating the drop as a fixed 0.7 V is accurate enough for almost all work.

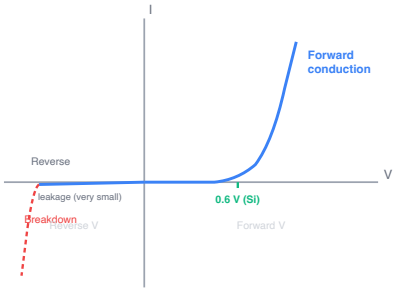
The same construction is what makes a Zener regulator, a clipper and a rectifier all analysable by hand, and it will reappear in the amplifier material where the load line is used to set a transistor's bias point.

How Strong the Field Across the Junction Actually Is

It is worth appreciating the scale of the field involved, because it explains why breakdown happens at voltages that sound harmless. A depletion region a fraction of a micrometre thick, with a few volts across it, produces a field strength of hundreds of thousands of volts per centimetre — enormously higher than the field that would flash over an air gap.

This is why a reverse voltage that is unremarkable as a circuit voltage is dramatic inside the crystal, and why the two breakdown mechanisms need such different structures. A heavily doped junction concentrates its whole reverse voltage across a very thin layer, so the field reaches bond-breaking strength while the terminal voltage is still only a few volts. A lightly doped junction spreads the same voltage over a much wider layer, so the field stays lower — but the extra width gives a carrier the running room it needs to accelerate to ionising energy before its next collision. Each mechanism is enabled by the very geometry that suppresses the other, which is why they dominate in different voltage ranges rather than competing throughout.

Diode Characteristics and Parameters



Important Diode Parameters

Parameter	Description
Forward voltage drop (V_F)	Voltage across the diode when conducting: ~0.6 V for Si, ~0.3 V for Ge, ~0.2 V for Schottky
Peak Inverse Voltage (PIV)	Maximum reverse voltage the diode can withstand without breakdown
Maximum forward current ($I_{F\,max}$)	Maximum continuous current in forward direction before damage
Leakage current (I_R)	Small current that flows in reverse bias (due to minority carriers); increases with temperature
Power dissipation	$P = V_F \times I_F$. Must not exceed the diode's thermal rating.
Reverse recovery time	Time to switch from forward conduction to reverse blocking. Important at high frequencies.
Temperature sensitivity	V_F decreases ~2 mV/°C; leakage current roughly doubles for every 10°C rise

Reading the Characteristic Curve

The curve above is one continuous line through two quadrants, and each part of it answers a different question. Working along it from the right:

- **The forward foot**, from the origin out to the knee, is nearly flat. Some current does flow here — the exponential law never gives exactly zero — but it is microamperes, far too little to matter, and the diode is treated as off.
- **The knee** is where the curve turns sharply upward. It is the forward voltage at which significant current begins to flow, and it is what a data sheet and an examiner both mean by the threshold, cut-in or turn-on voltage.
- **The forward conduction region** above the knee is nearly vertical. The slope of the curve is the reciprocal of the diode's resistance, so a near-vertical line means a very low resistance. It is never truly vertical, because the plain resistance of the silicon on each side of the junction remains in series.

- **The reverse region** to the left of the origin is nearly horizontal, sitting a hair below the axis. A near-horizontal line means a very high resistance. That it is horizontal rather than sloping is the graphical statement that leakage barely depends on reverse voltage.
- **The breakdown knee** at the far left is where the curve plunges. Beyond it the diode conducts heavily in the reverse direction, and the line is again nearly vertical — the resistance has collapsed.

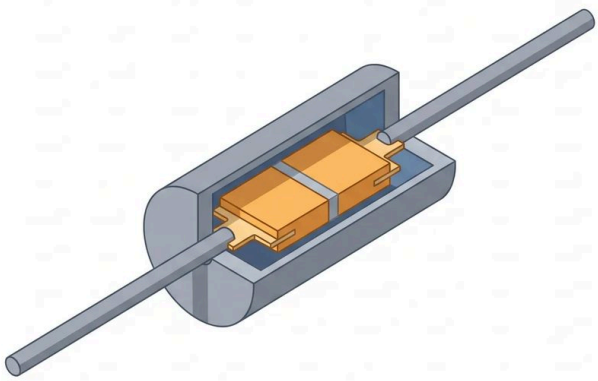
An **ideal** diode would be the two axes themselves: perfectly horizontal in reverse and perfectly vertical in the forward direction from the origin, that is, **zero forward resistance and infinite reverse resistance**, with no knee and no breakdown. Every departure of a real curve from that pair of straight lines is a departure worth naming, and the parameters in the table above are the names.

Three Models, and When to Use Each

Circuit analysis uses whichever simplification is good enough for the question in hand. Choosing badly wastes time or gives a wrong answer:

Model	The diode is treated as	Use it when	Typical error it introduces
Ideal	A perfect switch: a short circuit when forward biased, an open circuit when reverse biased	The supply voltage is very much larger than 0.7 V, or only the shape of a waveform is wanted	Overstates the output by one diode drop per conducting diode
Constant-voltage	A perfect switch in series with a fixed 0.7 V battery (about 0.3 V for germanium, and 0.2 to 0.3 V for a Schottky)	Almost all practical work, and almost every calculation in this note	Ignores the small rise in forward drop at high current
Complete	0.7 V plus the bulk resistance of the semiconductor material in series	High-current work, and any calculation of the diode's own power dissipation	Very little — but it needs a data-sheet figure

The **bulk resistance** named in the third row is worth defining because it is often confused with the junction. It is simply the ordinary ohmic resistance of the neutral P and N material and of the internal connections, excluding the junction altogether. It is what stops the forward characteristic from being vertical, and it is what makes a rectifier's forward drop climb from about 0.7 V at a few milliamperes to appreciably above 1 V at its full rated current.



Bulk resistance is the plain P and N material and internal connections, in series with the junction.

DC Resistance and AC Resistance Are Different Numbers

A diode has no single resistance, and quoting the wrong one is a common error. There are two, they are calculated differently, and they can differ by a factor of thirty at the same operating point:

- **DC (static) resistance** is simply the operating voltage divided by the operating current, $R_{DC} = V_F / I_F$. It describes the diode's effect on the steady bias conditions.
- **AC (dynamic) resistance** is the slope of the characteristic at that point — how much the voltage changes for a small change in current. For a forward-biased junction at room temperature it works out very close to $r_d \approx 26\ \mathrm{mV} / I_F$, with the current in amperes. It describes the diode's effect on a small signal riding on the bias.

The 26 mV in that expression is the **thermal voltage**, the natural voltage scale of a junction at room temperature. It is not a property of silicon in particular; it depends only on temperature, and it appears in the diode equation as well.

Worked Example — The Same Diode, Two Resistances

A silicon diode is carrying 10 mA with 0.7 V across it. Find both resistances, and then the AC resistance at 1 mA and at 100 mA.

DC resistance: $0.7 \div 0.010 = 70\ \Omega$

AC resistance at 10 mA: $0.026 \div 0.010 = 2.6\ \Omega$

AC resistance at 1 mA: $0.026 \div 0.001 = 26\ \Omega$

AC resistance at 100 mA: $0.026 \div 0.100 = 0.26\ \Omega$

Two lessons. First, the AC resistance is about twenty-seven times lower than the DC resistance at the same point, so which one is quoted matters enormously. Second, the AC resistance is **inversely proportional to the current** — ten times the current gives one tenth the resistance. That is the property exploited in a PIN diode used as a radio-frequency attenuator, where a DC control current sets the effective resistance seen by the signal.

The underlying law behind all of this is the **diode equation**, which states that the current is an exponential function of the applied voltage: $I = I_S(e^{V/nV_T} - 1)$, where I_S is the reverse saturation current, n is the ideality factor (between 1 and 2 for real junctions) and V_T is the thermal voltage. The equation is worth recognising rather than manipulating. Notice what it says in reverse bias: make V negative and large, the exponential term falls to nothing, and the current settles at $-I_S$ — a constant. That is the reverse saturation current, and the equation predicts the flat reverse characteristic without any extra assumption.

Peak Inverse Voltage in Service

PIV is a maximum, not a target, and there is no graceful degradation on the far side of it. Three points decide whether a diode survives in a real circuit:

- **The stress is set by the circuit topology, not by the supply voltage.** The section on rectifier circuits below shows that for the same DC output voltage the peak inverse stress is one peak voltage in a bridge rectifier and twice that in a centre-tapped rectifier. Choosing the diode requires knowing the circuit.
- **Transients count.** Switching an inductive load, a lightning-induced surge on a supply feeder or a contactor opening can put a spike on a line many times the nominal voltage and lasting microseconds. The diode does not care that it was brief.
- **Margin is normal practice.** A rectifier is selected with a PIV rating comfortably above the worst calculated stress, and where transients are expected the line is additionally protected by a suppression device rather than by hoping the rectifier will absorb them.

Exceed the PIV and the junction avalanches. In a Zener that is normal service; in a rectifier there is nothing to limit the resulting current, so the junction overheats locally and fuses into a permanent short circuit. That failure mode — **short, not open** — is the usual one for an overstressed rectifier and it is what makes a single failed diode in a bridge so destructive: it places a short across the transformer secondary on alternate half-cycles.

Current Ratings: Average, Repetitive and Surge

A data sheet gives more than one current figure and they are not interchangeable:

- **Average forward current** is the continuous rating — what the diode can carry indefinitely without exceeding its junction temperature. This is the number that appears in a catalogue description such as "3 A rectifier".
- **Repetitive peak forward current** is higher, and applies to the recurring current pulses that a rectifier feeding a capacitor filter actually experiences: conduction happens only in short bursts near the peaks of the waveform, so the peak current is several times the average.
- **Non-repetitive surge current** is higher still and is quoted for a single cycle. It exists because of what happens at switch-on: a discharged filter capacitor looks like a short circuit, and the inrush through the rectifier as it charges is limited only by the transformer and wiring resistance. A rectifier that is adequate in steady state can be destroyed in the first half-cycle after the bus is energised, which is why large power supplies include an inrush-limiting element.

Power Dissipation and Thermal Derating

The power the junction must get rid of is the product of the forward drop and the forward current, and it all appears in a region a fraction of a millimetre across. Getting it out to the air is a chain of thermal resistances — junction to case, case to heatsink, heatsink to ambient — and the junction temperature is the ambient temperature plus the total power multiplied by the total thermal resistance.

Worked Example — Will This Diode Survive the Bay It Is Fitted In?

A rectifier carries 1.0 A continuously with a forward drop of 1.0 V at that current. Suppose its data sheet quotes a junction-to-ambient thermal resistance of 50°C per watt and a maximum junction temperature of 150°C, and the equipment bay reaches 70°C.

Power dissipated: $1.0 \times 1.0 = 1.0$ W

Temperature rise above ambient: $1.0 \times 50 = 50$°C

Junction temperature: $70 + 50 = 120$°C

Margin to the limit: $150 - 120 = 30$°C — acceptable

Now repeat it for a 25°C bench. The junction sits at 75°C with a 75°C margin, and the same component looks twice as safe as it is. The figures used here are illustrative and every one of them must come from the actual data sheet, but the structure of the calculation does not change: **a current rating is only meaningful alongside the ambient temperature it was quoted at.**

Derating is the formal expression of that. Data sheets carry a curve showing the maximum permitted current falling as ambient temperature rises, reaching zero at the maximum junction temperature. Operating a

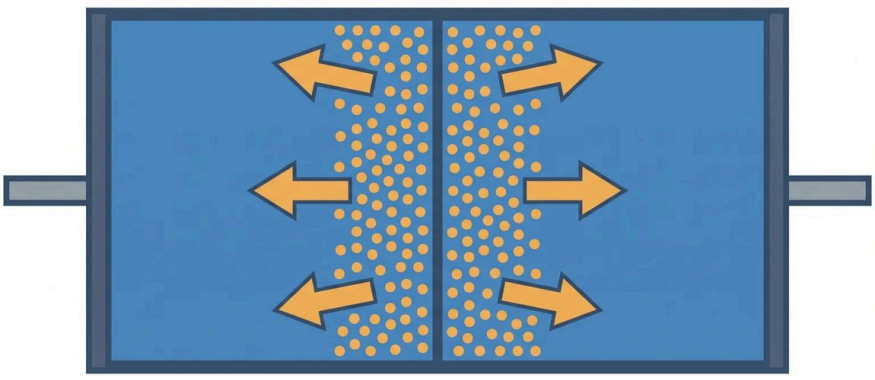
component at its full catalogue rating in a hot bay is one of the more common ways to build in an early failure, and it is invisible to any bench test done at room temperature.



A hot equipment bay raises junction temperature, so a current rating means little without its ambient.

Frequency: Why a Rectifier Has a Speed Limit

A diode that rectifies perfectly at 400 Hz may pass alternating current straight through at a megahertz. Two separate effects impose the limit:



Stored minority carriers flood the junction region and must be swept out before the diode can block.

- **Reverse recovery.** While conducting, the region either side of the junction is flooded with injected minority carriers. When the applied voltage suddenly reverses, that stored charge must be removed before the junction can block, and until it is, the diode conducts *backwards*. The reverse recovery time is how long

that takes. If the reverse recovery time is a significant fraction of the period, a large part of every cycle is spent conducting the wrong way and rectification collapses.

- **Junction capacitance.** A reverse-biased junction is a small capacitor, as the earlier section explained, and a capacitor's reactance falls as frequency rises. At a high enough frequency the capacitance simply carries the signal past the blocking junction.

This is what separates the diode types an engineer meets in a stores list:

Type	Recovery behaviour	Where it belongs
General-purpose rectifier	Slow — microseconds	Mains and 400 Hz supply rectification, where the period is milliseconds
Fast-recovery rectifier	Fast — tens of nanoseconds	Switching power supplies and inverter circuits, where turn-off must be quick
Soft-recovery rectifier	Fast, but with a deliberately gentle current tail	Circuits where a sharp recovery snap would radiate interference; chosen for electromagnetic compatibility rather than for raw speed
Schottky	Essentially none — there is no stored minority charge to remove	High-frequency switching, and anywhere the forward drop must be minimised

✈ **Aviation context:** aircraft AC systems run at 400 Hz rather than 50 or 60 Hz because transformers and motors can be made much smaller and lighter at the higher frequency. For rectification this is entirely benign — 400 Hz is still slow enough for ordinary rectifiers, and the higher ripple frequency makes smoothing easier. The demanding applications on board are elsewhere: switched-mode power supplies inside avionics units, which run at tens or hundreds of kilohertz and need fast-recovery or Schottky devices.

Reading a Diode Data Sheet

Every parameter in the table above appears on a manufacturer's data sheet under an abbreviated symbol, and knowing the symbols is what makes a substitution decision possible on the shop floor. The names that matter are these:

Symbol	Name	What it constrains
V_{RRM}	Repetitive peak reverse voltage	The reverse voltage the device may see on every cycle indefinitely. This is the number normally meant by PIV
$I_{F(AV)}$	Average forward current	The continuous rating, and always quoted alongside a case or ambient temperature

Symbol	Name	What it constrains
I_{FSM}	Surge forward current	A single non-repetitive half-cycle. Relevant only at switch-on
V_F	Forward voltage	Quoted *at a stated current*. A figure read at 1 mA is meaningless for a device that will run at 3 A
I_R	Reverse current	Quoted at a stated reverse voltage and temperature. The room-temperature figure is the optimistic one
t_{rr}	Reverse recovery time	How fast the device can be switched off; it decides the usable frequency
C_j	Junction capacitance	Quoted at a stated reverse bias, since it changes with voltage
R_{th}	Thermal resistance	Junction to case, or junction to ambient. It converts dissipated watts into a temperature rise
T_j	Junction temperature	The absolute limit that every other rating is ultimately protecting

⚠ **A rating with no stated conditions is not a rating.** Almost every figure on a data sheet is quoted at a specific current, voltage or temperature, and the difference between the quoted condition and the actual one is where substitutions go wrong. A device whose forward voltage looks identical to the original at 1 mA may be considerably worse at the current the circuit really uses, and a reverse current quoted at 25°C says nothing about a component sitting in a hot bay. Compare like with like or do not compare at all.

Where the Losses Go in a Rectifier

A diode's losses fall into two groups, and which one dominates tells you immediately what kind of device the circuit needs:

- **Conduction loss** is the forward drop multiplied by the current, present for as long as the device conducts. It does not depend on frequency.
- **Switching loss** is the energy wasted during each turn-off, while the reverse recovery charge is being swept out with voltage already appearing across the device. It is proportional to how often that happens, so it rises directly with frequency.

At 400 Hz the switching loss is negligible and the choice is made purely on forward drop. At a hundred kilohertz the switching loss can exceed the conduction loss, and a device with a slightly higher forward drop but much faster recovery will run cooler than one chosen on forward drop alone. This is the practical reason the type table earlier in this section exists.

Worked Example — Efficiency Cost of a Bridge

A bridge rectifier delivers 5 A at 28 V to a load. Each conducting diode drops 1.0 V at that current, and two are in series in the conduction path at any instant.

Power delivered to the load: $28 \times 5 = 140$ W

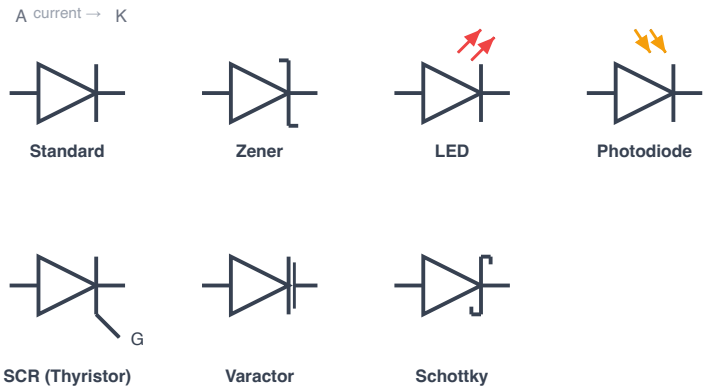
Power lost in the diodes: $2 \times 1.0 \times 5 = 10$ W

Total power drawn from the transformer secondary: $140 + 10 = 150$ W

Rectifier efficiency: $140 \div 150 = 93.3\%$

Now consider the same circuit delivering 5 A at 5 V. The load takes 25 W, the diodes still waste 10 W, and the efficiency collapses to $25 \div 35 = 71\%$. The loss did not change — the output did. That is why low-voltage supplies use Schottky rectifiers with their 0.3 V drop, and why a bridge is sometimes abandoned in favour of a centre-tapped circuit with only one diode drop in the path.

Diode Symbols



In all diode symbols, the **triangle (arrow)** points from anode (A) to cathode (K), showing the direction of **conventional current flow**. The cathode is the bar (line). On a physical diode, the cathode is typically marked with a band.

A Method for Reading Any Diode Symbol

Every symbol in the figure above is built from the same two strokes — a solid triangle and a bar across its tip — with a modifier added. Identifying a device from its symbol is therefore not a memory test over seven unrelated pictures but a three-step inspection, always in the same order:

- **Look at the bar first.** A plain straight bar means an ordinary diode. Anything bent, doubled or turned at the ends identifies a specific type, and the shape of the bend is the whole of the identification.

- **Then look for arrows beside the symbol.** Arrows mean light, and their **direction** says whether the device produces it or responds to it.
- **Then count the terminals.** A third lead means the device can be controlled, which restricts it to the thyristor family.

This is worth practising in that order, because a candidate who starts by looking for something familiar tends to fasten on the triangle — which every one of them has — and then guesses.

Symbol	What to look for	Why the modifier has that shape
Rectifier diode	A plain, straight cathode bar and nothing else	Nothing is added because nothing is special — the reference case
Zener diode	Two short flags at the ends of the bar, bent in **opposite** directions, so the bar reads as a squared letter **Z**	Z for Zener. Which end carries which flag varies between drawing conventions and means nothing electrically; what identifies the device is that the two flags oppose each other
Schottky diode	The same two flags, but each one **turns a second corner**, so the bar reads as a squared letter **S**	S for Schottky. This is the one that is most often confused with the Zener, and the extra corner at each end is the only thing that separates them
LED	Two parallel arrows **pointing away** from the symbol	The device emits — energy leaves it
Photodiode	Two parallel arrows **pointing toward** the symbol	The device receives — energy arrives at it
Varactor	The cathode bar doubled into two parallel lines	Two parallel plates are the symbol for a capacitor, and a voltage-controlled capacitor is exactly what the device is
SCR (thyristor)	An ordinary diode with a **third lead** leaving the cathode end of the bar	The gate. Its position at the cathode end is not decorative — the gate really does connect to the layer next to the cathode, and the trigger pulse is applied gate-to-cathode

⚠ **Zener and Schottky are the pair to get right.** Both modify the cathode bar with small flags and at a glance they look alike, so the distinction has to be made deliberately: **the Zener's flags each run straight and stop; the Schottky's flags each turn again.** Z has two corners, S has four. If a question shows a symbol and offers both as options, that second corner is the entire answer.

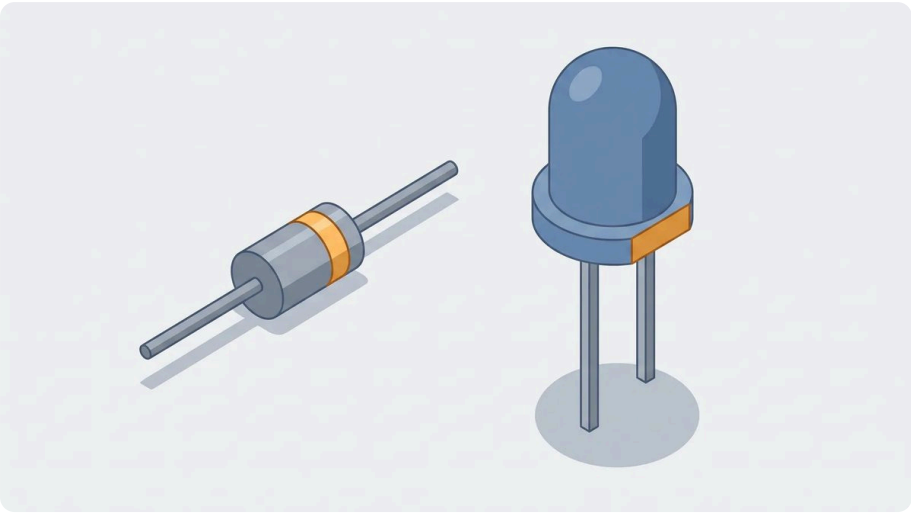
Symbols Not in the Figure Above

Three more appear regularly in aircraft schematics and in examination papers, and all three are built from the same parts:

- **TRIAC** — two diode triangles pointing in *opposite* directions, joined so that they share a common bar, with a gate lead brought out at one side. The back-to-back arrangement is a direct picture of what the device is: two thyristors in inverse parallel, able to conduct in either direction. Because there is no meaningful anode or cathode, its two main terminals are called **main terminal 1** and **main terminal 2** rather than anode and cathode.
- **DIAC** — the same back-to-back pair of triangles with **no gate lead**. Two terminals only. Having no gate, it cannot be triggered by a control signal and instead conducts of its own accord once the voltage across it reaches its breakover value. Its usual job is to fire the gate of a TRIAC.
- **Bidirectional suppressor** — two Zener symbols drawn back to back, cathode to cathode. Whichever way the transient arrives, one device is forward biased and the other is in breakdown, so the pair clamps in both directions. This is the transient voltage suppressor found across avionics supply inputs.

Finding the Cathode on the Actual Component

Reading the symbol is only half the job; the device then has to be fitted the right way round, and the marking convention differs by package:



The cathode is the banded end of an axial diode, and the shorter lead and rim flat of an LED.

Package	How the cathode is identified	Trap
Axial glass or plastic body	A single band printed around the body at one end	Some older or high-voltage parts also carry several colour bands as a type code, and the conventions vary between makers and eras — do not read polarity from the band pattern on one of these; use the type number, or measure
Surface-mount	A printed bar or notch at one end, matching the bar on the board silkscreen	The marking is small and easily obscured by flux residue — clean and inspect before assuming

Package	How the cathode is identified	Trap
Stud-mounted power rectifier	Polarity is part of the type number, since either terminal can be made the stud	Two physically identical-looking studs may be opposite polarity; the part number is the only reliable check
LED	The **shorter** lead and the **flat** on the rim of the body are both the cathode	Leads get trimmed. The flat survives trimming; the lead length does not
Bridge rectifier module	Four marked terminals: two AC, one positive DC and one negative DC	The AC pair is interchangeable; the DC pair is not

When the marking is unreadable, measure rather than guess. A multimeter in diode-test mode identifies the cathode in one step: the leads that produce a reading of about 0.6 V have the meter's positive lead on the anode, so the cathode is the end touched by the black lead. This works on an unknown device as readily as on a marked one, and it is faster than looking for a part number under a coat of conformal coating.

The Arrow Is a Convention, and It Is Worth Being Precise About

The triangle points the way **conventional** current flows — from positive to negative through the external circuit and from anode to cathode inside the device. It does *not* point the way the electrons go; they travel the opposite way, entering at the cathode. Both descriptions are correct and they describe the same physical current; only the sign convention differs.

The same convention governs the arrows on every other semiconductor symbol an engineer will meet, so it is worth fixing here rather than relearning it later. Wherever an arrow appears on a device symbol, it marks a PN junction and it points from the P material toward the N material, which is the direction in which that junction conducts conventional current.

Diodes on an Aircraft Wiring Diagram

A schematic in a maintenance manual is drawn to a different set of conventions from a textbook, and three of them regularly cause difficulty:

- **Reference designators differ by drawing office.** European practice normally labels a diode **D** and may use **VR** or **ZD** for a Zener; much North American practice, including a great deal of aircraft documentation,

labels every semiconductor diode **CR**, for "crystal rectifier". Encountering CR12 on a wiring diagram and looking for a relay or a resistor is a common and time-consuming mistake.

- **The triangle may be filled or open.** Some standards draw it solid, others as an outline. The difference is purely stylistic and carries no electrical meaning; the orientation of the triangle relative to the bar is the only thing that matters.
- **The device may not be drawn at all.** Diodes built into a connector, a relay socket or a lamp assembly are often shown only as a note or as part of a block, so a suppression diode across a coil may exist in the aircraft and be almost invisible on the drawing. A low reading in both directions across such a coil is normal rather than a fault, because the meter is reading the copper winding, and it does not reveal whether that diode is fitted.

Where several diodes are combined in one physical part — a bridge module, a dual diode, a diode array in a connector shell — the schematic usually draws the individual junctions with a dashed outline around them showing the package boundary. That outline is the signal that the parts inside it cannot be replaced separately, which changes what the fault-isolation procedure can achieve: the module is proved good or bad as a unit and replaced as a unit.

What the Type Number Tells You

Two numbering schemes are in general use and they encode different amounts of information. Knowing which one is on the component decides how much can be deduced before opening a catalogue:

- **The American (JEDEC) scheme** uses a leading digit that counts *junctions*, followed by the letter N and a registration number. A leading **1N** therefore means a single-junction two-terminal device — a diode of some kind — and a leading 2N means a two-junction device, that is, a transistor. The number after the N is allocated sequentially as devices are registered and **encodes nothing at all**: a higher number does not mean a higher voltage, a higher current or a newer part.
- **The European (Pro-Electron) scheme** is more informative. The **first letter gives the material** — A for germanium, B for silicon, C for gallium arsenide — and the **second letter gives the function**. Among the second letters, A is a small-signal or detector diode, B a variable-capacitance diode, T a controlled rectifier, Y a power rectifier diode and Z a voltage-reference or Zener diode. A type beginning BZ is therefore a silicon Zener before any reference is consulted, and one beginning AA is a germanium signal diode.

Neither scheme carries the ratings, so the type number narrows the search and never ends it. In particular, no coding indicates the voltage of a Zener or the current rating of a rectifier — those live only in the suffix or the data sheet, and a device fitted from a bin of visually identical parts is a device of unknown rating.

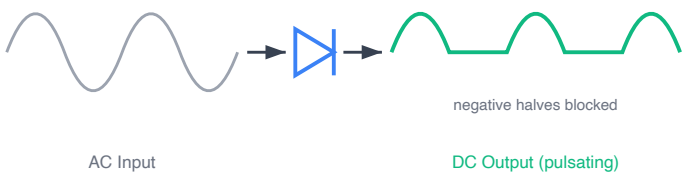
On an aircraft, the coding scheme is not the authority in any case. The part fitted is the part the illustrated parts catalogue calls for, and an electrically equivalent device from another scheme is not automatically an approved alternative. Type-number decoding is a diagnostic aid for understanding a circuit, not a basis for substitution.

Diodes in Rectifier Circuits

The most common application of diodes is **rectification** — converting alternating current (AC) to direct current (DC). This is essential in aircraft power supplies, battery chargers, and electronic systems.

Half-Wave Rectifier

A single diode passes only the positive half-cycles of the AC input and blocks the negative half-cycles. The output is pulsating DC with significant ripple.



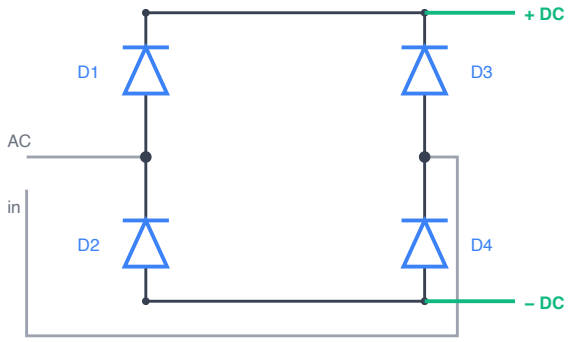
Disadvantages: Only 50% of the input energy is used. The ripple frequency equals the supply frequency. Not efficient for most applications.

Full-Wave Rectifier (Centre-Tapped Transformer)

Uses two diodes and a centre-tapped transformer. Each diode conducts on alternate half-cycles, so both halves of the AC input contribute to the output. The ripple frequency is double the supply frequency, making filtering easier.

Bridge Rectifier

The most common full-wave rectifier uses **four diodes** in a bridge arrangement. It does not require a centre-tapped transformer and is the standard circuit in most power supplies.



During the positive half-cycle, D1 and D4 conduct. During the negative half-cycle, D2 and D3 conduct. The load always receives current in the same direction.

✈ **Aviation context:** Bridge rectifiers are used in aircraft generator systems (transformer-rectifier units, or TRUs) to convert AC from the generators to 28 V DC for bus power. They are also found in voltage regulators and battery chargers.

The Three Circuits Compared in Numbers

Choosing between the three arrangements is a trade of diode count against transformer complexity, and the figures that decide it are all derived from one input peak voltage. Every quantity below assumes a sinusoidal input and ideal diodes, so that the comparison is between the topologies and not between components:

	Half-wave	Full-wave, centre-tapped	Bridge
Diodes	1	2	4
Transformer	Ordinary secondary	**Centre-tapped**, and only half of it is in use at a time	Ordinary secondary, fully used
Diode drops in the conduction path	One	One	**Two** — the current passes through two diodes in series
Average (DC) output	$0.318 \times V_{pk}$	$0.637 \times V_{pk}$	$0.637 \times V_{pk}$
Ripple frequency	Same as the supply	**Twice** the supply	**Twice** the supply
Ripple factor (unfiltered)	1.21	0.48	0.48
Form factor (unfiltered)	1.57	1.11	1.11
Peak inverse voltage per diode	V_{pk}	$2 \times V_{pk}$	V_{pk}

Two rows in that table decide most real designs. The **PIV row** is the reason the bridge is preferred at high voltage: the centre-tapped circuit stresses each diode with twice the peak secondary voltage, because the non-conducting diode has the whole secondary winding across it, while in a bridge the non-conducting diodes are held off by only one peak. The **diode-drop row** is the reason the centre-tapped circuit survives at low voltage: losing 1.4 V in a bridge matters little from a 115 V supply and matters a great deal from a 5 V one.

The **form factor** in that table is the ratio of the RMS value of a waveform to its average value, and it is a pure measure of how lumpy the output is. A perfectly smooth DC output would have a form factor of exactly 1, because its RMS and average values would be the same number. Full-wave rectification reaches 1.11 and half-wave only 1.57, which is the same conclusion the ripple factor reaches by a different route.

Worked Example — Output of a Bridge from a 400 Hz Aircraft Supply
A transformer secondary delivers 26 V RMS at 400 Hz into a bridge rectifier. Find the peak, the unfiltered average output, the ripple frequency and the PIV each diode must withstand.
Peak secondary voltage: $26 \times 1.414 = 36.8$ V
Peak available at the load, after two diode drops: $36.8 - 1.4 = 35.4$ V
Average unfiltered output: $0.637 \times 35.4 = 22.5$ V
Ripple frequency: $2 \times 400 = 800$ Hz
Peak inverse voltage on each diode: 36.8 V, so a 100 V part gives comfortable margin
Now add a capacitor filter and the average output changes completely, rising toward the peak value of 35.4 V rather than the 22.5 V average of the raw waveform. The 0.637 figure applies only to the unsmoothed waveform, and quoting it for a filtered supply is a standard mistake — roughly a 13 V error in this example.

Why the Ripple Frequency Is the Most Useful Number on the List

Doubling the ripple frequency does two good things at once. The gaps between peaks are halved, so a filter capacitor has less time to discharge and the ripple voltage falls for the same capacitance; and the higher frequency is itself easier to filter, because a capacitor's reactance falls as frequency rises. This is why the comparison above says filtering is easier with full-wave rectification — the benefit is arriving twice over.

It is also the single most useful diagnostic signal in a power supply, because **the ripple frequency identifies which rectification is actually happening, regardless of what the schematic says should be happening.** Put an oscilloscope on the output of a supply fed from 400 Hz:

- **800 Hz ripple** — full-wave rectification is taking place. All four bridge diodes are working.
- **400 Hz ripple** — only one half of each cycle is reaching the load. The circuit has degenerated into a half-wave rectifier, which in a bridge means **one diode has failed open circuit**. The output voltage will also have sagged and the ripple amplitude will have roughly doubled.

That single measurement separates a failed rectifier from a failed filter capacitor without disassembling anything. A tired capacitor raises the ripple *amplitude* but leaves the *frequency* alone; an open diode changes the frequency.

⚠ **The other bridge failure is far more violent.** A diode that fails *short* circuit puts the transformer secondary directly across the conducting pair on alternate half-cycles. The result is a fault current limited only by the winding resistance, and it usually takes out the remaining diodes, the fuse and sometimes the transformer with it. When one diode in a bridge is found short circuited, **all four are replaced** and the transformer is checked before power is reapplied — the survivors have been through the same fault current.

Smoothing: What the Filter Capacitor Actually Does

A rectifier alone produces pulsating DC — unidirectional, but nowhere near constant. The remaining alternating component is the **ripple**, and removing it is the filter's job. The simplest filter is a single capacitor across the output, and its operation is a two-part cycle:

- **Near each peak** the rectifier output is higher than the capacitor voltage, so a diode conducts and charges the capacitor rapidly up toward the peak.
- **Between peaks** the rectifier output falls below the capacitor voltage, every diode is reverse biased, and the capacitor alone supplies the load, drooping slowly as it discharges.

The output therefore sits close to the *peak* voltage rather than the average, with a sawtooth droop between peaks. For a full-wave supply the ripple voltage is approximately the load current divided by the product of the ripple frequency and the capacitance.

Worked Example — Sizing a Smoothing Capacitor

The bridge above supplies 1.0 A to its load, with 800 Hz ripple, through a 4700 μF capacitor. Find the ripple voltage, then the percentage ripple on a 35 V output.

Ripple voltage: $1.0 \div (800 \times 0.0047) = 1.0 \div 3.76 = 0.27$ V peak to peak

Percentage ripple: $(0.27 \div 35) \times 100 = 0.77\%$

Repeat at 50 Hz for comparison: $1.0 \div (100 \times 0.0047) = 2.1$ V, that is **eight times worse**

This is a large part of why aircraft use 400 Hz. The same capacitor is eight times more effective, or equivalently the same ripple can be achieved with a capacitor one eighth the size and weight. The saving compounds through the transformer and the smoothing choke as well.

Increasing the capacitance **reduces** the ripple, and the relationship is direct: double the capacitance and the ripple voltage halves. But there is a competing effect that stops the designer simply fitting the largest capacitor available. A bigger capacitor is charged in a shorter, sharper burst near each peak, so the **peak repetitive**

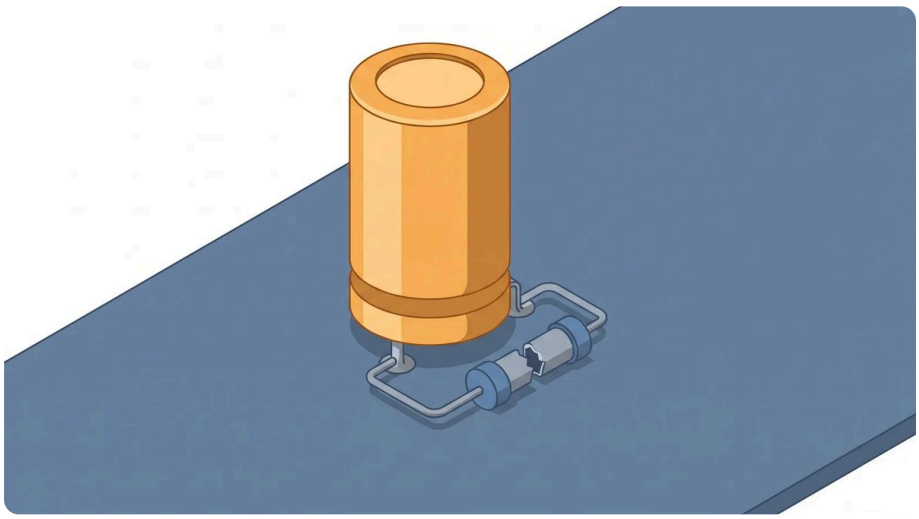
current through the diodes rises, and the inrush at switch-on rises with it. Beyond a certain point the rectifier and the transformer are being punished to buy a ripple improvement that a following regulator would have removed anyway.

Better Filters, and the Bleeder Resistor

Filter	Arrangement	Behaviour
Capacitor input	A single capacitor across the output	Output near the peak voltage; simple; poor regulation, since the output sags as load current rises; high peak diode current
Choke input	An inductor in series, then a capacitor across the output	The inductor opposes changes in current, so conduction is spread over the cycle. Lower peak diode current and better regulation, but the output settles nearer the average than the peak
Pi filter	Capacitor, then inductor in series, then a second capacitor — drawn as the Greek letter pi	The best ripple reduction of the three: the first capacitor takes the bulk of it, and the inductor and second capacitor form a second stage. Heavier and larger

A **bleeder resistor** is often fitted permanently across the filter capacitor. It serves two purposes and the safety one is the reason it is mandatory in high-voltage equipment: without it, a large capacitor holds its charge for a long time after the supply is removed, and the first person to put a hand inside the unit completes the circuit. The bleeder guarantees the capacitor is discharged within a defined time. Its second function is to keep a minimum load on the supply at all times, which prevents the output from rising to the full unloaded peak voltage and improves regulation.

⚠ **Never assume a bleeder resistor has done its job.** A bleeder that has failed open circuit gives no indication of any kind, and the equipment works perfectly with it missing. Filter capacitors are proved discharged by measurement before any work inside the unit begins, every time, regardless of how long the equipment has been isolated.



A bleeder that fails open across the filter capacitor gives no sign, so the capacitor may stay charged.

Regulation: the Number That Describes a Supply's Stiffness

Regulation is the ability of a supply to hold its output voltage constant when the load current or the input voltage changes, and it is quoted as the percentage by which the output falls between no load and full load. A perfect supply would have zero per cent regulation; a capacitor-input rectifier on its own is typically poor, because the output starts at the peak voltage with no load and droops as soon as current is drawn.

Filtering and regulation are separate jobs and are done by separate parts of the circuit. A filter removes the ripple that is *already* in the waveform; a regulator holds the average value steady against changes that the filter cannot see. A rectifier and filter alone can be quiet and still drift; a regulator following them is what makes the output both quiet and fixed. In practice that regulator is either a Zener diode with a series resistor for small loads, as the section on special diode types below describes, or a three-terminal regulator device that accepts an unregulated higher input and delivers a fixed output at useful current.

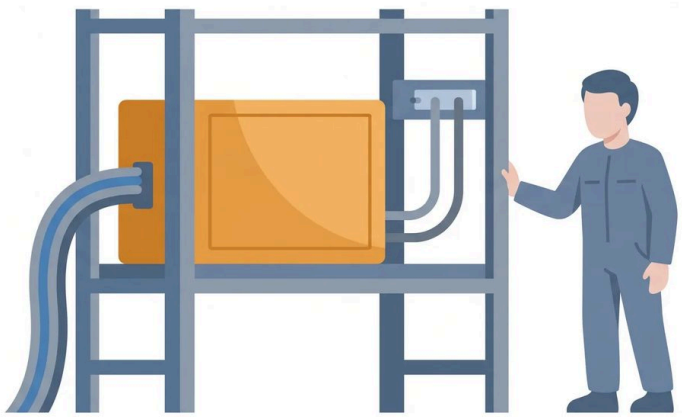
Finally, note that the same diode-and-capacitor combination appears elsewhere doing a different job. A **peak detector** is a rectifier and capacitor arranged to charge to the peak of an input signal and hold it, used to measure a signal's amplitude rather than to power anything. And where a signal is smaller than a diode's forward drop and would not rectify at all, a **precision rectifier** places the diode inside the feedback loop of an operational amplifier, which effectively divides the forward drop by the amplifier's gain — a technique that belongs with the amplifier material rather than here, but which is worth knowing exists.

Three-Phase Rectification and the Aircraft TRU

Aircraft main power is normally 115 V, 400 Hz, three-phase, and rectifying all three phases at once is markedly better than rectifying one. A three-phase bridge uses six diodes and produces **six output pulses per input**

cycle, so from a 400 Hz supply the ripple appears at 2400 Hz. Two consequences follow: the ripple amplitude before any filtering at all is already small, because the output is never far below the peak; and what remains is at a frequency that a modest filter removes easily.

Larger transformer-rectifier units go further and use two three-phase bridges fed from secondary windings connected in star and in delta, whose outputs are offset from one another. The combination produces twelve pulses per cycle instead of six, halving the ripple again and reducing the harmonic current drawn back into the aircraft's AC system — which matters, because that harmonic current is a load on the generators and a source of interference in its own right.



A transformer-rectifier unit takes three-phase AC from the generators and feeds DC to the aircraft bus.

The Filter Changes the Diode's Reverse Stress

The peak inverse voltage figures given in the comparison above are for an unfiltered output. Adding a smoothing capacitor changes one of them, and it changes it by a factor of two — enough to destroy a diode that was correctly chosen for the circuit without the capacitor.

Follow the half-wave case through at its worst instant. The capacitor has charged to the positive peak and holds it. Half a cycle later the transformer terminal has swung to the *negative* peak. The diode now has the capacitor's stored positive peak on one side and the negative peak on the other, so the reverse voltage across it is the sum of the two:

- **Half-wave with a capacitor-input filter: PIV becomes $2 \times V_{pk}$**, not V_{pk}.
- **Bridge with a capacitor-input filter: PIV stays at V_{pk}**, because the non-conducting diodes are clamped by the conducting pair and never see more than one peak.

- **Centre-tapped with a capacitor-input filter: PIV stays at** $2 \times V_{pk}$, unchanged, since it was already the sum of the two half-windings.

This is one more reason the bridge dominates in practice: it is the only one of the three whose diodes never see more than one peak of reverse voltage, filtered or not.

Worked Example — The Same Transformer, Two Circuits

A transformer has a 52 V RMS secondary with a centre tap, so each half delivers 26 V RMS. Compare a centre-tapped full-wave rectifier using both halves against a bridge rectifier connected across the whole winding.

Centre-tapped. Peak per half-winding: $26 \times 1.414 = 36.8$ V. One diode drop in the path, so the peak at the load is $36.8 - 0.7 = 36.1$ V, and the unfiltered average is $0.637 \times 36.1 = 23.0$ V. PIV on each diode: $2 \times 36.8 = 73.6$ V.

Bridge across the whole winding. Peak: $52 \times 1.414 = 73.5$ V. Two diode drops, so $73.5 - 1.4 = 72.1$ V at the load, and the unfiltered average is $0.637 \times 72.1 = 45.9$ V. PIV on each diode: 73.5 V.

The bridge delivers almost exactly twice the output voltage from the same transformer — because the centre-tapped circuit only ever uses half the winding at a time while the bridge uses all of it on every half-cycle. It costs two extra diodes and one extra diode drop. Note also that the PIV per diode came out almost identical in the two cases here, which is not a coincidence but a general result — twice the peak of half a winding is the peak of the whole winding, so one centre-tapped secondary used either way stresses its diodes equally, the 0.1 V difference here being nothing but rounding. The comparison that matters is always against the voltage each circuit is actually producing, and per volt of output the centre-tapped circuit stresses its diodes twice as hard.

Rectification Efficiency

A separate figure of merit answers the question of how much of the AC power arriving at the rectifier emerges as useful DC power in the load. For a resistive load and ideal diodes the theoretical maxima are **40.6 per cent for half-wave** and **81.2 per cent for full-wave**, whether the full-wave circuit is centre-tapped or a bridge.

Two cautions go with those numbers. They are *ideal* figures with no diode drop and no winding resistance included, so a real circuit does worse — the worked calculation of rectifier losses earlier in this note shows a bridge losing several per cent in the diodes alone. And they apply to a resistive load: adding a capacitor-input filter changes the current waveform completely, so the figures no longer describe what is happening. They are best used as they are intended, as a statement that full-wave rectification extracts almost exactly twice as much useful DC from the same waveform as half-wave does.

Clippers, Clampers, and Voltage Multipliers

Clippers (Limiters)

A **clipper** circuit removes (clips) a portion of a waveform above or below a certain level. A diode conducts when the signal exceeds the threshold, diverting the excess away from the load.

- **Positive clipper** — removes the positive portion above a set voltage
- **Negative clipper** — removes the negative portion below a set voltage
- **Biased clipper** — a DC voltage in series with the diode sets the clipping level

Clippers are used for waveform shaping and protecting sensitive circuits from voltage spikes.

Clampers (DC Restorers)

A **clamper** shifts the entire waveform up or down by adding a DC offset, without changing its shape. It consists of a capacitor and a diode. The capacitor charges to the peak value, then adds this as a DC level.

- **Positive clamper** — shifts the waveform upward so its negative peak sits at 0 V (or another reference)
- **Negative clamper** — shifts the waveform downward

Voltage Doublers and Triplers

Voltage **multiplier** circuits use combinations of diodes and capacitors to produce a DC output that is a multiple of the AC peak input voltage. A **voltage doubler** produces approximately $2 \times V_{peak}$; a **voltage tripler** produces approximately $3 \times V_{peak}$. They are used where high DC voltages are needed from a low AC supply, at relatively low current.

Series and Shunt Clippers Are Not the Same Circuit

Two arrangements both go by the name clipper and they work in different ways, so it is worth separating them before looking at the levels:

- **Series clipper** — the diode sits *in line* between source and load. It simply refuses to pass one polarity, so the load sees nothing at all during those half-cycles. This is the same circuit as a half-wave rectifier; the only difference is that it is being used to shape a signal rather than to make DC.

- **Shunt clipper** — the diode sits *across* the load, with a resistor in series with the source. When the signal reaches the level at which the diode conducts, the diode becomes a near short circuit and the series resistor takes up the surplus voltage, so the load voltage stops rising. The excess is diverted, exactly as the description above says.

The shunt form is far more common in practice, because the series resistor gives it somewhere to put the diverted energy. A shunt clipper without that resistor is a short circuit across the source, and the diode is the component that discovers this.

Where the Clipping Level Actually Sits

A plain shunt clipper with its diode returned to ground does not clip at 0 V. The diode has to be conducting to do anything, and a conducting silicon diode has about 0.7 V across it, so the level is 0.7 V beyond ground — a small offset, but one that shows up immediately on an oscilloscope and is a routine examination point.

Adding a DC source in series with the diode moves the level, and the two voltages simply add:

Worked Example — A Biased Clipper

A 10 V peak sine wave is applied through a series resistor to a shunt clipper. The diode has its **anode at the output node and its cathode returned to a +5 V supply**. What does the output look like?

The diode conducts only when its anode is above its cathode by the barrier voltage, that is, when the output tries to exceed $5 + 0.7 = 5.7$ V.

Positive half-cycles are therefore flattened at **+5.7 V** instead of reaching +10 V.

Negative half-cycles reverse-bias the diode, which does nothing, so they pass through **unaltered to –10 V**.

Reverse the diode and return it to –5 V instead and the mirror image results: negative peaks flattened at –5.7 V and positive peaks untouched. Note that the bias supply sets the level and the diode's orientation decides which side of the waveform is affected — two independent choices, and questions routinely change one without the other.

Putting two biased clippers back to back — one for each polarity — produces a **double-ended limiter** that flattens both peaks. Drive it hard with a large sine wave and the output approximates a square wave, which is one of the standard ways of generating one. Two Zener diodes connected cathode to cathode do the same job with fewer parts: whichever way the input swings, one device is in breakdown and the other is forward biased in series with it, so the pair clips symmetrically at the Zener voltage plus one forward drop.

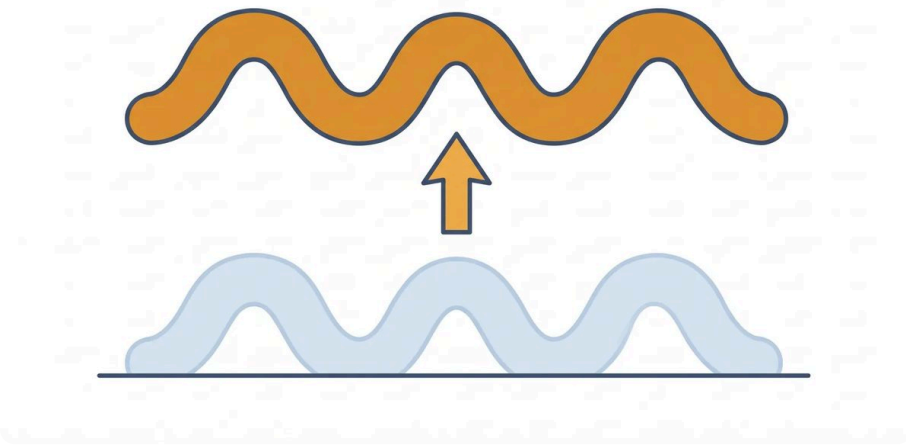
The most important application is not waveform shaping at all but **input protection**. A pair of diodes from a sensitive input to the positive and negative supply rails will conduct the instant the input strays beyond either rail, holding the input pin within about 0.7 V of the supply no matter what arrives on the wire. Almost every integrated circuit contains such a pair on every pin.

How a Clamper Moves a Waveform Without Distorting It

A clamper looks similar to a clipper — a diode and one passive part — but it does something categorically different, and the distinction is the point of the two circuits:

	Clipper	Clamper
Passive component	Resistor	**Capacitor**
Effect on waveform shape	Alters it — part of the waveform is removed	**Preserves it exactly**
Effect on peak-to-peak amplitude	Reduces it	**Unchanged**
Effect on DC level	Changes as a side effect	**This is the whole purpose**
Other name	Limiter	DC restorer

The mechanism runs in two stages. On the first excursion that forward-biases the diode, the capacitor charges rapidly through it — the diode's forward resistance is very low, so this takes almost no time. Once charged, the capacitor holds that voltage, and from then on the diode stays reverse biased for virtually the whole cycle. The charged capacitor is now simply a DC source sitting in series with the signal, adding a fixed offset to everything that passes through. Because the offset is constant, every point of the waveform moves by the same amount and the shape survives untouched.



A clamper adds a fixed DC offset, so every point of the waveform moves equally and the shape survives.

Working out which way a given circuit shifts the waveform takes one question: **which excursion turns the diode on?** Whichever it is, that excursion is the one that gets pinned near zero, and the rest of the waveform is pushed to the other side of it.

Worked Example — Reading a Clamper from the Diode's Orientation

A 10 V peak sine wave passes through a series capacitor. A diode is connected from the output node to ground with its **anode at ground and its cathode at the output node**. Which way does the waveform move?

The anode is at 0 V, so the diode conducts whenever the cathode — the output node — goes **below** ground.

The output is therefore prevented from going appreciably negative: its **negative peak is held near 0 V**.

With a 20 V peak-to-peak input and the bottom pinned at 0 V, the output swings from **0 V to +20 V**.

Average value before: 0 V. After: **+10 V**. Peak-to-peak before and after: 20 V, unchanged.

The waveform has been shifted upward, so this is a positive clamper — and it is identified by the fact that the diode's anode is the terminal connected to ground. Turn the diode round, putting the cathode to ground, and it conducts on positive excursions instead: the positive peak is pinned near 0 V, the output swings from 0 V down to –20 V, and the circuit is a negative clamper.

The Clamper's Time Constant

The clamp holds only for as long as the capacitor holds its charge, so the discharge path matters. Between the brief conduction pulses, the capacitor can only discharge through whatever resistance is across it — the load, any deliberate resistor, and the diode's own reverse resistance.

The requirement is that the capacitor must lose almost none of its charge between one conduction pulse and the next, which means the time constant has to be **long compared with the period** of the signal — a factor of ten as an absolute minimum and usually far more. If it is too short, the capacitor droops noticeably between pulses and the flat parts of the waveform develop a visible tilt, a distortion known as sag.

Removing the resistor entirely takes this to its limit. The only remaining discharge path is the diode's reverse leakage, which is an extremely high resistance, so the time constant becomes very long indeed and the clamping level is held almost perfectly. That is why a clamper intended for a low-frequency or intermittent signal is built with as little parallel resistance as the following stage will allow.

The classic use is in the name — **DC restoration**. Passing a signal through a coupling capacitor strips out its DC content and leaves it centred on zero, which is a problem whenever the DC level carried information. A clamper puts a known DC level back, referenced to a chosen part of the waveform rather than to its average, so the reference stays correct even when the waveform's duty cycle changes.

Inside a Voltage Doubler

A multiplier is a clamper and a peak detector working in series, and seeing it that way makes the arithmetic obvious rather than magical. Take the half-wave doubler, which uses two diodes and two capacitors:

- **On the first half-cycle**, one diode conducts and charges the first capacitor to the peak input voltage. That capacitor is now a DC source of V_{pk} sitting in series with the input — a clamper.
- **On the next half-cycle**, the input swings the other way and *adds* to the voltage already stored on the first capacitor, so the second diode sees $2 \times V_{pk}$ and charges the second capacitor to that value — a peak detector.
- **The output** is taken across the second capacitor and sits at approximately twice the peak input.

The full-wave doubler arranges things differently: two capacitors are charged on opposite half-cycles, each to the peak value, and the output is taken across the two *in series*. It reaches the same doubled voltage but draws from both halves of the cycle, so its ripple is at twice the frequency and is easier to filter. A tripler and a quadrupler are simply further identical stages cascaded onto the same chain, each adding another V_{pk} to the total.

⚠ **A multiplier multiplies voltage, not power.** The energy has to come from the source, so a doubler that produces twice the voltage can supply at best half the current, and in practice much less. These circuits have a very high output impedance: the output voltage collapses as soon as any real current is drawn, and the quoted multiple is only reached at almost no load. They belong where a few hundred volts at microamperes is wanted from a small supply and nowhere near a power circuit. Note also that the diodes in a doubler must be rated for at least twice the peak input voltage, not once.

ℹ **A charged multiplier is a hazard long after power is removed.** The stack of capacitors holds the full multiplied voltage, the load that would normally discharge it draws almost nothing by design, and there may be no bleeder path at all. Treat any high-voltage multiplier as live until each capacitor has been individually proved discharged.



A voltage multiplier stays charged after power is removed, so each capacitor is proved discharged in turn.

The Clamped Level Is Not Exactly Zero

The worked example above gave the output as swinging from 0 V to +20 V, which is the idealised answer. The real one is displaced by a diode drop, and knowing which way it goes is worth a mark.

The diode only conducts once it has about 0.7 V across it, so it cannot hold the waveform at exactly 0 V — it lets the peak go 0.7 V past the reference before it starts to clamp. In the positive clamper of the example, whose diode conducts when the output goes negative, the negative peak therefore settles at about **–0.7 V** rather than 0 V, and the whole waveform sits 0.7 V lower than the ideal analysis predicts. The negative clamper is displaced by the same amount in the opposite direction, with its positive peak at about +0.7 V.

Reference levels other than ground are obtained by putting a DC source in series with the diode, exactly as in a biased clipper, and the levels simply add. A positive clamper with a +3 V source in series with its diode holds the negative peak at about +2.3 V — the 3 V reference less the diode drop — and a 20 V peak-to-peak input then swings from +2.3 V to +22.3 V. Note that the peak-to-peak value is untouched at 20 V no matter what the bias is: a clamper moves a waveform and never stretches it.

Clipping With the Diode in Series

Worked Example — Series Clipper Against Shunt Clipper

A 10 V peak sine wave is applied to a load through a diode whose anode faces the source. What reaches the load, and how does it differ from the shunt clipper examined earlier?

Positive half-cycles: the diode is forward biased and conducts, so the load sees the input less one diode drop — peaks reach $10 - 0.7 = 9.3$ V.

Negative half-cycles: the diode is reverse biased, no current flows and the load sees **0 V** throughout.

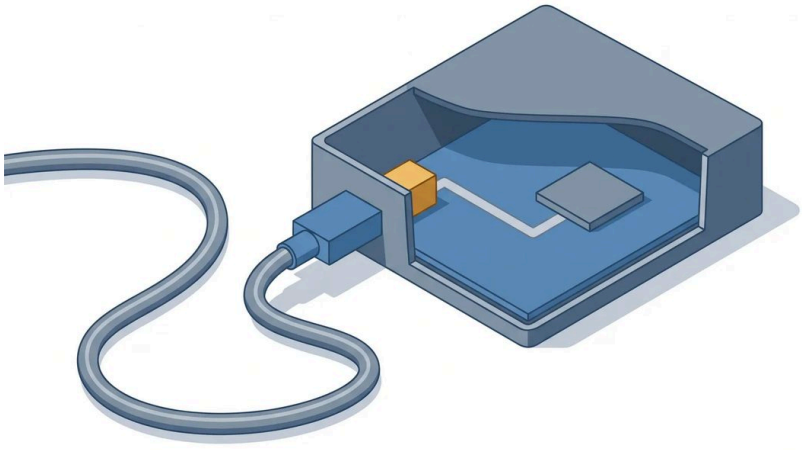
Output: a series of positive humps separated by flat gaps — which is exactly the half-wave rectified waveform shown earlier in this note.

The contrast with the shunt clipper is the point. The shunt circuit removed only the part of the waveform beyond its threshold and passed everything else unaltered, including the whole of the other half-cycle. The series circuit removes an entire half-cycle and cannot be set to a threshold at all without adding a bias source. A shunt clipper limits; a series clipper gates. Choosing between them is a question of whether the surviving part of the waveform must arrive intact.

Where Clippers Earn Their Place

- **Noise limiting.** In a communications receiver, impulse noise — ignition interference, static discharge, switching transients — arrives as short spikes far larger than the wanted signal. A clipper set just above the peak of the wanted audio removes the tops of those spikes, and the result is far less objectionable to listen to than the spikes themselves, even though the clipping distorts the signal slightly.

- **Squaring a waveform.** Driving a double-ended limiter hard flattens both peaks and produces something close to a square wave from a sine, which is the classic simple way to derive a timing edge from an oscillator.
- **Protecting a meter movement.** A pair of inverse-parallel diodes across a sensitive moving-coil movement conducts on any overload and diverts the excess, while doing nothing at all at normal indication levels because the voltage across the movement is well below a diode drop.
- **Defining a logic level.** Clipping an incoming signal to the supply rails guarantees that whatever arrives on a long cable stays inside the range the receiving circuit can accept.



A signal arriving on a long cable is clipped to the supply rails before it reaches the receiving circuit.

Recognising These Circuits on a Diagram

Examination questions frequently show a small circuit and ask what it does. All the circuits in the table below are built from diodes, capacitors and resistors, so the identification comes from *how* they are arranged rather than from what parts are present. Three checks settle it:

Look for	What it indicates
A capacitor **in series** with the signal path, with a diode across the line after it	**Clamper** — the series capacitor is the giveaway, but count them — a clamper has exactly one, while a doubler or multiplier also puts one across the output
A resistor in series and a diode across the line, with no series capacitor	**Shunt clipper**
A single diode in line, with the load and any capacitor across the output only	**Half-wave rectifier** — or a series clipper, which is the same circuit used for a different purpose
Four diodes in a closed loop with AC on one diagonal and the output on the other	**Bridge rectifier**

Look for	What it indicates
Two diodes and **two capacitors**, one capacitor in series with the input and one across the output	**Voltage doubler** — count the capacitors, because a rectifier has at most one
A repeating ladder of diodes and capacitors	**Voltage multiplier**, and the number of stages gives the multiple

The most reliable single discriminator is the **capacitor count and position**. A rectifier has no capacitor in the signal path and at most one across the output; a clamper has exactly one, in series; a doubler has two, one in each position. Counting diodes alone is not enough, because a half-wave rectifier and a series clipper have one apiece and a doubler and a centre-tapped rectifier have two.

Special Diode Types

Zener Diode

A Zener diode is designed to operate in **reverse breakdown** at a specific, stable voltage (the Zener voltage, V_Z). When reverse voltage reaches V_Z, the diode conducts freely in reverse, maintaining a near-constant voltage across itself regardless of current variations.

Primary use: Voltage regulation. A Zener diode in parallel with a load provides a stable reference voltage even when the input voltage or load current fluctuates. Common Zener voltages: 3.3 V, 5.1 V, 6.8 V, 12 V, 15 V, 24 V.

Worked Example — Zener Regulator

A 5.1 V Zener diode with a 470 Ω series resistor is fed from a 12 V supply. What is the current through the resistor and the voltage across the load?

Voltage across resistor: $V_R = 12 - 5.1 = 6.9$ V

Current: $I = \frac{6.9}{470} = \mathbf{14.7}\ \mathrm{mA}$

Voltage across load: **5.1 V** (held constant by the Zener)

Silicon Controlled Rectifier (SCR / Thyristor)

An SCR is a four-layer (PNPN) device with three terminals: **anode, cathode, and gate**. It acts as a latching switch:

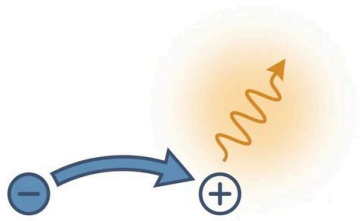
- It remains OFF (non-conducting) until a small current pulse is applied to the gate

- Once triggered, it latches ON and conducts like a normal diode (gate loses control)
- It turns OFF only when the anode current falls below a minimum value (the **holding current**) — typically at the zero crossing of an AC waveform

Uses: Power control, motor speed control, lighting dimmers, overvoltage protection (crowbar circuits). In aviation, SCRs are used in TRUs and generator control units (GCUs).

Light Emitting Diode (LED)

An LED emits light when forward-biased. Electrons recombining with holes release energy as photons. The colour depends on the semiconductor material:



When an electron recombines with a hole, the energy it gives up leaves as a photon of light.

Colour	Material	Forward Voltage
Red	GaAsP	~1.8 V
Green	GaP	~2.2 V
Blue	InGaN	~3.3 V
White	Blue LED + phosphor	~3.3 V

Aviation uses: Cockpit indicator lights, annunciator panels, position lights (navigation lights), landing lights on modern aircraft, panel backlighting. LEDs offer long life, low heat, and high reliability compared to incandescent bulbs.

Schottky Diode

Uses a metal-semiconductor junction instead of a PN junction. Key advantages:

- Very low forward voltage drop (~0.2–0.3 V vs 0.6 V for standard silicon)
- Very fast switching speed (no minority carrier storage)
- Used in high-frequency circuits, switching power supplies, and as clamping/protection diodes

Photodiode

A photodiode is operated in **reverse bias**. When light falls on the junction, it generates electron-hole pairs, increasing the reverse (leakage) current in proportion to light intensity. Used in optical sensors, smoke detectors, fibre-optic receivers, and light meters.

Varactor (Varicap) Diode

A varactor is operated in **reverse bias**. The width of the depletion region (and therefore its capacitance) changes with the applied reverse voltage. Used as a voltage-controlled capacitor in automatic frequency control (AFC) circuits, FM modulators, tuners, and phase-locked loops (PLLs).

Varistor (MOV — Metal Oxide Varistor)

A varistor is a voltage-dependent resistor. At normal voltages, its resistance is very high (effectively open circuit). When the voltage exceeds a threshold (the **clamping voltage**), its resistance drops sharply and it conducts, diverting the surge energy safely to ground.

Use: Transient voltage suppression. Protects sensitive electronic circuits from voltage spikes caused by lightning, switching, or static discharge. Essential in aircraft avionics protection.

Designing the Zener Regulator: What the Series Resistor Has to Satisfy

The regulator above works because the Zener holds its terminal voltage constant and the **series resistor absorbs everything else**. That resistor is doing two jobs at once, and both set limits on its value:

- It drops the difference between the unregulated input and the regulated output, and that difference changes whenever the input changes.

- It limits the current, and therefore the power, in the Zener — without it the diode is directly across the supply in breakdown and is destroyed.

The current the resistor delivers is fixed by the voltages across it, so it does not change when the load changes. What changes is how that fixed current is *shared*: the Zener passes whatever the load does not take. That single sentence is the whole of the circuit's behaviour, and both of its failure modes follow from it.

Worked Example — Worst Case for the Same Regulator

Take the 12 V supply, 5.1 V Zener and 470 Ω resistor from the calculation above, which delivers 14.7 mA. Two worst cases have to be checked, and they are at opposite ends of the load range.

Load disconnected. All 14.7 mA now flows in the Zener. Its dissipation is $5.1 \times 0.0147 = 0.075$ W, that is 75 mW.

Input rises to 15 V with the load still disconnected. Current becomes $(15 - 5.1) \div 470 = 0.0211$ A and the dissipation is $5.1 \times 0.0211 = 0.107$ W, that is 107 mW. A 400 mW Zener is comfortable.

Load drawing 10 mA. The Zener takes the remaining $14.7 - 10 = 4.7$ mA and regulation is maintained.

Load drawing 16 mA. The resistor can only supply 14.7 mA, so there is nothing left for the Zener. It comes out of breakdown, stops regulating, and the output falls below 5.1 V and follows the load.

The maximum Zener dissipation occurs at maximum input voltage with the load removed, and the minimum output occurs at minimum input voltage with maximum load. Checking only the nominal condition passes a design that fails at both ends.

A Zener regulator that has lost regulation gives no warning. The output simply becomes a resistive divider and drifts with load, which on a meter looks like a low reading rather than a fault. If a Zener-regulated rail is found low, measure the current in the series resistor before condemning the diode: an overloaded output and a failed Zener produce the same symptom, and only the current distinguishes them.

Two refinements are worth knowing. The Zener is not a perfect voltage source — it has a small dynamic resistance, typically a few ohms, so its terminal voltage does creep up slightly as its current increases. And its temperature coefficient can be cancelled: because the coefficient is positive for Zener voltages above about 5.6 V and an ordinary forward-biased silicon junction has a coefficient of about –2 mV per degree Celsius, putting a plain diode in series with the Zener makes the two drifts oppose each other. That is how a temperature-compensated reference is built.

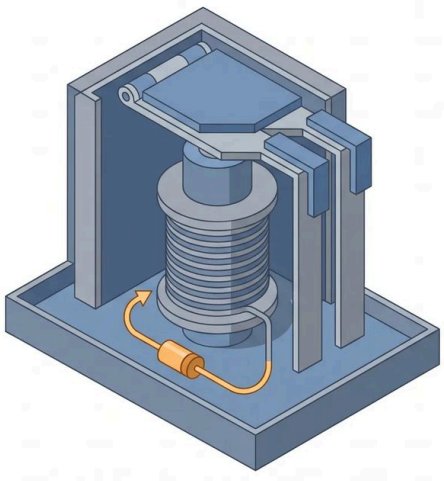
Vocabulary: voltage regulation and **voltage stabilisation** mean the same thing — holding an output voltage steady against changes in input or load. Different textbooks and different examination papers prefer one word or the other, and a question offering both is not asking you to separate them.

Other Jobs the Zener Does

- **Voltage reference.** A regulator supplies current; a reference only has to be accurate. A lightly loaded Zener in the 5 to 6 V range, where the two breakdown mechanisms cancel, is the classic stable reference against which a larger regulator compares its output.
- **Limiter.** Two Zeners connected cathode to cathode across a signal clip it symmetrically at the Zener voltage plus a forward drop — the double-ended limiter of the previous section, built with two components.
- **Overvoltage protection.** A Zener sensing an output rail can fire the gate of a thyristor which then short-circuits the supply, blowing the fuse deliberately. Because it destroys the fuse to save the equipment, this arrangement is called a **crowbar**.
- **Level shifting.** Placed in series with a signal path, a Zener subtracts a fixed voltage, moving a signal into the range a following stage can accept.

Suppressing an Inductive Spike

Switching off a relay, contactor or solenoid produces a large reverse voltage spike, because the collapsing magnetic field drives the coil current to keep flowing and the coil will generate whatever voltage it takes to do so. Left unchecked that spike destroys the switching transistor and radiates interference across the aircraft. A diode connected across the coil solves it, and there are two versions with a real difference between them:



At switch-off, a diode across the relay coil lets the coil current circulate and decay instead of spiking.

	Plain freewheel (flyback) diode	Zener or bidirectional suppressor
Connection	Across the coil, reverse biased in normal operation so it does nothing while the coil is energised	Across the coil or across the switching device, chosen to break down above the normal supply voltage

	Plain freewheel (flyback) diode	Zener or bidirectional suppressor
What it does at switch-off	Provides a path for the coil current to circulate and decay, clamping the spike to about one diode drop above the supply	Clamps the spike at the breakdown voltage instead — higher, but still safely below the switching device's rating
Effect on release time	**Slower.** Only about 0.7 V opposes the circulating current, so the field decays gently and the contacts release later	**Faster.** A much larger voltage opposes the current, so the field collapses quickly
Where it is preferred	Where release timing does not matter and simplicity does	Where the relay must drop out promptly, and on AC circuits, where a bidirectional device is needed because there is no fixed polarity

Note the competing effect in that table, because it is the reason both types exist. A lower clamping voltage protects the switching device better but slows the release; a higher clamping voltage releases faster but demands a switching device that can stand the higher voltage. The designer picks the compromise, and fitting the wrong one in the field changes the relay's timing without producing any other symptom.

Inside the SCR: the Two-Transistor Model

The latching behaviour described above is not a property that has to be taken on trust; it falls out of the four-layer structure once the structure is redrawn. Split the PNPN sandwich down the middle and it becomes two interconnected transistors — a **PNP and an NPN** — with each one's collector feeding the other's base. That cross-connection is positive feedback, and it explains every part of the device's behaviour:

- **Why a pulse is enough.** The gate current starts the NPN conducting. Its collector current becomes the PNP's base current, turning the PNP on. The PNP's collector current then feeds the NPN's base — which is where the gate current came from. The pair now sustains each other and the gate has become irrelevant.
- **Why the gate cannot turn it off.** The NPN is being driven by the PNP, not by the gate, so removing or even reversing the gate signal changes nothing. There is no path by which the gate can interrupt the internal loop.
- **Why the anode current has to be reduced instead.** The loop sustains itself only while enough current circulates to keep both transistors in conduction. Starve it and the feedback fails, both transistors turn off, and the device latches out.
- **Why the trigger pulse must be positive.** It has to forward-bias the NPN's base-emitter junction, which is the gate-to-cathode junction of the SCR. A negative pulse reverse-biases it and does nothing.

Two similarly named currents come out of this and they are not the same figure:

- **Latching current** — the minimum anode current that must be flowing at the moment the gate pulse is *removed*, for the device to stay on by itself. It is the larger of the two, because the internal loop has not yet fully established.
- **Holding current** — the minimum anode current needed to keep an already fully conducting SCR in conduction. Fall below it and the device turns off.

Turning an SCR off therefore means getting the anode current below the holding value, and there are only a few ways to do it: let the AC waveform pass through its natural zero crossing, open the supply, reverse the anode voltage so the anode goes negative with respect to the cathode, or divert the current around the device with a switched capacitor. In an AC circuit the first of these happens automatically twice per cycle, which is why SCR phase control is so much simpler on AC than on DC.

⚠ **Two ways an SCR turns on when nobody asked it to.** If the forward anode voltage is raised beyond the device's **forward breakover voltage**, it conducts with no gate signal at all. And a fast enough *rate of rise* of anode voltage will do the same, because the internal junction capacitance couples the transient straight into the gate region — the effect known as dV/dt triggering. Both are why thyristor circuits carry snubber networks, and why an intermittent, unexplained turn-on in service points at supply transients rather than at the trigger circuit.

The Rest of the Thyristor Family

The SCR conducts in one direction only, which halves what it can do with alternating current. Three relatives fill the gaps, and a fourth item in this list is not a semiconductor at all but is put here because it appears alongside them in examination options:

Device	Terminals	Conducts	Triggered by	Turned off by	Typical job
SCR (thyristor)	Anode, cathode, gate	One direction only	A positive gate pulse with respect to the cathode	Anode current falling below the holding current	Controlled rectification, phase control, crowbar protection
TRIAC	Main terminal 1, main terminal 2, gate	**Both directions** — it behaves as two SCRs in **inverse parallel**	A gate pulse of either polarity	Main-terminal current falling below the holding current, which happens at every AC zero crossing	Full-wave AC power control, lamp dimming, motor speed control

Device	Terminals	Conducts	Triggered by	Turned off by	Typical job
DIAC	Two only — **no gate**	Both directions	Nothing external: it breaks over of its own accord once the voltage across it reaches its breakover value	Current falling below the holding value	Generating the sharp trigger pulse that fires a TRIAC's gate
GTO thyristor	Anode, cathode, gate	One direction only	A positive gate pulse, as an SCR	**A negative gate pulse** — unlike an ordinary SCR, the gate can switch it off	DC circuits, where there is no zero crossing to commutate an ordinary SCR
Variac	Three — and it is *not* a semiconductor	Not applicable	Not applicable	Not applicable	A variable autotransformer with a sliding contact, giving a cheap and convenient continuously variable AC supply for test purposes

The Variac row is included deliberately. Its name appears alongside DIAC and TRIAC in examination options, but it contains no junction and no semiconductor of any kind — it is a wound transformer whose output tapping is moved by a brush. When a question asks which device gives a cheap and convenient *variable power supply*, that is the Variac; when it asks about controlling power in *both directions of an AC waveform*, that is the TRIAC.

Thyristors are, above all, **current-handling** devices. The whole point of the latched state is that the conduction path is a fully saturated four-layer structure with a very low forward drop, so a physically small device can pass a large current with a modest power loss. In the off state it is the reverse — a very high resistance blocking the full circuit voltage — and it is that combination of high off-state resistance and low on-state resistance that makes it a switch rather than an amplifier. In aircraft systems they appear in transformer-rectifier units, generator control units and any circuit where a substantial current must be switched without a moving contact.

Why an LED's Colour Is Fixed by Its Material

An ordinary silicon diode passes exactly the same recombination current and emits no useful light at all, so the light cannot be a property of the junction as such. What happens on recombination is that the electron gives up an amount of energy equal to the **band gap** of the material it is in, and in the right materials that energy leaves as a photon. Photon energy determines wavelength, and wavelength is colour. It follows that:

- **The colour is set by the semiconductor compound and its doping** — the active element of the device. It is not set by the plastic encapsulation, which only tints, diffuses or focuses what is already there; and there is no gas inside an LED to be excited.
- **An LED emits a narrow band of wavelengths**, essentially one colour determined by its type, not the broad spectrum of a heated filament. Fitting a red lens to a green LED gives a dim green-blocked output, not a red one.
- **A wider band gap means a higher photon energy, hence a bluer colour and a higher forward voltage.** The forward voltage is the price of the colour and cannot be separated from it — which is why red and green types run at around 1.6 to 2.2 V while a blue or white one needs roughly 3.0 to 3.5 V.
- **Silicon cannot be used at all.** Its band structure causes recombination to release energy as lattice vibration instead of light. LEDs are built from compound semiconductors — gallium arsenide phosphide, gallium phosphide, indium gallium nitride and their relatives.

Because the characteristic is that of a diode, the forward current rises almost vertically once the threshold is passed. An LED therefore **must** have a current-limiting element in series; connecting one straight across a supply destroys it. Brightness is set by current, not by voltage, and where dimming is required it is normally done by pulsing the current on and off rapidly rather than by reducing it, because the colour of some materials shifts slightly with current.

Worked Example — An Indicator LED on a 28 V Bus

A red LED with a forward drop of 2.0 V is to run at about 15 mA from the 28 V DC bus. Find the series resistor and its power rating.

Voltage across the resistor: $28 - 2.0 = 26$ V

Required resistance: $26 \div 0.015 = 1733\ \Omega$, so use the next standard value up, 1.8 kΩ

Actual current: $26 \div 1800 = 0.0144$ A, that is **14.4 mA**

Resistor dissipation: $0.0144^2 \times 1800 = 0.373$ W, so a 0.5 W part is needed

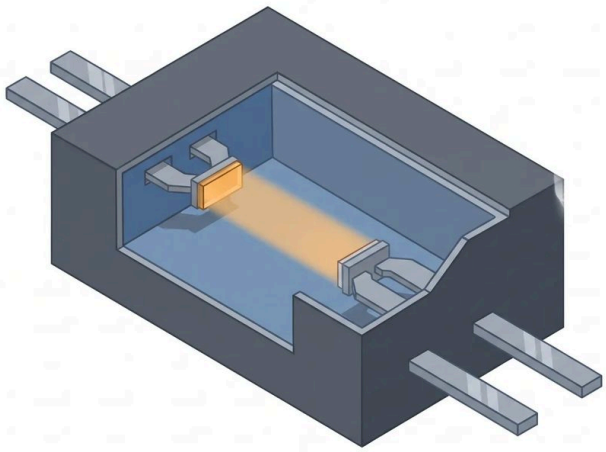
The resistor dissipates thirteen times as much power as the LED — 373 mW against $2.0 \times 0.0144 = 0.029$ W, that is 29 mW. This is why panel indicators on a 28 V bus are usually driven from a lower local rail or by a constant-current device: at 28 V, almost all the energy is wasted as heat in the resistor rather than as light.

⚠ **An LED's reverse breakdown voltage is very low** — a few volts only, far below that of an ordinary rectifier. Connect one across an AC supply and it is destroyed on the first negative half-cycle even though the forward half-cycles are within rating. The standard protection is an ordinary diode connected in **inverse parallel** with the LED: on the reverse half-cycle that diode conducts and holds the LED's reverse voltage to about 0.7 V, well inside its capability. Note what the protective diode is for — it guards against the reverse half of the AC waveform, not against excessive forward current, which remains the series resistor's job.

LEDs in Aircraft, and Two Devices Built Around Them

The advantages listed above — long life, low heat, high reliability — are joined by two that matter particularly in an airframe: an LED has no filament to fail under vibration, and it reaches full brightness instantly with no warm-up. Against that, an LED is a semiconductor and is therefore sensitive to over-temperature and to reverse transients in a way a filament lamp is not, and a failed LED usually goes open circuit and dark rather than dimming first, so there is no warning of impending failure.

Two composite devices are built from an LED and are examined as separate items:



An optocoupler: an LED and a photodetector face each other in one opaque package, with light the only link.

- **Optocoupler (optoisolator)** — an LED and a photodetector sealed facing each other in one opaque package. The signal crosses as light, so there is *no electrical connection whatever* between input and output. That is its entire purpose: transferring a signal while keeping the two circuits electrically isolated, which breaks earth loops, protects a low-voltage circuit from a high-voltage one, and stops interference on one side from reaching the other. It does not amplify.
- **Laser diode** — an LED-like junction built inside an optical cavity so that emission is **stimulated** rather than spontaneous. The output is **coherent and monochromatic**, where an ordinary LED's is incoherent and spread over a band of wavelengths. That coherence is what allows the beam to be focused tightly and modulated at very high rates, so laser diodes drive long-distance fibre-optic links while plain LEDs serve short ones.

The Photodiode's Two Operating Modes

The photodiode is the LED run backwards, in both senses: it converts light into electricity rather than the reverse, and it is used with the opposite bias. Light of sufficient photon energy falling on the depletion region

breaks covalent bonds there, and the junction field immediately sweeps the resulting electron-hole pairs apart before they can recombine. The result is a current proportional to the illumination, and it is added to the ordinary thermal leakage current.

That gives two distinct ways to use the device:

	Photoconductive mode	**Photovoltaic mode**
Bias	**Reverse biased**	**Zero bias** — no external supply at all
Output	A reverse current proportional to light intensity	A voltage across the terminals, as a solar cell produces
Depletion region	Wide, because reverse bias widens it, so more of the incoming light is absorbed where it is useful	At its natural width
Speed	**Fast** — the wide depletion region has low capacitance and a strong sweeping field	Slower
Dark current	Present, and it sets the sensitivity limit	Essentially none
Used for	Measurement and high-speed detection: fibre-optic receivers, smoke detectors, optical position sensors	Generating power, and very low-light measurement where dark current would swamp the signal

The reverse-biased arrangement is much the more common, and it is the one meant whenever a device is described as a **photoconductive diode** — the name refers to the mode, not to a different component. Note the wording carefully, because it is a favourite examination trap: the photodiode is *reverse* biased and it *conducts* when light falls on it. Both halves have to be right. A statement that it is reverse biased so as *not* to conduct in light, or that it is forward biased, is wrong on one half each.

Two figures describe how good a photodiode is, and they answer different questions. **Responsivity** is the output current divided by the incident light power, quoted in amperes per watt, and it tells the designer what signal to expect. **Quantum efficiency** is the number of electron-hole pairs generated per incident photon, and it tells how much of the arriving light is being used at all. A high responsivity can come from high quantum efficiency or simply from a favourable wavelength; the two are related but not interchangeable.

Where the available light is extremely weak, an **avalanche photodiode** is used instead. It is operated at a reverse voltage close to breakdown, so each photo-generated carrier is accelerated hard enough to ionise further atoms on impact — the avalanche mechanism described earlier, used deliberately as an amplifier. One absorbed photon produces many carriers, giving the device internal gain and making it far more sensitive than a plain photodiode. The price is a high, closely controlled bias supply and more noise.

A photoconductive cell is not a photodiode. The light-dependent resistor found in older ambient-light sensors is a slab of photosensitive material with no junction at all: light frees carriers throughout the bulk and its **resistance falls**. It has no polarity, generates nothing on its own, and responds slowly. A photodiode has a junction, is polarity sensitive, generates a current, and is fast. Questions distinguish them by the word used — a device that *changes its resistance* with light is the cell; one that *generates a current or a voltage* is the diode.

What the Schottky Trades Away for Its Speed

The advantages listed above come from one structural fact: with a metal on one side of the junction there is **no P-type region to inject holes into the N-type material**. Conduction is carried out entirely by majority carriers, so no minority-carrier charge is stored during forward conduction, and there is therefore nothing to remove when the voltage reverses. Reverse recovery is not merely fast — the mechanism that causes it does not exist.

That is not a free advantage, and knowing what is given up is what separates a correct answer from a plausible one:

	Schottky	Ordinary silicon PN rectifier
Forward voltage drop	**Lower** — roughly 0.2 to 0.3 V	About 0.7 V, rising above 1 V at full current
Reverse recovery	**Essentially none**	Present, and it limits the usable frequency
Reverse leakage current	**Higher**	Lower
Peak inverse voltage rating	**Lower**	Higher — ordinary rectifiers reach far greater reverse voltages
Behaviour at high temperature	Leakage climbs steeply, and thermal runaway is a real risk in a hot design	More tolerant

So the Schottky is chosen wherever the forward drop or the switching speed dominates, and rejected wherever the reverse voltage or the leakage does. In a switched-mode power supply it wins on both counts — the lower drop cuts conduction loss and the absent recovery cuts switching loss — which is why almost every low-voltage switching supply rectifies its output with one. Its other classic use is detecting very small signals: with a threshold of about 0.2 V it produces a usable output from an input that would not even reach the knee of a silicon junction. The alternative name **hot carrier diode** refers to the energetic majority carriers crossing the metal-semiconductor barrier.

A newer variant, the **silicon carbide Schottky**, removes most of the disadvantages: silicon carbide's much wider band gap gives a far higher breakdown voltage, much lower leakage and a considerably higher usable junction temperature, while keeping the absent reverse recovery. It is displacing silicon fast-recovery rectifiers in high-power converters, at a higher purchase price.

Varactor Behaviour, Stated with Its Direction

The varactor's capacitance and its reverse bias move in **opposite** directions, and this is the single fact the exam tests about the device. The chain of reasoning was laid out with the junction earlier: reverse bias widens the depletion region; the depletion region is the dielectric; a capacitor with a wider gap between its plates has less capacitance. So:

- **To decrease the capacitance, increase the reverse bias.**
- **To increase the capacitance, decrease the reverse bias.**

Check it against the end of the range: reduce the reverse bias to zero and the depletion region is at its narrowest, so the capacitance is at its maximum. That agrees. The relationship is *not* a straight line — capacitance falls off with the square root or cube root of the applied voltage depending on how the junction is graded — which is why the tuning of a varactor-controlled oscillator is never linear with control voltage and has to be compensated in the circuit around it.

⚠ **A varactor is never forward biased in service.** Forward bias collapses the depletion region, destroys the capacitance effect entirely and lets a large current flow through a device that is not built to carry one. In a tuning circuit this shows up as an oscillator that stops or jumps frequency when the control voltage strays too close to zero, and it is a design fault rather than a component failure.

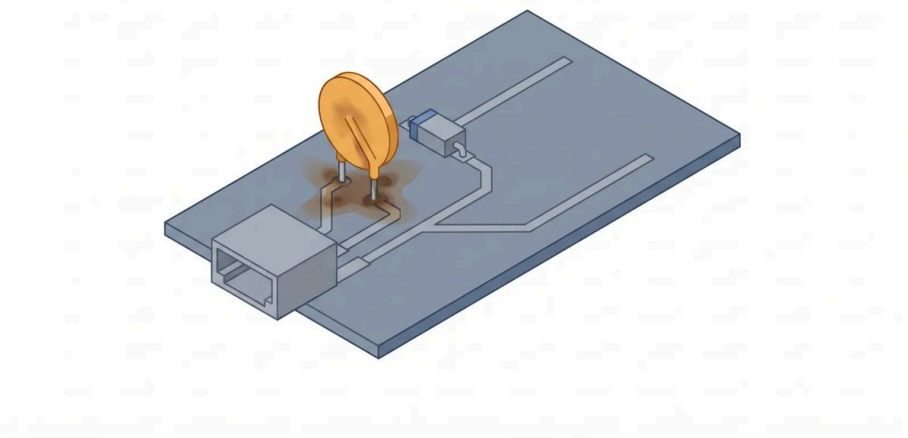
Surge Protection: Varistor Compared with Suppressor Diode

The varistor described above is one of two standard ways of clamping a transient, and they are chosen for different reasons. The comparison matters because they are often found together on the same input, doing different parts of the job:

	Metal oxide varistor	Transient voltage suppressor diode
How it clamps	Its bulk resistance falls sharply above the clamping voltage	It avalanches at a specific reverse voltage — a Zener built for surge energy rather than for regulation

	Metal oxide varistor	Transient voltage suppressor diode
Direction	Inherently symmetrical; the same in both polarities	Single-polarity as standard; a **bidirectional** version is two devices back to back, for AC lines and bipolar signals
Energy it can absorb	**Large** — this is its main advantage	Smaller
Speed of response	Slower	**Very fast**
Tightness of the clamp	Loose — the clamping voltage rises considerably as the surge current increases	Tight and well defined
Life	**It wears out.** Each surge degrades it a little; leakage rises and the clamping voltage drifts down until it eventually conducts at normal working voltage and overheats	Does not degrade in normal service

The wear-out row is the one with maintenance consequences. A varistor that has absorbed many surges gives no external sign until it begins to conduct at working voltage, at which point it heats, may char its board and can fail short circuit — so on equipment with a history of surge exposure, varistors are treated as life-limited rather than fit-and-forget. The usual arrangement in a well-protected input puts the varistor first to absorb the bulk of the energy and the suppressor diode behind it to catch what the varistor's loose clamp lets through.



A worn varistor gives no sign until it conducts at working voltage, overheats and can char its board.

Rectifier Diodes and Signal Diodes Are Different Products

Both are ordinary PN junctions, and a catalogue separates them because they are optimised in opposite directions. A **rectifier diode** has a large junction area to carry current and shed heat, which gives it a high

current rating but also a large junction capacitance and a slow reverse recovery. Where a high peak inverse voltage is wanted as well, it comes from a separate choice: the light doping that gives the junction a wide depletion region. A **small-signal diode** has a tiny junction, so it switches quickly and has very little capacitance, but it will pass only milliamperes.

Substituting one for the other is a routine field error with two distinct outcomes. A small-signal diode fitted where a rectifier belongs fails almost immediately from overcurrent. A rectifier fitted where a signal diode belongs survives indefinitely and merely stops the circuit working properly at frequency — the harder fault to find, because the component measures perfectly good on a meter.

The Specialist Diodes

The remaining types each exist because one property of the junction has been pushed to an extreme. They are rarely encountered on a line maintenance task but appear regularly in examination options, and each is best remembered by the one thing that was changed:

Device	What was changed	Resulting behaviour	Where it is used
Tunnel diode	Doping raised to an extreme on *both* sides, giving an extremely narrow depletion region	Carriers tunnel straight through the barrier at low forward voltage, producing a region where increasing the voltage **decreases** the current — a **negative resistance** region	Very high frequency oscillators and amplifiers; historically in microwave receivers
PIN diode	An undoped **intrinsic** layer inserted between the P and N regions	A very wide depletion region, hence very low capacitance and high reverse voltage; and at radio frequencies its resistance is controlled by the DC forward current through it	**Fast switching** of radio-frequency signals, and voltage-controlled attenuators; also as a photodetector
Step-recovery (snap-off) diode	The doping profile shaped so that stored charge is held right at the junction	Conducts normally in the forward direction, then **stops abruptly** when reversed, releasing its stored charge in an extremely short burst	Generating very fast pulses, and **harmonic generation** at microwave frequencies, because a sharp edge is rich in harmonics
Gunn diode	Not a junction device at all — a slab of bulk gallium arsenide	The transferred-electron effect produces a negative resistance in the bulk material, so it oscillates when biased in a suitable cavity	**Generating microwave oscillations**, for example in radar and Doppler sensors

Device	What was changed	Resulting behaviour	Where it is used
IMPATT diode	Structure arranged so avalanche multiplication and carrier transit time together produce a phase lag	Presents a negative resistance at microwave frequencies, and **must be operated in that negative-resistance range** to work at all	Higher-power microwave sources than a Gunn diode, at the cost of much more noise
Constant-current diode	A field-effect structure in a two-terminal package	Holds a **nearly constant current** over a wide range of applied voltage — the dual of what a Zener does for voltage	Driving an LED at fixed brightness regardless of supply voltage; biasing
Voltage reference diode	A Zener selected in the range where the two breakdown mechanisms cancel, or a bandgap circuit in the same package	An exceptionally stable output voltage against both temperature and current	The reference against which regulators, analogue-to-digital converters and test equipment measure everything else

Three of those — the tunnel, Gunn and IMPATT diodes — share the property of **negative resistance**, and it is worth pausing on what that means, because it sounds impossible. Over the relevant part of its characteristic, an increase in applied voltage produces a *decrease* in current. A device behaving that way delivers energy to a resonant circuit connected across it instead of absorbing energy from it, which is exactly the condition for sustained oscillation. That is why all three are found as microwave sources and not as rectifiers.

Choosing the Device for the Job

Examination questions and real fault-finding both tend to run the other way round from this section — they start from a function and ask which device performs it. Read the table backwards for practice:

The job	The device	The reason no other will do
Convert AC to DC	Rectifier diode	It conducts in one direction only, which is the definition of rectification
Hold a supply rail at a fixed voltage	Zener diode	Reverse breakdown is the only diode mechanism that holds a constant voltage at a value the designer can choose; forward conduction does it too, but only in fixed steps of one junction drop
Switch a large current with no moving contact	SCR, or a TRIAC for AC	The latched four-layer structure has a very low on-state drop and needs no continuous drive
Produce light from a current	LED	Recombination in a direct-band-gap compound emits a photon; silicon does not
Produce a current from light	Photodiode	Light generates carriers in the depletion region, which the junction field sweeps out

The job	The device	The reason no other will do
Tune a circuit electrically	Varactor diode	Its depletion width, and hence its capacitance, is set by the applied reverse voltage
Absorb a large voltage surge	Varistor, backed by a suppressor diode	Only these are built to dissipate surge energy rather than to block or pass current
Stop the spike when a relay coil is switched off	A freewheel diode, or a suppressor for a faster release	The coil current needs a path to circulate in while its field collapses
Rectify at very high frequency, or detect a very small signal	Schottky diode	No stored minority charge to remove, and a threshold of only about 0.2 V
Pass a signal between two circuits with no electrical connection	Optocoupler	The signal crosses the package as light; no other device here provides true isolation
Generate a microwave signal	Gunn or IMPATT diode	Both present a negative resistance, which sustains oscillation in a resonant cavity

How Good Is a Zener Regulator? Line and Load Regulation

A Zener regulator is judged by two separate numbers, answering two separate questions, and both come back to the diode's small but non-zero dynamic resistance:

- **Line regulation** — how much the output moves when the *input* voltage moves. The Zener and the series resistor form a divider for the change, so only a fraction of any input change reaches the output.
- **Load regulation** — how much the output moves when the *load* current changes. Any current the load takes is current the Zener does not, and the change in Zener current moves its terminal voltage by that current multiplied by its dynamic resistance.

Worked Example — Putting Figures on Both

Take the 5.1 V Zener with its 470 Ω series resistor again, and suppose the diode's dynamic resistance is 5 Ω.

Line regulation. The input rises from 12 V to 15 V, a change of 3 V. The change divides between the resistor and the Zener's dynamic resistance in proportion: $3 \times 5 \div (470 + 5) = 0.032$ V, that is **32 mV** at the output.

A 3 V disturbance has been reduced to 32 mV — a rejection of about 95 to 1.

Load regulation. The load current changes by 10 mA, so the Zener current changes by the same 10 mA in the opposite sense. Output change: $0.010 \times 5 = 0.050$ V, that is **50 mV**, or 0.98 per cent of 5.1 V.

Both numbers improve if the dynamic resistance falls, which is why a data sheet's Zener impedance figure matters as much as its voltage tolerance. Note also that the two mechanisms are independent: a circuit can have excellent line regulation and poor load regulation, and a measurement made with a fixed load will never reveal the second.

Driving More Than One LED

An annunciator panel or a lighting assembly needs many LEDs, and how they are wired decides how evenly they light and what happens when one fails:

	LEDs in series	LEDs in parallel
Current through each	**Identical by definition** — a series circuit has one current	Unequal, and set by each device's own forward voltage
Brightness matching	Excellent, with no selection needed	Visibly uneven unless the devices are matched or each has its own resistor
Supply voltage needed	The sum of all the forward voltages plus headroom for the resistor	Only one forward voltage plus headroom
Effect of one failing open	**The whole string goes out**	Only that device goes out
Resistors required	One for the string	One **per LED** — a single shared resistor does not work

The last row is the one that gets built wrong. Putting several LEDs in parallel behind one resistor is exactly the current-hogging arrangement described later in this note for parallel diodes, and it fails in the same self-reinforcing way: the device with the lowest forward voltage takes most of the current, runs hottest, and its forward voltage falls further. The others sit dim while the first one is overdriven. Series wiring avoids the problem entirely because a series circuit cannot have unequal currents, which is why long LED strips are built as strings of a few devices in series, each string with its own resistor, rather than as one large parallel bank.

The better solution again is to drop the resistor altogether and drive the string from a **constant-current** source, which holds the current steady regardless of supply voltage and of how the forward voltages drift with temperature. That is why aircraft LED lighting assemblies contain a driver module rather than a resistor: on a bus that varies between about 22 V and 30 V in normal operation, a resistor would give a brightness that visibly follows the bus voltage.

Getting a Protective Diode the Right Way Round

Whether a protection device conducts under normal conditions depends on which one it is. Some are arranged to do **nothing at all** until the event they guard against, while others carry current the whole time the circuit is running — and what fitting one backwards costs varies in the same way, from a device that conducts when it should be idle to one that blocks the supply and leaves the equipment dead. So the orientation rule is worth stating once for each case, and the reasoning is the same every time: *ask what the device should be doing while the circuit is behaving.*

- **Zener as a voltage regulator or suppressor.** It works in reverse breakdown, so it must be **reverse biased** in normal service: its **cathode goes to the more positive point** and its anode to the more negative one. If a diagram marks two points either side of a Zener suppressor and asks about their relative polarity, the point at the *anode* end must be the more negative of the two. Fit it the other way round and it is simply a forward-biased diode holding the rail at 0.7 V.
- **Freewheel diode across an inductive coil.** Again reverse biased while the coil is energised, so its **cathode goes to the positive end of the coil** and its anode to the negative end. When the supply is interrupted, the collapsing field reverses the voltage across the coil, the diode becomes forward biased, and the current circulates through it. Fitted the wrong way round it short-circuits the supply the moment the coil is energised.
- **Protective diode in inverse parallel with an LED.** Its **anode goes to the LED's cathode** and vice versa, so that the two face in opposite directions. Whichever half-cycle reverse-biases the LED forward-biases the protective diode.
- **Series blocking diode in a supply feed.** Forward biased in normal operation, so its **anode goes toward the source** and its cathode toward the load.

The first two of those are reverse biased in normal use, the third alternates with every half-cycle of the AC supply, and the fourth is forward biased throughout, which is exactly why one rule cannot be memorised for all of them. What can be carried across every case is the question: work out what the circuit does when it is behaving normally, decide whether the diode should be conducting then, and the orientation follows.

Diodes in Series and Parallel

Series Connection

Diodes connected in series increase the total reverse voltage capability: the PIV rating adds up. However, the forward voltage drops also add up. Used when the reverse voltage exceeds the rating of a single diode.

⚠ **Important:** When connecting diodes in series for reverse voltage sharing, equalising resistors should be placed in parallel with each diode. Without these, the reverse voltage may not divide equally due to differences in leakage current.

Parallel Connection

Diodes connected in parallel increase the total current-carrying capability. However, because no two diodes have exactly the same V_F, current will not divide equally. Small series resistors (ballast resistors) are added to ensure equal current sharing.

Why the Reverse Voltage Divides Unequally

A series string carries the same current through every device in it, and in the blocking direction that current is the reverse leakage. Each diode then takes whatever reverse voltage it needs in order to pass that particular leakage current — and since no two diodes have identical leakage, no two take the same voltage.

The direction of the imbalance is the opposite of what intuition suggests, so it is worth working through. Think of each reverse-biased diode as a very large resistance. In a series circuit the largest resistance takes the largest share of the voltage. The diode with the **lowest** leakage current is the one with the **highest** reverse resistance, so **the best diode in the string is the one that ends up with the most voltage across it** — and is therefore the first to be driven past its peak inverse voltage. It then breaks down, its share of the voltage transfers to the remaining diodes, and the string fails in a cascade within microseconds.

Worked Example — Sizing Equalising Resistors

Two 400 V diodes are placed in series to block 600 V. Ideally each would take 300 V, leaving a 100 V margin. Suppose the worst-case leakage is 50 μA. Choose equalising resistors so that the division is set by the resistors and not by the diodes.

Rule of thumb: make the resistor current about ten times the worst-case leakage, so $10 \times 50 = 500\ \mu$A.

Resistor value: $300 \div 0.0005 = 600\,000\ \Omega$, that is 600 k$\Omega$ across each diode.

Power in each resistor: $300^2 \div 600\,000 = 0.15$ W, so a 0.5 W part gives margin.

The reasoning behind the factor of ten is the whole point. Once ten times as much current is flowing through the resistors as through the diodes, the resistors dominate the division and the leakage mismatch becomes a small perturbation on it. Choose the resistors too large and they cannot do that; choose them too small and they waste power and heat the assembly continuously. The figures above are illustrative — the leakage value must come from the data sheet at the actual junction temperature, which is where it will be highest.

Static resistors solve only half of the problem. During the switching instant, the diodes come out of conduction at slightly different times because their reverse recovery times differ, and whichever recovers first is momentarily blocking the entire string voltage on its own. Resistors are far too slow to help with a transient of that speed, so a capacitor is added in parallel with each diode as well: the capacitors share the fast transient between them while the resistors handle the steady state. Any high-voltage series stack in service will be found to carry both.

Deliberate Uses of Diodes in Series

The additive forward drop mentioned above is not only a penalty. It is used on purpose in three ways an engineer will meet:

- **As a small voltage reference.** Three silicon diodes in series hold about 2.1 V, and the value is far more stable against load current than a resistive divider would be, because each junction's voltage barely changes with current.
- **As a level shifter.** Placed in a signal path, a series diode subtracts a fixed 0.7 V, moving a signal into the range a following stage needs.
- **As a temperature sensor.** Because a junction's forward drop falls by about 2 mV for every degree Celsius, a forward-biased diode fed with a constant current is a linear temperature transducer, and a cheap and rugged one.

A fourth use is the **blocking or isolating diode**, which is series-connected only in the sense that it sits in line with a supply. Where two sources feed a common bus — a battery and a transformer-rectifier unit, for example — a diode in each feed lets both supply the bus while preventing either from back-feeding into the other. Whichever source is at the higher voltage takes the load, and the other is simply reverse biased and idle, with

no switching, no contacts and no control logic. The cost is one forward drop in each path, which on a 28 V bus is a loss of about 2.5 per cent.

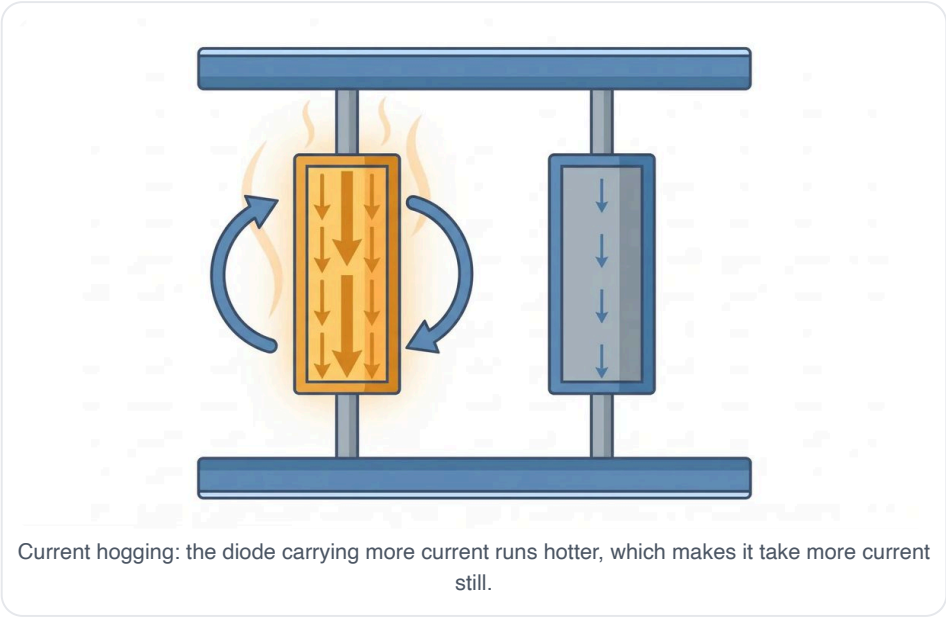
Why Parallel Sharing Fails, and Why It Gets Worse By Itself

Diodes in parallel all have the same voltage across them, and the forward characteristic is exponential, so a very small difference in forward voltage produces a very large difference in current. A mismatch of only 30 mV — well within normal manufacturing spread — is enough to make one diode carry two or three times the current of the other. The warning above is therefore an understatement rather than a caution: without ballast resistors the current does not merely divide unevenly, it divides badly.

What makes it dangerous rather than merely untidy is that the imbalance **reinforces itself**. Follow one cycle of the loop:

- The diode with the slightly lower forward voltage takes more than its share of the current.
- More current means more power dissipated in that diode, so it runs hotter than its neighbour.
- A junction's forward voltage **falls** as it gets hotter, at about 2 mV per degree Celsius.
- So the hot diode's forward voltage drops further below its neighbour's, and at the shared terminal voltage it takes an even larger share of the current.

Each pass round that loop makes the next one worse. This is **current hogging**, and it ends with one device carrying nearly all the current and failing, after which the survivors inherit the whole load and fail in turn. Note that it is the same self-reinforcing thermal mechanism described in the opening section of this note, but arriving through the sharing of current between devices rather than within one junction.



Current hogging: the diode carrying more current runs hotter, which makes it take more current still.

Worked Example — How a Ballast Resistor Breaks the Loop

Two diodes in parallel share a 10 A load, and their forward voltages differ by 30 mV. A 0.05 Ω ballast resistor is fitted in series with each. How unequal is the sharing?

Both branches have the same total voltage across them, so $I_1R + V_{F1} = I_2R + V_{F2}$.

Rearranged: $(I_1 - I_2) \times 0.05 = 0.030$, so $I_1 - I_2 = 0.6$ A.

With $I_1 + I_2 = 10$ A, the currents are **5.3 A and 4.7 A** — an imbalance of 6 per cent.

Power in the busier ballast resistor: $5.3^2 \times 0.05 = 1.4$ W.

The resistor works by negative feedback, and that is the mechanism worth taking away. Any diode that starts to take more current develops more voltage across its own ballast resistor, which leaves less voltage for its junction and pushes its current back down. The self-reinforcing loop of the previous paragraph has been turned into a self-correcting one. The price is the 1.4 W burned in the busier resistor, which is why the ballast value is kept as small as will do the job.

Two further measures are used alongside ballast resistors, and both are cheaper than they look:

- **Matched devices.** Diodes selected from the same production batch, or supplied as a graded matched set, start with a much smaller forward-voltage spread.
- **A common heatsink.** Mounting parallel diodes on one thermal mass forces them to share a temperature. That does not equalise their forward voltages, but it removes the temperature *difference* that drives the hogging loop, so the imbalance stops growing. This is exactly why a bridge rectifier module — four diodes formed on one substrate in one package — shares current far better than four discrete diodes wired to look the same.

The maintenance rule follows from all of the above: devices connected in series or in parallel are replaced as a **set**, not individually. Fitting one new diode alongside three aged ones reintroduces exactly the mismatch the design went to trouble to avoid, and the new device — typically the one with the lowest leakage and the lowest forward drop — is the one that will be overstressed.

Two Ways to Protect Against Reversed Polarity

One of the commonest uses of a single diode is to guarantee that equipment connected backwards is not damaged, and there are two arrangements with quite different consequences:

	Series blocking diode	Shunt (crowbar) diode with a fuse
Where it goes	In line with the positive supply lead	Across the supply input, reverse biased in normal use
Normal operation	Conducts, dropping about 0.7 V	Does nothing at all — no loss

	Series blocking diode	Shunt (crowbar) diode with a fuse
On reversed connection	Reverse biased, so no current flows and the equipment simply does not work	Forward biased, so it short-circuits the supply and blows the fuse
Cost in normal use	**A permanent voltage drop and a permanent power loss**	None
Cost after a reversal	None — correct the wiring and it works	**A blown fuse to replace**, and the diode must survive the fault current
Suits	Low-current equipment where 0.7 V is affordable	Higher-current equipment where a continuous drop would be wasteful

The shunt arrangement looks alarming — it deliberately creates a short circuit — but that is exactly the point: it converts a slow, invisible failure into an immediate and obvious one. A reversed supply that merely fails to work may be mistaken for a dead unit and investigated for hours; a reversed supply that blows a fuse the instant it is connected identifies itself.

A practical caution applies to both. The diode has to be rated for the reversed condition, not for the normal one. In the shunt arrangement it must carry the full fault current for the whole time the fuse takes to clear, and in the series arrangement it must block the full reversed supply voltage without breaking down. A device chosen only against the normal operating figures will fail during the very event it was fitted to survive.

Functional Testing of Diodes

A diode can be tested with a multimeter set to the **diode test** mode:

Test	Connection	Good Diode Reading
Forward bias	Red lead → Anode, Black lead → Cathode	0.5–0.7 V (silicon) or 0.2–0.3 V (germanium/Schottky)
Reverse bias	Red lead → Cathode, Black lead → Anode	"OL" (overload / infinite resistance)

Fault indications:

- **Short circuit:** Low reading (near 0 V) in both directions
- **Open circuit:** "OL" reading in both directions
- **Leaky:** Some reading in reverse direction (should be OL)

✈ **Aviation context:** Diode testing is a routine part of component-level troubleshooting. When testing diodes in-circuit, other parallel paths may give misleading readings — it is often necessary to disconnect at least one lead of the diode for an accurate test. Always observe correct ESD (electrostatic discharge) precautions when handling semiconductor devices.

Why the Diode-Test Range and Not the Ohms Range

The two ranges look interchangeable and are not. A digital multimeter's **diode-test** function forces a small, defined current through the device and displays the **voltage** that results — which is why a good silicon diode reads about 0.6, with the meter showing volts rather than ohms. That figure is the forward drop at the meter's own test current, not a resistance, and reading it as ohms is the first of several traps.

The ohms ranges of a modern digital meter are deliberately designed to apply *less* than a diode drop, so that semiconductor junctions in the circuit under test do not turn on and corrupt a resistance measurement. That is helpful when measuring a resistor and useless when measuring a diode: the meter cannot reach the barrier potential, so **a perfectly serviceable silicon diode reads open circuit in both directions on an ohms range.** Many good components have been discarded on the strength of that reading.

⚠ **On an analogue multimeter the lead polarity may be the opposite of what is printed on it.** Analogue ohms ranges are powered by an internal battery, and on many instruments that battery is connected so that the **black** lead is the positive terminal on the resistance ranges — even though the same lead is negative on every voltage and current range. A forward-resistance check made with such a meter gives its low reading with the *black* lead on the anode. Confirm the instrument's polarity on a known-good diode before drawing any conclusion from an unknown one; the answer is written in the instrument's manual and is not consistent between makes.

What the Forward Reading Tells You Besides Good or Bad

The number itself identifies the technology, which is useful when the marking has been lost or when the part fitted is not the part on the drawing:

Forward reading	What is in front of you	Reverse reading expected
About 0.5 to 0.7 V	An ordinary silicon junction — rectifier, signal diode, or the gate-cathode junction of an SCR	OL
About 0.15 to 0.3 V	Germanium or Schottky	OL, but a Schottky's higher leakage may give a high, unstable reading rather than a clean OL, and that alone does not condemn it

Forward reading	What is in front of you	Reverse reading expected
About 1.5 to 3.5 V, and the device may glow faintly	An LED. The value rises with the emitted colour, so a red one reads lowest and a blue or white one highest	OL
OL in both directions on an LED	**Possibly a good blue or white LED.** Many meters cannot supply the 3 V or more that such a device needs, so it never conducts	Not conclusive — test with a current-limited bench supply instead
Roughly double the expected value	Two junctions in series — a dual diode in one package, or two devices measured through a common connection	OL

A meter reading is also a rough thermometer. Warm a forward-biased junction with a hot-air source or even a fingertip and the reading should fall by roughly 2 mV per degree, which confirms the junction is a junction and not a resistor. A device whose reading does *not* move with temperature is not a working PN junction.

Testing Devices That an Ordinary Diode Check Cannot Reach

- **Zener diode.** A diode check only exercises the forward direction, so it will confirm the junction exists but tells nothing at all about the Zener voltage — the meter cannot supply enough voltage to drive the device into reverse breakdown. The real test needs a variable DC supply, a current-limiting resistor and a voltmeter across the diode: raise the supply and the voltage across the Zener should climb and then stop dead at its rated value while the supply continues to rise. If it keeps climbing with the supply, the diode is not in breakdown.
- **Bridge rectifier module.** Four junctions are reachable from the four terminals. From each AC terminal to the **positive** DC terminal there should be a forward reading with the meter's positive lead on the *AC* pin; from the **negative** DC terminal to each AC terminal there should be a forward reading with the positive lead on the *negative* pin. Reverse each of those four and they read OL, as do both readings between the two AC pins and the reading taken from the **positive** DC terminal to the **negative** one. One combination is the exception: with the positive lead on the **negative** DC terminal and the negative lead on the **positive**, the path runs through two forward junctions in series, so a meter whose diode-test voltage covers two drops reads roughly double a single junction's value while one that cannot reach two drops reads OL. Either result is normal there. Four good forward readings, either result on that one DC-to-DC combination and OL everywhere else is a complete test of the module.
- **SCR.** The gate-to-cathode path reads like an ordinary diode. Anode to cathode reads OL in *both* directions, because the device has not been triggered — that is the correct result for a serviceable SCR, not a fault. Whether it will latch is a separate test, and most multimeters cannot supply enough current to hold one on even after a successful trigger, so a small lamp and a battery make a more trustworthy rig than the meter does.
- **Photodiode.** Its reverse reading is the test. Measured in the dark it should be OL or nearly so; shine a light on it and the reverse leakage rises measurably. A photodiode that reads the same covered and uncovered

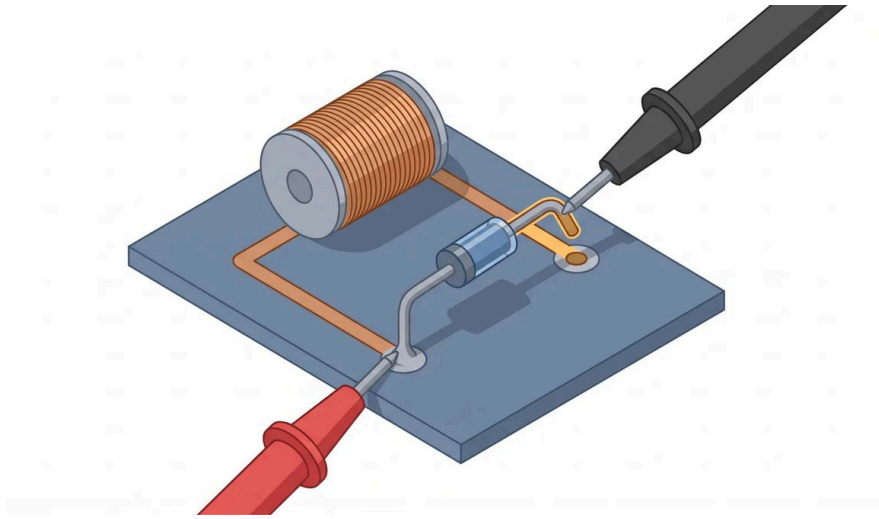
has lost its response even if both readings look reasonable.

In Circuit: What Else Is Connected Across the Diode

The guidance above is that parallel paths mislead, and it is worth knowing exactly which paths and how, because the readings they produce imitate specific faults:

What else is across the diode	Reading it produces	Fault it imitates
A transformer or relay winding	A low reading in both directions, since a winding is a few ohms of copper	A short-circuited diode
A smoothing or decoupling capacitor	A reading that drifts as the capacitor charges from the meter's test current, then settles	A leaky diode — and the drift may be mistaken for an intermittent contact
Another diode in inverse parallel, such as an LED's protection diode	A forward reading in *both* directions	A short circuit, since neither direction gives OL
A bleeder or equalising resistor	A finite reverse reading instead of OL	A leaky diode

In every one of those cases the reading is entirely correct for the circuit as built, and the diode may be perfect. Disconnecting one lead of the device removes the ambiguity in one step, and it is the only way to be certain.



Lifting one diode lead breaks every parallel path, so the meter sees the diode and nothing else.

⚠ **Two things must be true before any in-circuit measurement is taken:** the equipment is isolated from every power source, and the filter capacitors have been *proved* discharged rather than assumed to be. A charged capacitor discharging into a meter set to a low-resistance range will damage the meter, and in a high-voltage supply it will do considerably worse to the person holding it.

From Symptom to Cause

Component testing is usually the last step, not the first. In a power supply the fault can normally be narrowed to one component before anything is unsoldered, and the instrument for that is an oscilloscope on the output. The first two rows below apply the ripple-frequency test set out with the rectifier circuits earlier in this note; the remaining three cover faults that test cannot see:

Symptom	Most likely cause	Confirming check
Output low, ripple amplitude roughly doubled, **ripple frequency has halved**	One diode in the bridge has failed open circuit	Diode-test each of the four junctions; exactly one will read OL in both directions
Output low, ripple amplitude high, **ripple frequency unchanged**	The smoothing capacitor, not the rectifier — lost capacitance, or a high internal resistance	Substitute a known-good capacitor
Fuse blows immediately on power-up, transformer buzzes or gets hot	A diode has failed short circuit, placing a short across the secondary on alternate half-cycles	Test all four; replace the whole set, and check the transformer before reapplying power
Output correct but drifts with load	The regulating element, not the rectifier — a Zener out of breakdown or an overloaded output	Measure the current in the Zener's series resistor
Fault appears only when the equipment is hot, and clears on cooling	A leaky junction. Leakage roughly doubles for every 10°C rise, so a marginal device passes every bench test at room temperature	Repeat the measurement at operating temperature, with the equipment under load

That last row is the one that costs the most time in service. A component that only misbehaves hot will test perfectly on a cold bench, be signed off as serviceable, and produce exactly the same complaint on the next flight. When a reported fault is temperature-dependent, the measurement has to be made at temperature or it has not been made at all.

Finally, a diode is a semiconductor and its handling carries the same electrostatic precautions as any other — earthed wrist strap, conductive mat and conductive packaging — and the full handling regime belongs with printed-circuit board practice rather than here. Test results are recorded with the conditions under which they were taken, since a forward drop quoted without its test current and its temperature is not a measurement anyone can repeat.

Seeing the Characteristic Instead of a Single Point

A multimeter gives one point on the characteristic curve — the forward voltage at the meter's own test current — and that is enough for most purposes. Where it is not, the whole curve can be displayed on an

oscilloscope, which is what a dedicated curve tracer does.

The arrangement is simple in principle: a low-voltage AC supply drives the device through a small current-sensing resistor; the voltage *across the device* goes to the oscilloscope's horizontal input and the voltage *across the sensing resistor*, which is proportional to current, goes to the vertical input. With the instrument in X-Y mode the trace draws the current against the voltage — the characteristic curve itself, sixty or four hundred times a second. What it shows that a meter cannot:

- **The knee**, and therefore whether the device is silicon, germanium or Schottky, read directly off the horizontal axis.
- **The reverse breakdown**, if the supply voltage reaches it — which is how a Zener voltage is confirmed. A rectifier's PIV is far higher than a low-voltage rig can reach, so checking that needs a curve tracer with a high-voltage supply.
- **Excess leakage**, which appears as a reverse region that slopes instead of lying flat.
- **Two devices compared side by side** for matching, before they are put in series or parallel.

Measuring a Diode That Is Working

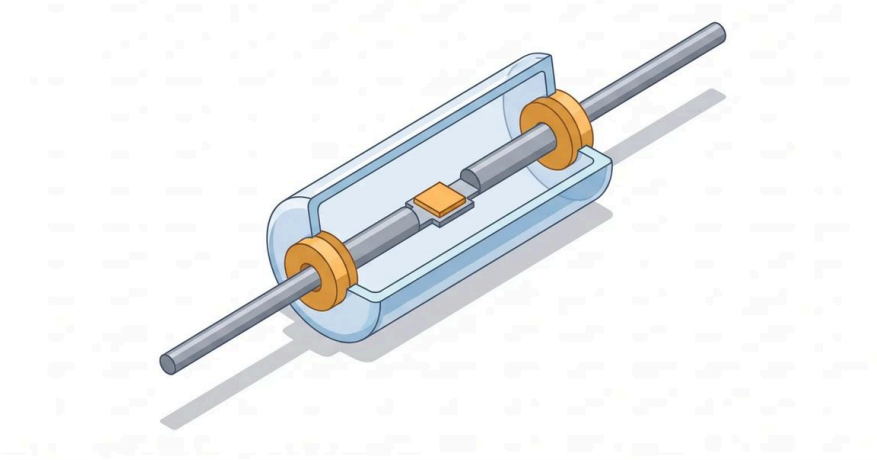
Some faults only exist under load, and for those the diode has to be measured in the powered circuit rather than tested out of it. In a DC circuit — a series blocking diode, a supply isolation diode, a suppression diode across a coil — a voltmeter across the device answers the question in one reading:

Voltage measured across the diode, powered	Conclusion
About 0.6 to 0.7 V (0.2 to 0.3 V for a Schottky)	Conducting normally — serviceable
Approximately zero	**Short circuited**, or being bypassed by something else in parallel
Approximately the full supply voltage, with no current reaching the load	**Open circuited** — the diode is dropping everything because nothing is flowing
Somewhere between, and unsteady	A high-resistance joint or a cracked device, often intermittent with vibration or temperature

The middle two rows are the useful ones, because they distinguish the two failure modes without disconnecting anything. Note that they are opposites in a way that is easy to reverse under pressure: a *short* shows almost no voltage across the device and a healthy current through the circuit, while an *open* shows the whole supply across the device and no current at all. Confirming the load current as well as the diode voltage removes the ambiguity entirely.

Removing and Refitting Without Destroying the Replacement

A semiconductor junction can be damaged by the heat of the operation that installs it, and the damage does not always show up immediately — a junction cooked during soldering may work on test and fail weeks later. Three precautions apply to any small semiconductor:



In a glass-bodied diode the junction sits close to the leads, and a seal at each end keeps moisture out.

- **Work quickly with a hot iron rather than slowly with a cool one.** A properly hot bit makes the joint in a couple of seconds and puts less total heat into the device than a marginal one held there for ten.
- **Use a heat shunt on short leads.** A pair of long-nosed pliers or a purpose-made clip gripping the lead between the joint and the body conducts heat away before it reaches the junction. This matters most on axial glass-bodied devices, whose leads are short and whose junction is close to the joint.
- **Do not bend a lead closer to the body than the manufacturer allows**, and never bend it at the seal. The glass-to-metal or plastic-to-metal seal is what keeps moisture out, and a cracked seal admits contamination that raises leakage over months.

Finally, a measurement is only useful if it can be repeated. A forward-voltage figure recorded without its test conditions cannot be compared with anything later, so the record carries the instrument used, the range or function selected, and the temperature of the component when the reading was taken. Test equipment used to make a maintenance decision must itself be within calibration; an out-of-calibration meter produces readings that look entirely normal and are not evidence of anything.

Module 5 — Digital Techniques / Electronic Instrument Systems

5.1 — Electronic Instrument Systems

Electronic instrument systems have replaced traditional analogue "steam gauge" instruments in modern aircraft. These systems use computers, digital data buses, and electronic displays to present flight, engine, and systems information to the crew. Understanding the architecture, components, and layout of these systems is essential for any aircraft maintenance engineer working on current-generation aircraft.

Evolution from Analogue to Digital

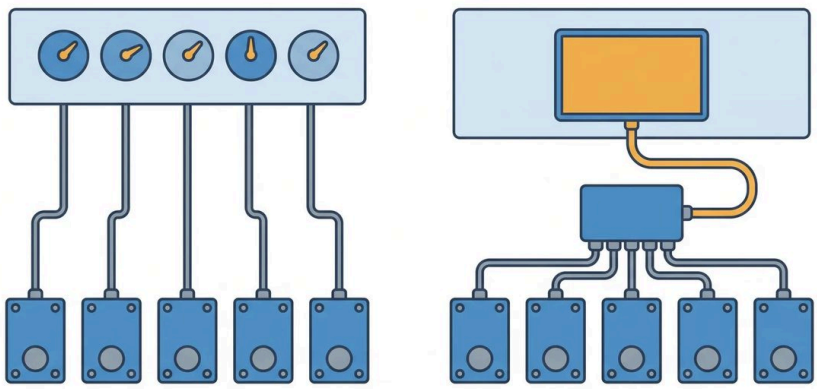
Traditional cockpits used individual electromechanical instruments — each driven by a dedicated sensor and displaying a single parameter on a dial or pointer. A typical analogue cockpit might contain over 100 individual instruments.

Modern **glass cockpits** replace these with a small number of large electronic displays driven by digital computers. This provides:

- **Reduced weight and space** — fewer instruments, less wiring.
- **Improved reliability** — fewer moving parts.
- **Flexibility** — displays can be reconfigured, and information presentation can change with flight phase.
- **Reduced pilot workload** — integrated displays combine related information.
- **Built-in test** — digital systems can continuously monitor themselves and report faults.

Why the One-Instrument-One-Parameter Cockpit Ran Out of Room

The limit on an electromechanical flight deck is geometry, not electronics. Every dial needs a panel cut-out inside the seated pilot's reach and visual scan, its own transducer, its own wiring run from that transducer to the panel, its own integral lighting and its own supply. Panel area is fixed by where the pilot's eyes and hands are, so adding a system means shrinking or deleting something already there. Weight, wiring length and the number of items that can fail all rise roughly in proportion to the number of parameters displayed. That proportionality is why the last generation of large analogue aeroplanes needed a third crew member sitting at a dedicated systems panel: the indications would not fit in front of two pilots. The Boeing 747-400 deleted the flight engineer's station that the earlier 747 required, and it did so by drawing the engineer's parameters on shared electronic displays instead of giving each one a gauge.



Each dial needs its own sensor and wire; a shared screen takes the same sensors through one computer.

An electronic display breaks that proportionality because it is a general-purpose drawing surface. What it can present is limited by the data the computers behind it receive, not by how many holes are cut in the panel, so adding a parameter becomes a software and data-bus task rather than a structural one and costs no extra panel area. The same property is what allows a single screen to change its content with flight phase, which no dial can do.

✈ **Aviation context:** electronic instrument systems replaced the *indicators*, not the *sensing*. Pitot and static pressure, total air temperature probes, thermocouples, tachometer probes, synchros and gyros are all still fitted. What changed is that their outputs are digitised, sent over a data bus and drawn as symbology, rather than being wired directly to a movement in a dial. This is worth holding on to when troubleshooting: a wrong indication on a screen has exactly the same sensing chain behind it as a wrong indication on a gauge, plus several computers.

Three Generations of Flight Deck

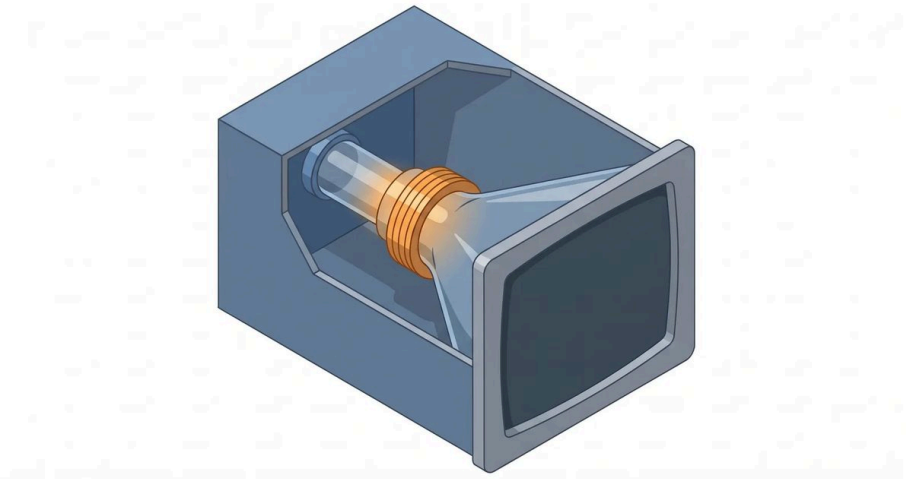
A line engineer meets all three of these, sometimes on the same ramp, and the fault-finding method differs sharply between them. The distinction that matters is not how old the aeroplane looks but where the information is turned into a picture.

Generation	How the indication is produced	How data reaches the panel	Fault-finding character
Direct-reading and electromechanical	Capsules, bourdon tubes, gyros, moving-coil and synchro-driven movements inside each instrument	Pneumatic lines, thermocouple leads and individual analogue wiring runs, one chain per indication	The faulty indication identifies its own chain: one sensor, one wire, one instrument
Early electronic and hybrid	First-generation EFIS: colour CRT attitude and navigation displays, with conventional airspeed, altimeter and engine instruments alongside them	Digital data buses appear alongside analogue signalling; sensor data is concentrated into computers before it reaches a display	Split personality — some indications are still one-to-one, others already pass through a computer
Integrated glass	Large flat-panel displays drawing every flight, navigation and system indication, reconfigurable in flight	High-rate switched data networks feeding shared computing cabinets that host many functions as software	The indication no longer identifies the unit at fault; the chain must be traced through buses, computers and software configuration

First-generation EFIS is the generation the classic exam questions describe, and it is worth being precise about what it did and did not replace. It substituted two cathode ray tubes for the attitude director indicator and the horizontal situation indicator in front of each pilot, but it left the airspeed indicator and altimeter as separate mechanical instruments. The fully integrated primary flight display — attitude in the centre with airspeed and altitude presented as moving tapes flanking it — belongs to the following generation. That history explains why exam questions about the electronic attitude director indicator list decision height, radio altitude, autoland status and flight director bars but not airspeed: on the aircraft the question was written for, airspeed was not on that tube.

From Cathode Ray Tube to Active-Matrix LCD

The display unit itself went through its own generational change, and both technologies still fly. A colour CRT display unit works exactly as a colour television tube does: three electron guns, one per primary colour, fire beams that are steered by magnetic deflection coils and land on a screen coated with triads of red, green and blue phosphor dots. A perforated metal shadow mask sits just behind the phosphor layer so that each gun's beam can only strike its own colour of dot. The anode operates at tens of kilovolts to accelerate the beams, and that high tension is the single most important hazard the unit presents to a maintenance engineer.



A colour CRT display unit: the long glass tube, with deflection coils at its neck, sets the unit's depth.

Two quite different writing methods are used, often on the same tube in the same picture:

- **Stroke writing** (also called cursive or calligraphic) steers the beam directly along the outline of the symbol to be drawn, in the way a pen is moved on paper. It produces bright, smooth, high-contrast lines with no stair-stepping, which is what attitude scales, pointers, alphanumerics and pitch ladders need.
- **Raster writing** scans the beam over the whole screen in a fixed pattern of horizontal lines, exactly as a television does, and modulates its brightness as it goes. It is the only practical way to fill an area, so it is what produces weather radar returns, a shaded sky-ground horizon, terrain shading and any video image.

Neither method alone is sufficient, so display units interleave them: a raster field paints the filled areas, then a stroke field overwrites the symbology on top of it, and the eye integrates the two into one picture. This is the reason the classic description of an electronic flight instrument display lists a raster *and* stroke generator as one of its four building blocks, alongside the mode or control panel that selects the format, the processor that decides what to draw, and the sensor data inputs that tell it the values.

An active-matrix liquid crystal display works on an entirely different principle. It emits no light of its own: a backlight illuminates the panel from behind, and each pixel acts as a controllable shutter. The liquid crystal layer sits between two polarisers, and applying a voltage across a pixel changes the twist of the crystal and therefore how much light that pixel passes. "Active matrix" means each pixel has its own thin-film transistor holding its state, which is what allows a large, sharp, flicker-free colour picture without the crosstalk that a passively addressed panel suffers.

Property	Colour CRT display unit	Active-matrix LCD display unit
Depth, weight and panel area	Deep and heavy — the tube's length sets the depth, which limits how large the visible area can be made	Flat and light; the visible area can be made large enough to be divided into several independent windows
Supply	Needs an extra-high-tension supply of tens of kilovolts; significant power draw and heat	Low-voltage electronics; most of the power goes into the backlight
Susceptibility	Beam position is set magnetically, so an external magnetic field can distort the geometry or upset the colour purity of the picture	Immune to magnetic deflection effects; sensitive instead to temperature at the cold end of the range
Ageing and wear	Phosphor ages and a static symbol can leave a ghost of itself burned into it; cathode emission falls with hours, so brightness decays	Backlight output falls with hours and individual pixels can fail; the crystal itself does not wear in the same way
Typical degraded symptom	Dim, discoloured or defocused picture, convergence errors showing as coloured fringes on symbols	Overall dimming with correct symbology, slow response when cold, isolated dead or stuck pixels
Maintenance consequence	Treat as a high-voltage unit even when disconnected; never open the case in the field	Treat as a sealed optical assembly — the face carries an optical coating and the electronics behind it are static-sensitive

Cold-temperature behaviour is the one respect in which the flat panel is the more delicate of the two. Liquid crystal becomes sluggish as it cools, so a panel left cold-soaked overnight can be slow, smeared or blank at first power-up. Display units therefore contain heaters, and the correct action after a cold night is to allow the specified warm-up before condemning a unit that is simply not yet at temperature. Reporting such a display as failed, when the maintenance manual allows several minutes of warm-up, is a common way to generate an unnecessary component removal.

⚠ **High tension:** a cathode ray tube display unit holds a charge on its final anode after power has been removed, and the tube is also an evacuated glass envelope that can implode if it is struck or scratched. Do not open, probe or attempt bench repair of a display unit on the aircraft, and handle a removed CRT unit by its case with the screen protected. Flat-panel units carry no such high tension but are static-sensitive assemblies, so the electrostatic-discharge precautions covered elsewhere in this module apply to them and to the cards inside the computers that drive them.

What Gets Harder

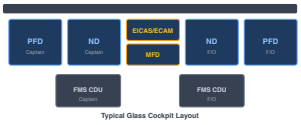
The advantages listed at the start of this section are real, and they are bought at a price that an engineer needs to understand as clearly as the benefits:

- **Common-mode failure replaces independent failure.** A hundred separate instruments fail one at a time and for a hundred separate reasons. A set of identical display units running identical software can, in principle, all be affected by the same design or software error at the same moment. Certification answers this with dissimilarity and with an independent standby chain rather than by adding more of the same unit.
- **Configuration becomes part of the aircraft's identity.** A display unit is defined by the software part number loaded into it and by its position programming as well as by the part number on its label. Two physically identical units are not necessarily interchangeable until the right configuration is in place.
- **Everything depends on electrical power and on cooling.** Dial instruments driven pneumatically or by a self-contained gyro keep working when the busbar does not. Electronic displays do not, which is why the powering arrangements and the standby chain covered in the last section of this note are as much a part of the instrument system as the screens.
- **Diagnosis moves upstream.** A parameter now travels sensor → data bus → computer → symbol generator → display unit. A wrong or missing indication identifies a chain, not a component, and the skill the job now demands is deciding which link owns the symptom.
- **Obsolescence bites the display, not the aeroplane.** Display hardware follows the commercial electronics lifecycle, which is far shorter than an airframe's, so a display system will typically be re-equipped at least once during the aircraft's service life.

Typical Glass Cockpit Layout

A modern transport aircraft cockpit typically features the following main displays:

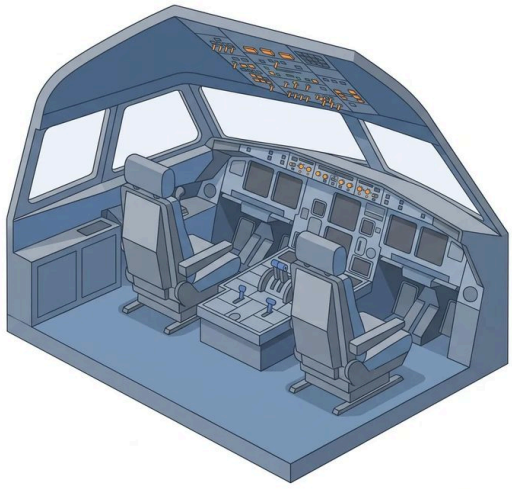
Display	Abbreviation	Function	Typical Location
Primary Flight Display	PFD	Attitude, airspeed, altitude, vertical speed, heading, flight director, ILS deviation	Directly in front of each pilot
Navigation Display	ND	Map, route, waypoints, weather radar, TCAS traffic, VOR/DME information	Beside each PFD (inboard)
Engine/Systems Display	EICAS or ECAM	Engine parameters (N1, N2, EGT, fuel flow), crew alerting, systems status	Centre panel, upper
Multi-Function Display	MFD	Secondary systems pages, checklists, synoptic diagrams	Centre panel, lower



Typical Glass Cockpit Layout

The Four Zones of a Transport Flight Deck

Display units are only part of the picture. A transport flight deck is organised into four zones, and knowing which zone owns which function saves a great deal of time when locating equipment or interpreting a defect report:



A flight deck has four zones: glareshield, main instrument panel, centre pedestal and overhead panel.

- **Glareshield** — the autopilot and autothrust target selector (Boeing calls it the mode control panel, Airbus the flight control unit), the electronic flight instrument control panels, and the master warning and master caution lights and their reset switches. Everything here is placed for a hand reaching forward without leaving the seat.
- **Main instrument panel** — the primary flight and navigation displays for each pilot, the centre displays carrying engine and systems information, the standby instruments, the landing gear lever and the primary annunciators. This is the zone the eyes live in.
- **Centre pedestal** — the flight management control and display units, radio and transponder panels, engine start controls, thrust levers, speed brake, flap lever and trim.
- **Overhead panel** — the switches and indicators that *configure* the aircraft systems: electrical, hydraulic, pneumatic and air conditioning, fuel, ice and rain protection, lighting, oxygen, APU and engine start selection.

The overhead panel is worth a sentence of its own because it is often assumed that glass cockpits did away with it. They did not, and the reason is a clean division of labour: the displays *show the state* of a system,

while the overhead panel *commands its configuration*. Removing it would mean routing every valve and pump selection through a menu on a screen, which is slower and less certain than a switch that can be found by touch. What did change is how the panel is read. Modern overhead panels follow a dark-cockpit philosophy: when the aircraft is correctly configured for the phase of flight, no light is lit. The crew scan for illuminated captions rather than reading each switch position, and an engineer restoring a system after maintenance is not finished until the panel is dark again.

The Basic T Survived the Change

Instrument panels have been standardised around the basic T arrangement since long before electronic displays: attitude at the top centre, airspeed to its left, altimeter to its right, and the heading indication directly below the attitude indicator. The arrangement exists because the pilot's scan radiates from attitude, and the four parameters needed most often should be one short eye movement away.

The primary flight display did not discard that geometry, it absorbed it. The attitude picture occupies the centre of the screen, the airspeed tape runs vertically down its left-hand edge, the altitude tape down its right-hand edge, and a heading or track scale sits across the bottom. A pilot trained on dials therefore uses the same scan pattern on glass, and an engineer looking at an unfamiliar type can find the four primary parameters immediately. Understanding this also explains why the primary flight display is *never* the screen chosen to carry a systems synoptic: its content is fixed by the scan it has to support.

Inside the Primary Flight Display

Beyond the four basic-T parameters, the primary flight display gathers everything needed to fly a precision approach without looking away:

- **Attitude** — a pitch ladder and horizon against a fixed aircraft symbol, a bank angle scale with a pointer, and a slip and skid indication beneath it.
- **Airspeed tape** — the current speed in a reading window with the scale moving past it, a speed trend vector showing the speed that will be reached in the next few seconds at the present acceleration, take-off and manoeuvring speed bugs, the selected speed, and low- and high-speed limit bands.
- **Altitude tape** — altitude with the selected altitude bug, the barometric reference setting in use, and the vertical speed indication adjacent to it.
- **Flight mode annunciation** — a band across the top of the display stating which autothrust, lateral and vertical modes are armed and which are engaged, plus the autopilot and flight director status. This has no equivalent on a dial cockpit, and it is the single most important addition the primary flight display made: with automation flying the aeroplane, knowing which mode is actually engaged matters more than any individual parameter.

- **Approach information** — localiser and glideslope deviation scales, radio altitude, the selected decision height or minimum, and autoland or approach capability status.
- **Guidance** — the flight director command bars, and the flight path vector or bird on types that provide one.
- **Failure indications** — flags and removed symbology when a source becomes invalid, which the last section of this note deals with in detail.

Inside the Navigation Display

Where the primary flight display answers "how is the aircraft flying", the navigation display answers "where is it, relative to what". Its content is selectable, and the modes offered are broadly the same across manufacturers even though the names on the selector differ:

Mode	Orientation and centring	Typical picture	Weather radar overlay
Expanded arc (full arc)	Heading or track up, aircraft symbol at the bottom, compass shown as a forward arc	The forward sector at the selected range, with route, navaids, traffic and terrain	Available
Rose (VOR or ILS)	Heading up, aircraft symbol in the centre of a full 360° compass rose	Raw radio navigation data — VOR bearing and deviation, or localiser and glideslope deviation	Available
Map	Heading or track up, aircraft symbol at the bottom or centred	The flight-management route with waypoints, airports, navaids, altitude and speed constraints, wind, ground speed and distance to go	Available
Plan	**North up**, centred on a selected flight-plan waypoint, not on the aircraft	A static, chart-like view of the route used to check the programmed flight plan waypoint by waypoint	**Not available**

The one mode that cannot carry weather radar is plan, and the reason is geometric rather than arbitrary. The radar antenna scans a sector ahead of the aircraft, so every return it produces is referenced to the aircraft's own position and heading. All the other modes are drawn around the aircraft symbol, so the returns can be laid straight onto them. Plan mode is oriented to north and centred on a waypoint that may be hundreds of miles away, so there is no correct place to put a return that was measured relative to the aeroplane. The radar picture is therefore suppressed rather than shown in the wrong place.

Range selection deserves the same care. The range control sets the scale of the picture and therefore also the depth of the radar sweep that is displayed, so a cell that is plainly visible at a long range disappears from the screen when a short range is selected — not because the weather has gone, but because the display is

no longer drawing that far ahead. A defect report of "weather radar not displaying" is very often a range or mode selection, and the check costs nothing.

The Multi-Function Display and Its Controls

A multi-function display is defined by what it can be told to become. Its content is selected by the crew from a menu of pages, and the mechanisms for making that selection are worth knowing by name because they appear in defect reports:

- **Bezel or line-select keys** — physical switches arranged around the edge of the screen. Each one does a different job on each page, because the label that defines its current function is drawn on the display immediately beside it. A key that appears dead may simply be on a page where it has no function assigned.
- **Display partitioning** — dividing one physical screen into independent windows so that, for example, a navigation map, an engine page and a checklist can be shown side by side. Partitioning is what makes a large flat panel worth fitting: the screen is not one indication made bigger, it is several indications made simultaneous.
- **Cursor control device** — a trackball or touchpad, usually on the pedestal or the armrest, that moves a pointer across the displays so the crew can open menus and select items exactly as a mouse does. It appeared for a structural reason: once displays absorbed enough functions, the number of dedicated switches needed to reach them would have started growing again, and a pointer is the general-purpose control that matches a general-purpose display. Its human-factors weakness is precision pointing in turbulence, which is why cursor devices are mounted with a hand or wrist rest and why on-screen targets are made deliberately large.
- **Touchscreen** — direct interaction by touching the displayed control. Found on the most recent flight decks and widely on business aircraft, it removes a layer of hardware but shares the same turbulence problem, and it obliges the crew to look at what they are touching.
- **Multi-purpose control and display unit** — the keyboard and screen on the pedestal used to load the flight plan, performance data and radio frequencies into the flight management system, and on many types also to interrogate the maintenance computer.

Head-Down, Head-Up and Enhanced Vision

Everything described so far is a **head-down display**: a screen mounted in the instrument panel, read by looking down and inside. A **head-up display** puts a second, smaller set of symbology on a transparent combiner glass mounted between the pilot's eyes and the windscreen. Two properties make it more than a convenience. The image is collimated, that is, focused at optical infinity, so the eye does not have to refocus between the symbology and the outside world; and the symbology can be drawn conformally, so that a flight path marker sits on the point on the ground the aircraft is actually going to, and a runway outline overlays the

real runway. That combination is what allows a head-up display to be credited for low-visibility take-off and approach operations.

Two related systems are frequently confused, and the difference is the source of the picture:

- **Synthetic vision** builds a computer-generated three-dimensional view of terrain, obstacles and runways from a stored *database*, positioned and oriented using the aircraft's own navigation and attitude data, and draws it as the background of the primary flight display. It shows what the database says is there, in any weather and in the dark, and its integrity depends entirely on position accuracy and database currency.
- **Enhanced vision** presents a real-time image from a *sensor* — typically infrared or millimetre-wave — so it shows what is actually there, including traffic and vehicles that no database contains, but only to the extent the sensor can penetrate the conditions.

Both feed the same displays, and combined installations exist, but for maintenance purposes they are different animals: a synthetic vision complaint points at navigation sources and database loading, an enhanced vision complaint points at a sensor, its window and its alignment.

Where the Standby Instruments Sit

The standby instruments are part of the panel layout, not an afterthought bolted to it. They are placed in the centre of the main instrument panel, low enough to leave the primary displays unobstructed but inside the scan of both pilots, because either pilot may have to fly on them. Their content and their independence are dealt with in the final section of this note; what belongs here is that the layout reserves that space deliberately, and that anything mounted so as to obstruct the standby indications — a tablet mount, a placard, a temporary installation — is a certification and safety issue rather than a cosmetic one.

✈ **Aviation context:** examiners test the split of information between the attitude display and the navigation display more often than any other point in this paragraph, and the discriminator is reliable. Anything referenced to the horizon or to the approach path — pitch and roll, flight director bars, slip, radio altitude, decision height, autoland status, glideslope and localiser deviation — is on the attitude display. Anything referenced to the ground plan — compass rose, selected heading, waypoints, distance to go, weather radar returns, traffic and terrain — is on the navigation display. If an answer option pairs a compass rose or weather radar with the attitude display, it is describing the navigation display.

EFIS — Electronic Flight Instrument System

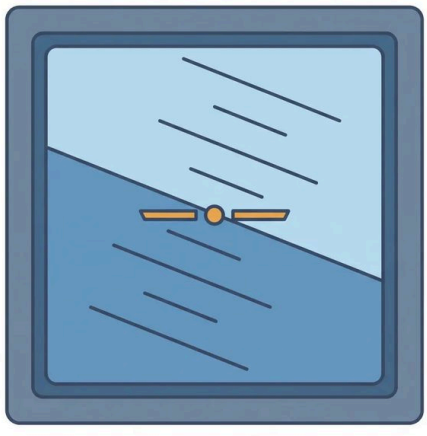
EFIS is the system that generates and displays the primary flight information. A typical EFIS installation includes:

- **Display units** — LCD screens showing PFD and ND information.
- **Symbol generators (SGs)** — computers that receive sensor data, process it, and generate the display imagery. A typical installation has fewer SGs than display units: each side's displays are normally driven by their own SG, with a further SG held as a standby that can be switched in to replace either one.
- **EFIS control panel** — allows the pilot to select ND range, mode (MAP, VOR, ILS, PLAN), and display options (weather, traffic, terrain).
- **Sensor inputs** — air data computers (ADCs), inertial reference systems (IRS), radio navigation receivers, FMS.

EADI, EHSI and Today's PFD and ND

Two sets of names are in use for the same two screens, and both appear in examinations. The older set names the instruments the tubes replaced: the **electronic attitude director indicator** (EADI) and the **electronic horizontal situation indicator** (EHSI). The current set names what the screens do: the primary flight display and the navigation display. A basic electronic flight instrument system is therefore correctly described as the attitude display, the navigation display and the symbol generators that drive them — the control panels and the sensors are associated equipment feeding that core, not part of it. A *complete* installation is a matched pair of these: an independent set for the captain and another for the first officer, each with its own displays and its own symbol generator, so that a failure on one side leaves the other side flying.

The attitude display carries an aircraft symbol that is **fixed to the case** as the datum; the horizon line and pitch ladder move behind it. This is inherited directly from the mechanical attitude director indicator and it is the reverse of what an untrained eye expects, because the aeroplane appears stationary while the world moves. Everything else on the display — flight director bars, deviation pointers, the rising runway — is positioned relative to that fixed symbol.



The aircraft symbol stays fixed and level to the case while the horizon tilts behind it in a turn.

Information	Which display carries it	Why it belongs there
Pitch and roll attitude	Attitude display (EADI / PFD)	The reference every other attitude-related cue is drawn against
Flight director command bars	Attitude display	They command a pitch and roll, so they must be read against the horizon
Slip and skid indication	Attitude display	Balance is flown with the same control inputs as bank
Radio altitude and decision height	Attitude display	Read during the last stage of an approach, when the eyes cannot leave the attitude picture
Autoland and approach capability status	Attitude display	Tells the crew what the automatics are currently able to do on this approach
Rising runway symbol	Attitude display	A height-above-ground cue tied to the approach
Glideslope and localiser deviation	Attitude display (also selectable as raw data on the navigation display in rose mode)	Flown by attitude changes; the navigation display can additionally present the raw signal
Airspeed, altitude, vertical speed and Mach	Attitude display on an integrated primary flight display	This is why air data computer outputs are routed to the attitude display; on the earliest systems these parameters were still on separate mechanical instruments
Compass rose, selected heading, bearing pointers	Navigation display (EHSI / ND)	A plan view of the horizontal situation. The primary flight display carries only a short heading or track strip, never a rose
Waypoints, route, distance to go, ground speed, wind	Navigation display	Flight-management information is inherently a plan-view picture
Weather radar returns, terrain, traffic	Navigation display	All three are measured or predicted around the aircraft's ground track
Traffic-alert resolution advisory pitch guidance	Attitude display	A resolution advisory commands a vertical rate, so the guidance is flown on the attitude display while the traffic that caused it is plotted on the navigation display

From Sensor to Screen

Every symbol on the screen has travelled the same road: **sensor → data bus → computer → symbol generator → display unit**. Knowing the road is what makes structured fault-finding possible, because each stage can be interrogated separately. The principal sources feeding the symbol generators are:

- **Air data computer** — takes pitot pressure, static pressure and total air temperature and computes calibrated airspeed, Mach number, barometric altitude, vertical speed, true airspeed and static air temperature. It is a computing unit, not a repeater: its outputs are corrected for position error and

compressibility, which is exactly why the same static source can feed both the computer and a standby altimeter and give slightly different readings.

- **Inertial reference system** — provides attitude, heading, position, ground speed and track. It must be aligned while the aircraft is stationary, during which it finds true north by gyrocompassing and takes its starting position from the crew or the flight management system.
- **Attitude and heading reference system** — the lighter alternative, using solid-state gyros and accelerometers with a magnetometer or flux valve for heading. It supplies pitch, roll and heading but *not* position or velocity, which is the practical difference from an inertial reference system.
- **Radio navigation receivers** — VOR, ILS or GLS, DME, ADF and the radio altimeter.
- **Flight management system** — the route, waypoints, constraints, computed position and predictions drawn on the map.
- **Weather radar, terrain awareness and traffic collision avoidance computers** — the overlays.
- **Engine and systems computers** — where the same display system also drives the engine and systems screens.

Reduced to its essentials, an electronic flight instrument display is built from four things: the **mode or control panel** that selects what is to be shown, the **data inputs** that supply the values, a **processor** that decides what to draw, and the **raster and stroke generation** that actually puts it on the screen. Everything else — air data computers, weather radar processors, navigation receivers — is a *source* feeding that display, not a part of it. Examination questions frequently mix the two categories, and the test is simple: if removing the box would leave the screen drawing a picture with one item missing, it is a source; if it would leave no picture at all, it is part of the display system.

Data Validity: How a Display Knows Not to Draw

A dial instrument that loses its input tends to fall to a stop or wander. A digital display has a far better option: it can decline to show anything. Every parameter arriving at a symbol generator is accompanied by information about its own trustworthiness, and the system is designed to be *fail-passive* — the wrong picture is more dangerous than no picture, so a doubtful parameter is removed and replaced by a flag, a dashed field or a cross rather than displayed.

On a typical transport aircraft the data arrives on ARINC 429 buses, and two fields in each 32-bit word do this work. A **sign/status matrix** lets the transmitting equipment label its own output as normal operation, functional test, failure warning, or *no computed data* — the last meaning "I am working, but I cannot compute this right now", as an inertial system reports during alignment. A **parity bit** in the last bit position guards the word against corruption in transit. ARINC 429 uses odd parity: the transmitter sets that bit so that the total number of logic ones across the whole word is odd, and the receiver counts them and rejects the word if the count comes out even.

Worked example — setting and checking an odd parity bit

Suppose the first 31 bits of a word (label, source identifier, data and status fields) happen to contain **twelve** logic ones. Twelve is even, so to make the total odd the transmitter sets the parity bit to **1**, giving thirteen ones in the complete 32-bit word. The receiver counts thirteen, finds it odd, and accepts the word.

Now corrupt any single bit in transit. If a one is read as a zero the count falls to twelve; if a zero is read as a one it rises to fourteen. Both are even, so the receiver rejects the word — a single-bit error is always caught. If the first 31 bits had instead contained **thirteen** ones, the transmitter would have set the parity bit to **0**, because the total is already odd.

Note what this does and does not buy. Parity *detects* an error, it cannot correct it, and it cannot see two errors in the same word because the second flip restores the odd count. That is acceptable here only because the data is repetitive: a rejected word is simply replaced by the next transmission of the same label a few tens of milliseconds later. If words keep failing, the receiving equipment declares the input invalid and the affected symbology is removed from the screen.

Symbol Generators: One Pool, Many Screens

The symbol generator is the heart of the system. It receives the digital data, checks it, decides what the currently selected format requires, and converts the numbers into the graphical symbols, scales, pointers, tapes and text that appear on the screen. Manufacturers give the function different names — symbol generator, display management computer, display processing computer, display controller — but the role is the same: it is the unit that turns data into a picture. The display unit itself contributes remarkably little intelligence; it accepts the generated video and drives the screen.

There are normally fewer symbol generators than display units, and this is a deliberate architecture rather than an economy. The Airbus A320 family, for example, drives its six display units from three display management computers. In normal operation each side of the flight deck is fed by its own generator — number one for the captain, number two for the first officer — and the third is a standby that is not routinely displaying anything. It is not a comparison channel and it is not a permanently active primary; it exists to be switched in to either side when a primary generator fails, which is precisely why the architecture is worth more than simply fitting one generator per screen: a spare that can serve either side covers twice as many failures as a spare dedicated to one side.

Because any generator can be routed to any display, the failure cases degrade gracefully rather than in steps. Only when every other generator has failed do both the left and right displays end up being fed by the same single generator, and that condition is the last resort rather than a normal configuration — it restores the picture on both sides but removes the independence between them.

Worked example — losing a symbol generator in flight

A three-generator system is in the normal configuration: number one driving the captain's displays, number two driving the first officer's, number three in standby.

Generator one fails. The captain's displays lose their picture while the first officer's are unaffected — that asymmetry alone identifies the failure as being on the captain's side of the architecture rather than in a shared sensor. Selecting the standby source on the captain's switching panel routes generator three to those displays and the picture returns. Both crew now have independent displays again, but there is no longer a spare.

Generator three then also fails. The only remaining source is number two. Selecting it to feed both sides restores symbology on all four screens, but the two sides are now drawing from one computer. The cross-comparison between them is meaningless, because a fault inside that generator would produce the same wrong picture on both sides simultaneously. The crew's remaining independent reference is the standby instrument set, and this is exactly the case those instruments exist for.

The maintenance lesson from the same sequence: a report of "captain's displays blank, restored by switching" is a symbol generator or its supply, not a display unit, because switching the source changed nothing about the screens themselves.

Comparators and Cross-Monitoring

Independence is only useful if somebody checks that the two sides agree. That checking is done by **comparators inside the symbol generators**, not in the display units and not in a separate box: each generator receives both its own side's data and the other side's, compares the two channels parameter by parameter, and raises an annunciation when a difference exceeds a set threshold and persists for a set time. Typical comparator annunciations cover airspeed, altitude, attitude, heading and, on approach, localiser and glideslope deviation.

A comparator warning says only that two sources disagree; on its own it cannot say which one is wrong. This is why the number of independent sources matters so much in this kind of architecture. With two sources a disagreement can be detected but not resolved. With three, the system can identify the outlier by majority and disregard it, which is the reason large transports carry three air data and inertial reference sources rather than two. Where only two are available, the resolution has to come from outside the pair — the standby instruments, or a parameter derived a different way.

Two Control Panels, Two Different Jobs

An electronic flight instrument system presents the crew with two distinct panels, and confusing them is a common examination error. One is the **display control panel**, which selects the *format*: navigation display mode and range, which overlays are shown, the barometric reference, and the decision height or minimum. The other is the **source or switching panel**, which selects *where the information comes from*: which symbol generator drives which display, and on many types which air data and inertial reference source feeds each

side. Neither is a brightness control and neither is a spare copy of the other: the first decides what the picture looks like, the second decides who draws it.

The source panel is the one that matters most to maintenance, because it is the only control on the flight deck that can silently change the architecture the aircraft is flying on. Its switches are normally left in a normal or automatic position, and a display fed from a transferred source looks exactly like a display fed from its own — which is why the position of these selectors is treated as part of the aircraft configuration in the closing section of this note.

Colour and Symbology Conventions

Colour on an electronic display is a coding system, not decoration, and it is applied consistently enough across manufacturers to be examinable. The conventional assignments are:

Colour	Meaning	Typical use on the flight displays
Red	Warning — unusable, or requiring immediate action	Warning messages, failure flags, barber-pole speed limits, a resolution advisory
Amber	Caution — abnormal, requiring awareness and timely action	Caution messages, comparator annunciations, a traffic advisory
Green	Normal operation, engaged and active	Engaged mode annunciations, normal system states, active values
Magenta	Computed or commanded — what the aircraft is being asked to do	Flight director command bars, ILS localiser and glideslope deviation pointers, selected values and bugs, the active flight-plan leg
Cyan	Selected, preselected or inactive information	Selected altitude and heading readouts, inactive route segments, advisory data
White	Present status, scales and labels	Compass and tape scales, unit labels, memo-type messages

The magenta convention is the one candidates most often lose a mark on: it marks information that has been *computed for the pilot to fly to*, which is why both the flight director command bars and the ILS deviation pointers are drawn in it. Individual types vary in the details of their symbology, so the aircraft's own description of its displays is always the authority, but the underlying logic — red and amber reserved for alerting, magenta for commanded guidance — is common ground.

Flight Director Command Bars

The command bars show **the required flight path with respect to the actual flight path**. That direction matters and is easy to invert. The bars are not a display of what the aircraft is doing; the attitude picture

behind them already does that. They are the output of the flight director computation, which takes the selected mode and target — a heading, a track, an altitude, a glideslope — compares it with what the aircraft is currently achieving, and displays the pitch and roll that would remove the difference. The pilot or the autopilot flies to centre the bars on the fixed aircraft symbol. Following the bars satisfies the required path; the bars themselves never show a bank angle for its own sake.

On an electromechanical attitude director indicator the bars are physically positioned by servos, and two feedback terms in that servo loop are examined separately because they control different things:

- **Position feedback** sets *where* the bars come to rest. The servo compares the bar's actual position with the demanded position and drives until that error is nulled, so the bars settle precisely on the commanded value instead of stopping short or overshooting it.
- **Rate or velocity feedback** sets *how fast* they get there. A signal proportional to the speed of movement is fed back in opposition to the drive, which damps the loop. Increase the rate feedback and the bars move more slowly and settle without oscillating; remove it and they chase the demand, overshoot and hunt about it.

On an electronic display the bars are drawn rather than driven, so there is no servo to fail, but the same two terms survive inside the computation as gain and damping. The distinction remains a fair examination question because it is really about control loops rather than about instruments.

Reading an ILS Approach on the Attitude Display

Deviation indications on the attitude display follow one rule, and applying it mechanically prevents the reversal candidates most often make: **the pointer shows where the beam is, relative to the aircraft.** The aircraft symbol is fixed, so the pointer moves and the pilot flies towards it.

- Glideslope pointer **below** the centre mark: the glidepath is below the aircraft, so the aircraft is **above** the glideslope and must descend to recapture it.
- Glideslope pointer above the centre mark: the glidepath is above the aircraft, so the aircraft is below it and must reduce its rate of descent.
- Localiser pointer displaced to the left: the runway centreline lies to the left of the aircraft, so the aircraft is right of the centreline and must turn left.

The **rising runway** symbol on the attitude display works on the same fly-towards-it principle but is driven by a different source. Its vertical position comes from the **radio altimeter**, not from pressure altitude and not from rate of descent, and it appears only in the low band where the radio altimeter is meaningful. As the aircraft descends the radio height decreases and the symbol rises towards the fixed aircraft symbol, so the two meet at touchdown. Test that against the opposite case to be sure of the direction: if the symbol is seen *moving*

down during the final stages of an approach, the radio height must be increasing — the aircraft is going the wrong way relative to the ground and must be flown down.

> ✈ **Aviation context:** the rising runway also gives the engineer a free functional check. Because it is driven by radio altitude, its behaviour during a radio altimeter test tracks the height the test set is injecting. A rising runway that does not follow the injected height, or that appears at the wrong height band, points at the radio altimeter chain rather than at the display.

Traffic and Weather Symbology

The two most heavily coded overlays on the navigation display are traffic and weather, and both use shape and colour together so that the picture can be read without interpretation.

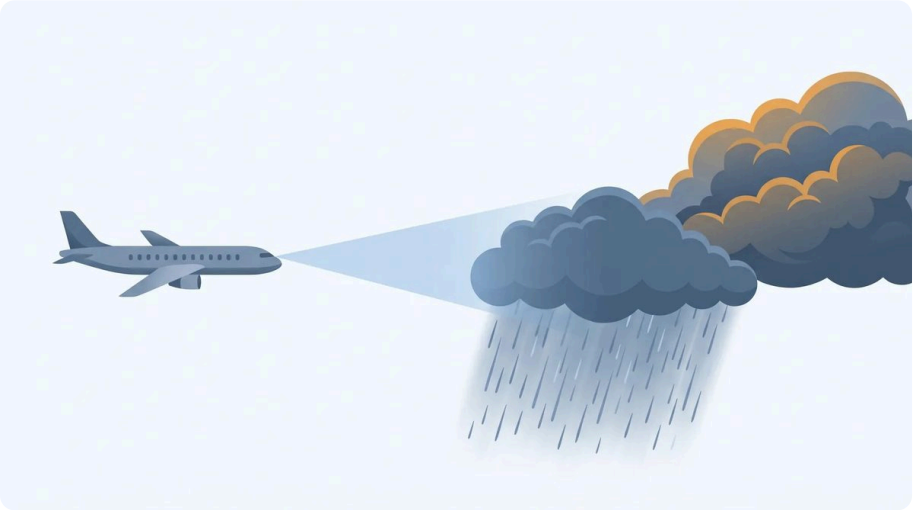
Airborne collision avoidance symbology escalates through four steps. Other traffic within the display's selection criteria is drawn as an **open white or cyan diamond**; traffic closing to within the proximate band is the same **diamond filled in**; a **traffic advisory** is shown as a **filled amber circle**, telling the crew to look for the aircraft; and a **resolution advisory** becomes a **filled red square**, at which point the vertical guidance to resolve it is flown on the attitude display while the traffic itself remains plotted on the navigation display. Each symbol carries the relative altitude in hundreds of feet, with a sign showing above or below, and an arrow when the intruder is climbing or descending.

Weather radar returns are coloured by the strength of the echo, which corresponds to the rate of precipitation the beam is being reflected from:

Colour on the display	What it represents
Black	No significant return — either clear air, or nothing reaching the antenna
Green	Light returns
Amber or yellow	Moderate returns
Red	Heavy returns — the core of a severe cell, shown as a red area against the black of the surrounding clear air
Magenta	Turbulence, on radars with the Doppler processing needed to detect it

Black deserves a caution rather than comfort. The radar sees precipitation, so a black area may be clear air, or it may be an area the beam never reached because a nearer cell attenuated it — a shadow behind heavy rain looks exactly like clear weather. This is a limitation of the sensor rather than of the display, but it is

displayed information that an engineer investigating a "radar not painting" report has to understand before deciding whether there is a defect at all.



A nearer heavy-rain cell attenuates the radar beam, leaving the area behind it in shadow that looks clear.

Brightness, Readability and Night Operations

A display that cannot be read is a failed display, and the range of conditions it has to be readable in is extreme: direct sunlight over the glareshield at one end, and a night approach at the other where any excess brightness destroys the crew's dark adaptation. Three mechanisms work together.

Automatic compensation. A light-dependent sensor mounted on the flight deck measures the ambient lighting and its signal is fed in parallel to the display units, so that all of the screens track the cockpit lighting together and none of them ends up brighter or dimmer than its neighbours. Using one flight-deck sensor for all displays rather than a separate sensor inside each unit is deliberate: what matters to the crew is that the screens are consistent with one another, and a sensor buried in one instrument panel position would read a different local light level from the others. The crew retain a manual brightness control to override the automatic setting.

Sunlight readability. Being bright is not sufficient by itself, because in daylight the dominant problem is reflected light washing out the image rather than a lack of emitted light. Sunlight-readable units combine a high-luminance backlight or beam current with anti-reflective and anti-glare treatment of the screen surface, and the glareshield above the panel exists for the same reason. The coating on the face of a display unit is therefore a functional part of it rather than a finish, with the consequences for handling and cleaning set out in the closing section of this note.

Night operations. Dimming for night reduces brightness so that the screens do not impair night vision, and on aircraft required to operate with night vision goggles a further constraint applies. Goggles amplify light in the near-infrared as well as the visible band, so an ordinary display emitting in the near-infrared saturates them and destroys the intensified image of the outside world. Compatibility is achieved by filtering the emitted spectrum of the displays and of the panel lighting to suppress that band, not by simply turning the brightness down. A cockpit modified for goggle use therefore has display and lighting components that are not interchangeable with the standard parts, however identical they look.

EICAS and ECAM

System	Manufacturer	Primary Function	Key Feature
EICAS	Boeing	Engine parameters + crew alerting; modern implementations also provide system synoptic pages and electronic checklists	Messages colour-coded by severity (red = warning, amber = caution, amber and indented one space = advisory; white = memo/comm)
ECAM	Airbus	Engine parameters + crew alerting + system synoptics	Automatically displays the relevant system page when a fault occurs; provides interactive checklists

One Job, Two Design Philosophies

Whatever the badge on the front, an engine and systems display has the same four jobs: keep the parameters needed to operate the engines permanently in view, tell the crew when something needs attention and how urgently, make the detail of any system available when it is wanted, and hand what happened to maintenance afterwards. The two families differ in how the crew are brought to the information and in how the computing is divided up, and those differences are what an engineer notices when moving between fleets.

Aspect	Boeing practice (EICAS)	Airbus practice (ECAM)
Upper centre display	Primary engine parameters and the crew alert message list	Engine/warning display — primary engine parameters, fuel on board, slat and flap position, the warning and caution messages, and memo items
Lower centre display	Secondary engine parameters, system synoptics, checklists, status	System display — one system synoptic page at a time, plus the status page
How a page is chosen	Crew selection from the display select panel is the normal route; secondary parameters can appear automatically when one goes out of limits	The relevant page is called up **automatically** by the failure or by the phase of flight; the crew can override the selection manually

Aspect	Boeing practice (EICAS)	Airbus practice (ECAM)
Where the alerting logic lives	Interface units on earlier types, and a shared computing cabinet on later ones, feeding the display processing	Flight warning computers generate the warnings, the aural alerts and the message text; system data acquisition concentrators gather the system data used for cautions and synoptics; display management computers draw the result
Crew procedure presentation	Alert message plus an electronic checklist called up for the failure	The message line itself expands into the actions to be carried out, item by item, cleared as they are completed
Redundancy of the alerting function	Duplicated interface or core computing, with the display set able to be reconfigured	Two flight warning computers and two data concentrators; the most critical warnings are generated by the warning computers from inputs taken directly, so they survive the loss of the concentrators

The automatic page call-up is the genuine philosophical difference and it is worth stating carefully, because it is easy to overstate into a false contrast. Both families provide system synoptics, both provide status information, and both provide electronic checklists on current implementations. What differs is who decides when the crew see the relevant system page: on one the failure brings the page up, on the other the crew go and get it. Everything else in the comparison is a matter of layout and naming rather than capability.

Alert Levels and How They Are Prioritised

Crew alerting is built on a hierarchy of urgency, because an undifferentiated list of messages would leave the crew doing the triage at exactly the moment they have least capacity for it. Three levels are used, and each is defined by the response it demands rather than by the system it comes from:

Level	What it means	Attention-getter	Expected crew response
Warning (red)	A condition requiring *immediate* action	Red master warning light plus an aural alert — a continuous repetitive chime, or a dedicated sound such as the fire bell	Act now, then read
Caution (amber)	A condition requiring immediate crew *awareness* and subsequent action	Amber master caution light plus a single chime	Note it, complete the current task, then deal with it
Advisory	A condition requiring crew awareness only, with monitoring rather than action	No master light and no aural alert — the indication is the message or the automatic appearance of the affected page	Monitor; no immediate procedure

Note how the attention-getters are graded with the level: an aural alert cannot be ignored and is therefore reserved for the conditions that must not be missed, while an advisory has no attention-getter at all because

getting attention is not what it is for. The certification requirement behind this is CS-25.1322, which prescribes red for warnings and amber or yellow for cautions and permits any colour except red or green for advisories. Within that framework each manufacturer has its own scheme, so the colour a message is drawn in tells you the level only once you know whose flight deck you are standing on.

Two further behaviours are common to both families and are regularly examined:

- **Ordering.** Messages are listed by priority class, warnings above cautions above advisories, and within a class the most recent message appears at the top of that class. The crew therefore always read the most urgent condition first, regardless of the order in which things went wrong.
- **Phase-related inhibition.** Alerts that are not immediately actionable are suppressed during the highest-workload moments — the take-off roll and rotation, and the final stage of the landing — so that the crew are not presented with a message at the point where their attention must be outside. The inhibited alerts are not lost; they appear once the aircraft is clear of the inhibited phase. An engineer running an engine or a systems test on the ground should expect a different set of messages from the same fault depending on the phase logic the aircraft believes it is in.

The ECAM Colour Code

The Airbus scheme is the most systematic of the two and is worth learning as a set, because the same code is used on the warning display, the system pages and the status page:

Colour	Meaning in the ECAM code
Red	A configuration or failure requiring immediate action
Amber	A condition the crew must be aware of, but with no immediate action
Green	Normal operation, and the memo items reminding the crew of a normally selected configuration
White	Titles and guidance text, and the items remaining to be completed
Blue (cyan)	Actions to be carried out and limitations to be observed
Magenta	Particular messages, such as an indication that an alert has been inhibited

Reading the code changes how a page is used. On a system synoptic, green means a component is doing its job and amber means it is not, so a page can be scanned for colour long before any label is read — which is the whole point of drawing systems as pictures. On the warning display, a line of white text is something still to be done and a line of blue text is the action itself, so the crew's place in a procedure is visible at a glance.

✈ **Aviation context:** the two schemes are both correct and they are not interchangeable. Boeing's alert list keeps advisories in the caution colour and distinguishes them by indenting them one space, with white reserved for memo and communication messages that are not crew alerts at all. Airbus keeps the alert list to red and amber and uses green, white, blue and magenta elsewhere in the code, so an advisory-level condition typically shows itself by calling up the affected page and pulsing the parameter concerned rather than by adding a coloured line. Learn one scheme thoroughly and know that the other exists; do not assume a colour carries the same meaning across the ramp.

Engine Indication

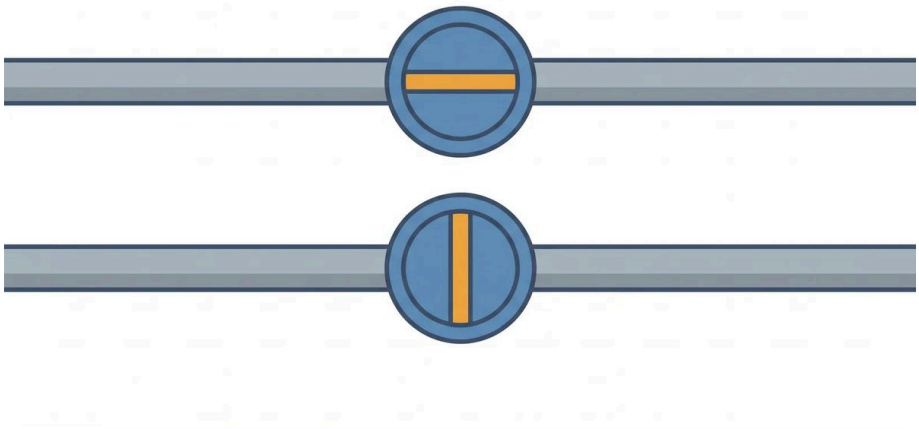
Engine parameters are split into two groups, and the split is about what must never be off the screen rather than about importance in the abstract:

- **Primary parameters** — the thrust-setting parameter (fan speed or engine pressure ratio, depending on the engine), exhaust gas temperature, core speed and fuel flow. These are needed on every take-off and every thrust change, so they are displayed permanently on the upper centre screen and cannot be deselected.
- **Secondary parameters** — oil pressure, oil temperature, oil quantity, vibration and similar. They are needed when a question arises rather than continuously, so they live on the lower screen. On many types they appear automatically if one of them goes out of limits, which is a small but revealing piece of design: the system decides that a parameter has become primary the moment it stops being normal.

Exceedance is shown by colour on the scale and on the digital readout: a normal band, an amber band where a limit is being approached or a time-limited region entered, and a red limit beyond which the parameter must not go. The readout itself changes colour and is usually boxed when it enters those regions, so the crew see the exceedance in the number as well as on the scale. On electronic displays the same parameter can be drawn as a round dial, an arc, a vertical tape or a bare digital readout on different types, and none of those is more correct than another — the format is chosen for the scan, and the underlying signal is identical.

System Synoptics and Status

A synoptic page is a schematic of one system drawn to be read as a picture: tanks, pumps, valves, ducts and electrical paths in their real topology, coloured by their current state rather than annotated with numbers. Valve symbols are drawn in line with the duct they sit in when they are open and across it when they are shut, so a fuel or bleed configuration can be verified in a single glance. Flow paths in a normal condition are drawn in the normal-operation colour, and a faulty or unpowered item changes colour where it sits, so the eye is taken to the location of the fault rather than to a list.



Synoptic valve symbols: the bar in line with the duct means open, the bar across the duct means shut.

Some parameters are colour-banded on these pages in the same way engine limits are. Brake temperature is the standard example: each brake's temperature is displayed with the colour changing as the temperature rises through defined bands, so the crew can see at a glance which brake is hot and how hot, without comparing numbers against a limit in a manual. That is a genuinely different design choice from displaying one fixed colour with a numeric value beside it: different temperatures are shown in different colours, and the colour is the indication rather than a decoration on it.

The status page is the one an engineer meets most often. It lists what is still inoperative and what limitations remain after all the actions have been completed — the residue of the failure rather than the failure itself. It is the crew's dispatch reference and it is also the fastest way, on arrival, to find out what the aircraft thinks is wrong with it before opening any manual.

Electronic Checklists

Electronic checklist systems present procedures on a centre display or a dedicated screen, covering *both* normal procedures and non-normal ones. Their significant feature is that many items are **sensed**: the system reads the actual state of the system concerned and marks the item complete by itself, so a switch that has been selected shows as done without the crew touching the checklist. Items that cannot be sensed — anything the aircraft has no way of detecting — must be actioned and closed by the crew. Items that cannot be completed now can be deferred and remain visible as outstanding.

The value is not that it saves reading. It is that the checklist state and the aircraft state cannot silently diverge: a paper checklist can be ticked while the switch is still in the wrong position, and a sensed electronic one cannot. It also links directly to the alert list, so the procedure for a failure is one selection away from the message announcing it.

What the Alerting System Hands to Maintenance

The crew see a *cockpit effect*: a warning, a caution, a flag, a lost indication. The maintenance system sees a *fault message*: the failure the monitoring detected, identified down to a unit or a wiring interface. The two are not the same and the whole value of the maintenance interface lies in the correlation between them — one root failure can produce several cockpit effects, and several cockpit effects reported by the crew can turn out to be one message in the maintenance record.

Fault messages are classified by their operational consequence. A fault with an associated cockpit effect is the highest class and normally has to be dealt with before further flight; a fault with no cockpit effect but a bearing on dispatch capability forms the next class, to be rectified within a defined period; and faults with neither are recorded for trend purposes only. Alongside this, aircraft condition monitoring records performance data and produces reports used for trend monitoring rather than for immediate rectification.

For the engineer, the practical consequence is a discipline: **record the cockpit effect exactly as the crew saw it, and read the fault messages separately.** A crew report of "engine display blank" and a maintenance message naming a data bus are two halves of the same event, and rectifying one without confirming it explains the other is how an aircraft leaves with the defect still latent.

✈ **Aviation context:** Modern aircraft have reversionary capability — if a display fails, its information can be transferred to another screen. For example, the captain's PFD information can be displayed on the centre MFD. Some aircraft also retain a small set of analogue standby instruments (attitude, airspeed, altitude) as an ultimate backup.

Redundancy and Reliability

Electronic instrument systems use multiple levels of redundancy:

- **Dual (or triple) symbol generators** — if one fails, another takes over.
- **Display switching** — any display unit can show information from any SG.
- **Independent power supplies** — displays fed from different electrical buses.
- **Standby instruments** — independent analogue or self-contained electronic instruments for attitude, airspeed, and altitude.
- **BITE (Built-In Test Equipment)** — continuous self-monitoring with fault logging for maintenance.

What the Redundancy Has to Achieve

The layers listed above are not there to improve dispatch reliability, although they do. They exist because the airworthiness requirements grade every failure condition by its effect and then demand a probability to match: the more severe the consequence, the less often it may be allowed to happen. Total loss of the primary flight information presented to both pilots is at the catastrophic end of that scale, so it must be shown to be extremely improbable — of the order of 10^{-9} per flight hour — and no single failure may be permitted to cause it.

That requirement drives the whole architecture, and it explains a distinction that is easy to miss. Duplicating equipment only helps if the duplicates cannot be taken out by the same event. Two display units fed by one air data computer are not two sources of airspeed, they are one source shown twice; two symbol generators supplied from the same busbar are not independent of a busbar failure; and two identical computers running identical software are not independent of an error in that software. Real independence therefore has to be built at four levels at once — the sensing, the computing, the electrical supply and the physical routing — and it is backed up by a standby chain that is *dissimilar* rather than merely additional.

Independent Sources, Not Just Independent Boxes

Behind the displays sit multiple independent sensing chains. A large transport typically carries three air data and inertial reference sources rather than two, and separate pitot and static systems for the left side, the right side and the standby instruments, with their probes physically separated on the airframe so that one impact, one bird or one blocked drain cannot affect them all. Heating is applied to each independently for the same reason.



Left, right and standby probes sit apart on the airframe, so one bird strike cannot reach them all.

Three is a deliberate number rather than generosity: as the discussion of comparators earlier in this note sets out, two sources can be compared but not adjudicated, and the third is what resolves the ambiguity. The source selection switches on the flight deck exist to make use of that: if one air data or inertial source is known to be faulty, the affected side can be transferred to a healthy one, which restores a valid picture without waiting for maintenance. The cost is that the two sides may then be sharing a source, and the crew have to know that their cross-check has become weaker.

Electrical Segregation and Power

Electronic displays stop when the electrical supply stops, so the powering arrangement is part of the instrument system rather than a separate subject. The displays and their symbol generators are split deliberately across different busbars, so that the loss of any one bus cannot darken both pilots' screens. The essential or emergency bus is arranged to keep a minimum flying set alive — typically one pilot's primary flight and navigation displays plus one engine and systems display — and that bus is the one that stays alive on the battery when normal generation is lost, through a static inverter where an alternating-current supply is still needed, and on many types on a ram air turbine driving an emergency generator.

The standby instruments are then arranged one level further out again, on the battery bus with their own internal reserve, so that they survive the loss of everything else for a defined period. When tracing a display defect it is always worth establishing which bus the affected unit is on before suspecting the unit, because the pattern of what is lost and what survives usually identifies the supply immediately: a fault that takes out one pilot's screens and nothing else is on that side's bus or that side's equipment, while a fault that takes out screens on both sides is upstream of the split.

Reversion and Switching in Practice

Reversionary capability is what makes a display failure an inconvenience rather than an emergency. In a reversionary mode the essential information from a failed screen is transferred onto a screen that is still working, so the crew keep attitude, airspeed, altitude and the engine indications even though there are fewer screens to put them on. The consequence to understand is that the surviving display is now doing two jobs, so the format is compacted: something has to give way, and what gives way is the information the crew can most afford to lose. A reverted display therefore shows less than the two displays it replaced, and that is a design choice rather than a fault.

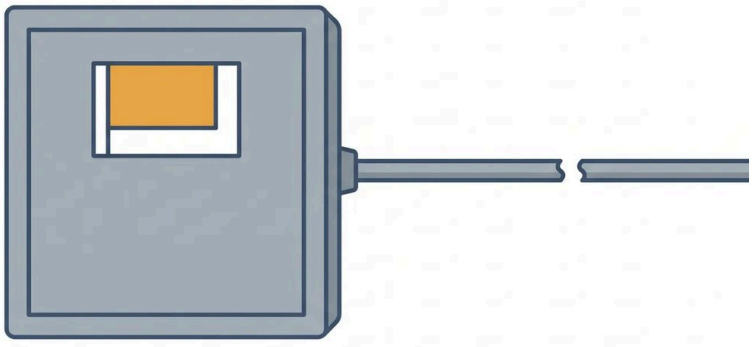
Two separate switching layers are involved and they are often confused in reports. Selecting a different **source** changes which symbol generator or sensor drives a display, and the picture stays where it was. Selecting a **reversionary format** changes what a display draws, moving information from one screen to another. A crew report that mentions switching should always be read carefully to establish which of the two was done, because they point at different halves of the system.

Recognising a Failure on the Flight Deck

Because everything travels the sensor, bus, computer, generator, display chain described earlier in this note, the *pattern* of what is missing localises the failure before any test equipment is connected. The reasoning is always the same: whatever is still working tells you which parts of the chain are healthy, and the fault must lie in something common to everything that is not working.

What is seen	Where the fault most probably lies	Why the symptom says so
One screen completely dark, the others normal	That display unit, its supply or its cooling	Nothing common to the other screens is affected, and a dark screen is not drawing anything at all
Screen alive but showing only a large cross	The symbol generator feeding it, or the switching to it	The display unit is healthy enough to draw the diagnostic cross — what it is not receiving is valid symbology
Part of one display missing, the rest correct	The symbol generator handling that symbology, or the input sensor supplying that parameter	The display unit and the basic generation are proven alive by everything still drawn around the gap
The same single item missing from *every* display	Something common to all of them — the display controller and symbol generator function, or the single source or bus that feeds that item to all of them	A fault inside one display unit cannot reach the others; only a shared element can remove one item everywhere
A flag in view, a dashed field or a boxed cross over one window	The source of that parameter, which is declaring its own data invalid	The data is being *refused*, not lost; the display is doing exactly what it should
Comparator caution such as airspeed, altitude or attitude	One of the two compared sources — not yet identified	The comparison only proves disagreement; the third source or the standby instruments identify the outlier
Correct symbology but a dim or discoloured picture on one unit	The display unit itself — backlight or tube ageing	The data is plainly arriving and being drawn correctly
All screens lost together	Power or cooling, upstream of everything	Independent units do not fail simultaneously; a common service does

The flag convention behind the flag-in-view row is old and worth stating explicitly because it is counter-intuitive at first: warning flags are biased **into** view and are pulled **out** of view by the system when the data being monitored is valid. A heading flag that is out of sight therefore means the heading information is good, and a flag in view means it is invalid or is not being monitored. The arrangement is deliberately fail-safe — lose the power or the signal that holds the flag away and it falls into view by itself, so the failure state is the default and no loss of supply or drive can leave the flag hidden while the data it monitors is invalid.



Lose the power or signal that holds it away and the warning flag falls into view by itself.

Standby Instruments

The standby set is the last line, and its value lies entirely in being unlike everything else. It provides the three parameters needed to fly the aircraft safely — attitude, airspeed and altitude — from its own sensing, over its own wiring, on its own supply, so that no failure of the display system, its computers, its buses or its normal power can affect it. On many current types the three functions are combined in a single integrated standby unit with a small electronic screen of its own, which typically adds a heading indication and ILS deviation, and which contains its own attitude sensor and its own reserve battery. The unit needs a short alignment on the ground before its attitude output is usable, which is why it is switched on early rather than at the last minute.

The dissimilarity is the point. An integrated standby unit is a screen, like the ones it backs up, but it does not share their sensors, their computers or their supply, and it comes from a separate design rather than being another copy of the one it backs up. That is what makes it a genuine backup rather than a fourth copy of the same failure. Note also what it does *not* provide: no navigation picture, no engine or systems indications, no flight director. It is the minimum needed to keep the aircraft the right way up and within its speed and altitude limits, and nothing more.

Built-In Test and the Maintenance Trail

Continuous self-monitoring is what allows this much redundancy to be managed at all: with several units able to cover for each other, the aircraft could otherwise dispatch with hidden failures that only reveal themselves when the last healthy unit goes. Built-in test works at three levels — a power-up test that exercises the unit before flight, continuous monitoring during operation that is what raises flags and comparator alerts in the first place, and an interactive ground test that the engineer initiates through the maintenance system.

Three practical points repay attention:

- **Faults are stored by flight leg.** The record is what makes an intermittent fault tractable, because a failure that will not repeat on the ground still leaves its trace. Reading the stored record before disturbing anything is nearly always the right first action, and clearing it before it has been read destroys the evidence.
- **The test tells you which monitor tripped, not always which unit failed.** A unit reporting a missing or invalid input is reporting a symptom whose cause lies upstream in the transmitting equipment or the wiring between them. Treating the reporting unit as the faulty one is the classic way to replace a serviceable box.
- **Ground tests need the aircraft in the configuration the test assumes.** Systems must be powered and pressurised as the manual requires, and safety precautions such as gear pins, area clearance and control-surface warnings apply, because a display test can command real system responses.

Line Maintenance of Display Units and Symbol Generators

Display units and symbol generators are line-replaceable, and on most current types the display units are identical hardware whose role — primary flight display, navigation display, engine display — is set by **software configuration or by position programming** in the rack rather than by the part number of the unit. That is why one spare covers several positions, and it is also why fitting a unit is not finished when the fasteners are torqued: an incorrectly configured display can power up and look entirely convincing while presenting the wrong format for its position.

- **Confirm the configuration and the software standard**, not just the part number on the label. Two physically identical units can hold different software.
- **Restore every source-select and reversion switch to its normal position** before signing, and confirm it visually. A transferred selector left over from troubleshooting leaves an aircraft dispatching in a degraded configuration that shows nothing abnormal on the screens.
- **Cross-check the new unit against the opposite side** with the aircraft powered: identical parameters, identical formats, no comparator annunciations. This is the test that proves the whole chain, where a self-test only proves the unit.
- **Check the cooling.** Display units and their computers are cooled by forced air, and a blocked filter or a failed fan produces overheating symptoms — dimming, automatic shedding or a shutdown — that mimic a failing unit.
- **Clean only with the agent named in the maintenance manual.** The face of a display carries an anti-reflective and anti-glare coating that solvents and household cleaners will attack, and a damaged coating permanently degrades sunlight readability on a unit that is otherwise perfectly serviceable.
- **Apply the electrostatic-discharge precautions** covered elsewhere in this module to the units and, above all, to any card handled inside them.
- **Judge failed pixels against the published criterion** in the maintenance manual, which sets how many may be tolerated and whether their position and clustering matter, rather than against an impression that the screen looks imperfect.

✈ **Aviation context:** because the system is so redundant, an aircraft can often be dispatched with a display unit or a symbol generator inoperative under the minimum equipment list, subject to its conditions and to any placarding and crew notification the list requires. Two consequences follow for the engineer. First, the aircraft leaves with one layer of protection already spent, so the deferral has to be recorded accurately — the next failure is not the first one the crew will have handled. Second, the remaining serviceable units are now doing more work: a fleet that habitually defers display defects will show it in a rising rate of consequential findings, which is exactly the sort of trend the recorded fault data exists to reveal.

⚠ **Maintenance note:** When performing display unit or symbol generator replacements, always follow the aircraft maintenance manual (AMM) procedures for initialisation, alignment, and testing. After installation, the display configuration may need to be programmed and a full functional test performed before return to service.

Module 6 — Materials and Hardware

6.1 — Aircraft Materials — Ferrous

Introduction to Ferrous Materials

Ferrous materials are metals and alloys whose primary constituent is **iron (Fe)**. The word "ferrous" comes from the Latin *ferrum*, meaning iron. Ferrous materials are characterised by their high strength, hardness, and — with the notable exception of stainless steel — their susceptibility to corrosion (rust). In aircraft construction, ferrous materials are used selectively where their superior strength and wear resistance are essential, despite their relatively high density compared to aluminium alloys.

The most important ferrous material in aviation is **steel** — an alloy of iron and carbon (typically 0.05% to 2.0% carbon by weight). By adding other alloying elements and applying specific heat treatments, the properties of steel can be tailored to a wide range of requirements.

Why an Aircraft Uses Steel at All

The first thing most students are told about steel is that it is heavy, and the density figure listed among its properties later in this note reinforces that impression. Taken on its own it is misleading, and it leads to the wrong mental model of when a designer will reach for a ferrous material. Weight per unit of *volume* is not the quantity that matters in a structure; weight per unit of *load carried* is. That is called **specific strength**, and it is simply the ultimate tensile strength divided by the density.

Material	Typical UTS	Density	Specific strength (UTS / density)	Specific stiffness (E / density)
Low-alloy steel, heat treated (e.g. 4340)	about 1,800 MPa	about 7,850 kg/m³	about 229 kN·m/kg	about 26 MN·m/kg
High-strength aluminium alloy	about 570 MPa	about 2,810 kg/m³	about 203 kN·m/kg	about 26 MN·m/kg
Titanium alloy	about 950 MPa	about 4,430 kg/m³	about 214 kN·m/kg	about 26 MN·m/kg

Two results in that table are worth pausing over, because both are counter-intuitive. First, a properly heat-treated low-alloy steel has a **higher** specific strength than a high-strength aluminium alloy, not a lower one. Per kilogram of material, the steel carries more tensile load. It is not a heavy material in the sense that matters to a designer; it is a dense material, which is a different thing. Second, the specific stiffness of steel, aluminium and titanium is **essentially identical** at around 26 MN·m/kg. Young's modulus and density rise together as you move between the common structural metals, so you cannot make a stiffness-critical part lighter simply by changing which metal it is made from. You change the *section* instead — and that is exactly why a wing skin is aluminium: for the same mass you get a thicker sheet, and a thicker sheet resists buckling far better.

So the real reasons an aircraft carries ferrous parts are these:

- **Load in a small envelope.** A landing gear leg must absorb an enormous load inside a bay that has to close. Steel puts the required strength into the smallest volume of any common airframe material, and volume, not mass, is what constrains a gear bay, a bearing housing or a bolt hole.
- **Bearing and wear surfaces.** Nothing else available at reasonable cost survives repeated rolling or sliding contact the way a hardened steel surface does. Ball races, gear teeth, cam faces, bushes and pins are ferrous almost without exception.
- **Temperature.** Aluminium alloys lose useful strength above roughly 150 °C. Steels, and stainless steels in particular, keep working far beyond that, which is why firewalls, exhaust systems, engine bay structure and hot-air ducting are ferrous.
- **Stiffness in a slender member.** Because Young's modulus is roughly three times that of aluminium, a steel rod of a given diameter is three times as stiff. For a long, thin, tension-loaded member such as a control cable or a tie rod, that matters more than density.
- **Weldability and repairability.** A welded steel tube fuselage or engine mount can be repaired with equipment available in a small workshop, which is why the light aircraft fleet still uses it.
- **Cost.** Steel is the cheapest structural metal by a wide margin, so it is the default anywhere the weight penalty is acceptable.

Where the Engineer Actually Meets Ferrous Materials

Aircraft area	Typical ferrous items	Why ferrous
Landing gear	Shock strut cylinders and pistons, side stays, drag braces, axles, torque links, pins and bushes	Very high load in a confined volume; wear surfaces; the highest-strength steels in the aircraft live here
Powerplant installation	Engine mounts, mount bolts, firewall, exhaust stacks and shrouds, turbocharger and bleed ducting, clamps	Strength plus resistance to heat, vibration and exhaust gas
Engine internals	Crankshafts, connecting rods, camshafts, gears, valve springs, bearings, turbine shafts and discs	Fatigue strength, wear resistance and hot strength
Flight controls	Control cables, turnbuckles, bellcrank bushes, push-pull rod ends, hinge pins	Tensile strength, stiffness and wear resistance in small sections
Structure and mechanisms	Flap and slat tracks and carriages, wing attachment fittings, door latches and hinges, seat tracks and studs	Concentrated point loads that would need an impractical volume of aluminium
Fasteners	Bolts, studs, screws, nuts, self-locking inserts, pins, some rivets	High clamp-up and shear capacity in a small diameter
Systems hardware	Hydraulic tubing in high-pressure zones, actuator rods, springs, valve internals, bearing races	Pressure containment, elasticity and wear resistance

The Ferrous Family

"Ferrous" is a family, not a single material, and the members differ enormously. Sorting them out by carbon content is the quickest way to hold the family in your head, because carbon content is what separates them.

Material	Approximate carbon	Character	Aircraft relevance
Pure iron	essentially none	Soft, very ductile, weak, rusts freely, strongly magnetic	None structurally; used for magnetic cores and shielding
Wrought iron	below about 0.08%	Almost pure iron containing threads of slag; tough, easily forged and welded, notably corrosion resistant	Obsolete. Found only on historic and vintage airframes
Plain carbon steel	0.05% to about 1.5%	Properties set almost entirely by carbon content and heat treatment	Tubing, brackets, general hardware, springs, tools
Alloy steel	usually 0.1% to 0.5%, though the tool grades and the through-hardening bearing grades run higher	Deliberate additions of chromium, nickel, molybdenum and others to control strength, toughness, hardenability and hot strength	The workhorse of highly stressed aircraft parts
Stainless steel	as little as 0.03% in the low-carbon welding grades, up to about 1.2% in the hardenable ones	At least 10.5% chromium, giving a self-repairing passive film	Firewalls, exhaust, fasteners, cable, hot and wet zones
Tool steel	0.6% to about 1.5%	Very hard and wear resistant after heat treatment; brittle	Workshop tooling, drills, reamers, dies; not airframe parts
Cast iron	2.0% to about 4.5%	Cheap, easily cast, excellent in compression, brittle in tension, good vibration damping	Effectively absent from airframes; some piston engine components and ground equipment

Note where the boundary between steel and cast iron falls. Above roughly 2% carbon the iron can no longer dissolve all of the carbon at any temperature, so the excess appears as a separate constituent that cannot be removed by heat treatment. That is the physical reason steel is defined as an iron-carbon alloy of up to about 2% carbon: it is the practical limit of the solid solution, not an arbitrary line.

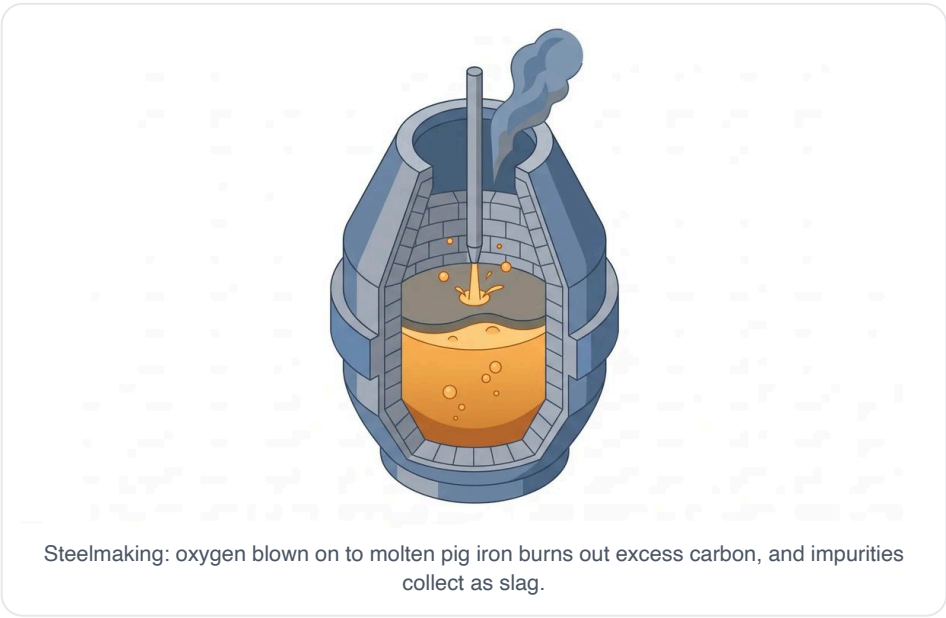
Why Pure Iron Is Not Used

Iron on its own is a poor structural material. It is soft, extremely ductile, and its tensile strength is only a small fraction of that of even a mild steel. It cannot be hardened by heat treatment either, because there is no carbon present to form the hard constituents that hardening depends upon. Note carefully that the objection to pure iron is **lack of strength**, not brittleness — pure iron is the opposite of brittle. It bends and stays bent.

Adding a fraction of one per cent of carbon transforms it, and adding the ability to heat treat it transforms it again.

How Steel Is Made

Understanding the manufacturing route explains several things an engineer meets later: why aerospace steel costs many times what commercial bar costs, why a mill certificate quotes a cast or heat number, and why some specifications insist on a particular melting process.



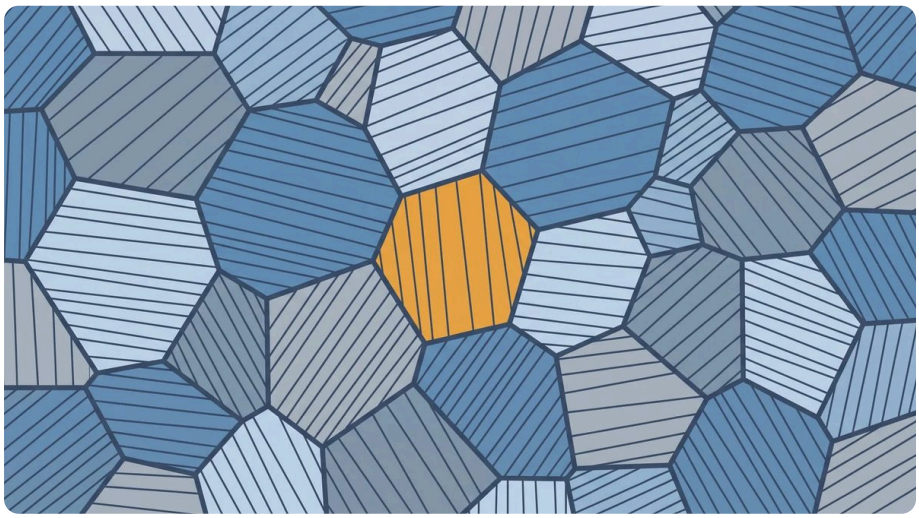
Steelmaking: oxygen blown on to molten pig iron burns out excess carbon, and impurities collect as slag.

- **Ironmaking.** Iron ore, coke and limestone are charged into a blast furnace. The coke reduces the ore to metallic iron and, in doing so, saturates it with carbon. The product, **pig iron**, carries roughly 4% carbon along with silicon, sulphur, phosphorus and manganese. It is hard, weak and brittle, and useless as it stands.
- **Steelmaking — removing carbon.** The molten pig iron is refined by blowing air or, in the modern basic oxygen process, pure oxygen through or on to it. The oxygen burns out the excess carbon as carbon monoxide, and also oxidises silicon, manganese and phosphorus into a slag. **The essential operation of steelmaking is the removal of carbon**, taking it down from about 4% to whatever the specification calls for. An electric arc furnace does the same job starting mainly from scrap.
- **Deoxidation and alloying.** Oxygen dissolved in the bath would form blowholes on solidification, so it is removed with aluminium or silicon. A fully deoxidised steel is called *killed*; the aerospace grades are always killed. Alloying elements are added at this stage.
- **Secondary refining.** For highly stressed aerospace parts, ordinary air-melted steel is not good enough. Non-metallic inclusions left over from melting act as internal notches and become fatigue crack starters. Premium routes such as **vacuum induction melting**, **vacuum arc remelting** and **electroslag remelting** re-melt the steel under vacuum or under a refining slag to strip out those inclusions and dissolved gases. The improvement shows up mainly in transverse ductility and fatigue life rather than in the tensile strength figure, which is precisely why a specification, and not the tensile test alone, has to control it.
- **Forming.** The steel is cast, then hot rolled, forged, drawn or extruded into bar, sheet, tube or forgings.

✈ **Why the melting route appears on a drawing:** when an aerospace specification calls for a steel to be "consumable electrode vacuum melted" or "VAR", it is not specifying a chemical composition — the chemistry would be identical without it. It is specifying *cleanliness*. Two bars of 4340 with identical certificates of analysis and identical tensile results can differ by a large factor in fatigue life if one is air melted and the other vacuum remelted. This is the clearest example in the module of why a material is defined by its specification and not by its name.

Four Words You Need Before Going Further

- **Alloy** — a metallic material made from two or more elements, of which at least one is a metal. Steel is an alloy of iron and carbon even though carbon is not a metal.
- **Solid solution** — one element dissolved into the crystal lattice of another while both remain solid, in the same sense that salt dissolves in water. The dissolved atoms are distributed uniformly and do not form a separate substance.
- **Phase** — a region of material that is uniform in both composition and crystal structure. A steel at room temperature normally contains two phases at once, and knowing which they are is the whole basis of heat treatment.
- **Grain** — a single crystal within the metal, typically a few hundredths of a millimetre across. A solid metal is a mosaic of grains, each with its lattice pointing in a different direction, meeting at **grain boundaries**. Grain size turns out to be one of the few things that improves strength and toughness at the same time.



A solid metal is a mosaic of grains, each a single crystal with its lattice pointing a different way.

The Iron-Carbon System

Everything that follows in this note — why carbon content decides a steel's character, why heat treatment works at all, why one steel can be hardened and another cannot — comes out of a single piece of physics: **iron changes its crystal structure when it is heated, and the two structures dissolve wildly different amounts of carbon.** Learn this section properly and the rest of the material stops being a list to memorise.

Iron Changes Its Crystal Structure With Temperature

Most metals keep one crystal structure from room temperature up to melting. Iron does not. It is **allotropic**, meaning it exists in more than one crystalline form, and it swaps between them at fixed temperatures on the way up:

Temperature range (pure iron)	Name	Crystal structure	Atoms per unit cell	Carbon it can dissolve
Room temperature to 910 °C	Alpha iron (ferrite)	Body-centred cubic (BCC)	Corner atoms plus one at the centre of the cube	Almost none — about 0.02% at best
910 °C to 1,394 °C	Gamma iron (austenite)	Face-centred cubic (FCC)	Corner atoms plus one at the centre of each face	Up to about 2.0%
1,394 °C to 1,538 °C	Delta iron	Body-centred cubic (BCC)	As alpha iron	Very little; of interest only in casting and welding
Above 1,538 °C	Liquid	None	—	Unlimited within the alloy range

The face-centred cubic structure is the more tightly packed of the two, which sounds as though it ought to have less room for carbon, not more. The resolution is that what matters is not how much empty volume there is in total, but how big the *individual* holes between the iron atoms are. FCC iron has fewer holes but each is larger, and a carbon atom fits into one of them. BCC iron has more holes but every one of them is too small, so a carbon atom forced into the BCC lattice has to push the surrounding iron atoms apart. That single geometric fact drives the whole of steel heat treatment.

The Constituents of Steel

The names below recur throughout the rest of this note, so learn them here. Each is a distinct constituent with its own structure and its own mechanical character.

Constituent	What it is	Crystal structure	Carbon content	Hardness and character	Magnetic?
Ferrite (alpha iron)	Almost pure iron; a solid solution holding a trace of carbon	Body-centred cubic	up to about 0.02%	Soft, weak, very ductile	Yes, strongly
Austenite (gamma iron)	Carbon fully dissolved in iron as a single uniform solid solution	Face-centred cubic	up to about 2.0%	Soft and very ductile at temperature; the starting point of every hardening treatment	No
Cementite (iron carbide, Fe3C)	A chemical compound, not a solution: three iron atoms combined with one carbon atom	Orthorhombic compound	6.67% by weight, fixed	Extremely hard, extremely brittle; the hard constituent of ordinary steel	Weakly, below about 210 °C

Constituent	What it is	Crystal structure	Carbon content	Hardness and character	Magnetic?
Pearlite	Not a phase but a mixture: alternating thin plates of ferrite and cementite	Lamellar (layered) mixture	0.83% overall	Strong without being brittle — the ferrite carries the ductility, the cementite the strength	Yes
Martensite	Carbon trapped in a distorted iron lattice by cooling too fast for it to escape	Body-centred tetragonal — a BCC cell stretched along one axis	Whatever the steel contained	The hardest constituent; also the most brittle and the most highly stressed	Yes
Bainite	Fine needles of ferrite with carbide between them, formed at cooling rates between those giving pearlite and martensite	Acicular (needle-like)	Whatever the steel contained	Hard, and notably tough for its hardness — a useful compromise between pearlite and martensite	Yes

Two of those entries are the ones examiners return to again and again, so state them to yourself precisely. **Austenite is carbon fully dissolved in iron and uniformly distributed in solid solution. Cementite is one carbon atom chemically combined with three iron atoms**, written Fe3C. And pearlite is neither: it is **cementite laminated with ferrite in alternate layers**, so it combines the properties of ferrite and cementite rather than being a substance in its own right.

✈ **The three-property triangle:** ferrite is *soft and weak*; cementite is *strong but too hard and brittle*; pearlite, being a mixture of the two, is *strong without being brittle*. Those three phrases answer a whole family of exam questions between them. The engineering point behind the wording is that neither constituent is usable alone — a component made entirely of ferrite would bend under load and one made entirely of cementite would shatter. It is the fine mechanical mixture of the two that makes steel useful, and heat treatment is the business of controlling how fine that mixture is.

Critical Points and Thermal Arrests

Heat a piece of plain carbon steel steadily and plot its temperature against time. The line does not rise smoothly. At certain temperatures it pauses, or at least slows markedly, even though the furnace is still supplying heat at the same rate. These pauses are called **thermal arrests** or **critical points**, and each marks a change of crystal structure. The heat going in is being consumed by the transformation itself — the latent heat of the change — instead of raising the temperature.

- **Lower critical point (LCP), also called A1.** The first arrest, at about **723 °C**. Here pearlite begins to transform into austenite. It sits at 723 °C for *every* plain carbon steel regardless of carbon content, which makes it a useful fixed landmark.
- **Upper critical point (UCP), also called A3.** The second arrest, reached after a further temperature rise. Here the last of the remaining ferrite finishes dissolving and the steel becomes **fully austenitic**. Unlike the

LCP it moves with carbon content: 910 °C in pure iron, falling steadily as carbon is added until it meets the lower critical line at 723 °C.

- **The magnetic change point (A2), about 770 °C.** Iron loses its ferromagnetism here without any change of crystal structure. It matters practically because a steel part heated above this temperature cannot be checked or inspected magnetically until it has cooled.

On a low-carbon steel, the interval between the two critical points is substantial — the temperature climbs a further 60 to 160 °C or so above the first arrest before the second one is reached, the gap being widest at the leanest carbon contents and closing steadily as carbon is added. Up to the pearlitic composition it can never be larger than the 187 °C between 910 °C and 723 °C, and it approaches that figure only in the limit of no carbon at all, where there is no pearlite to produce a first arrest. On a steel at the pearlitic composition the two arrests coincide and there is only one.

⚠ **Get the two critical points the right way round.** The *lower* critical point is where transformation to austenite *begins*; the *upper* critical point is where it is *complete*. It follows that fully austenitic steel exists only **above the upper critical point**. Heating to somewhere between the two leaves a partly transformed mixture, and a part quenched from that state hardens unevenly. Almost every hardening and normalising specification therefore quotes a soak temperature safely *above* the upper critical point for the grade concerned.

Reading the Iron-Carbon Diagram

Plot temperature on the vertical axis and carbon content on the horizontal, and the critical points trace out lines. This is the iron-carbon diagram, and three points on it carry names you must be able to separate.

Point	Composition	Temperature	What happens there	Where it matters
Eutectoid point	0.83% carbon (commonly rounded to 0.8%)	about 723 °C	Solid austenite transforms directly into two solids at once — ferrite and cementite — producing 100% pearlite. This is where the upper and lower critical lines meet	Steel. The reference composition for the whole of steel heat treatment
Eutectic point	4.3% carbon	about 1,147 °C	Liquid transforms directly into two solids at once, at the lowest melting temperature of the whole system	Cast iron. Far outside the steel range
Maximum solubility in austenite	about 2.0% carbon	about 1,147 °C	The most carbon austenite will ever dissolve. Beyond it, free carbide is present at every temperature	The dividing line between steel and cast iron

Eutectoid and eutectic are different words for different things, and the difference is worth fixing firmly. The tell is what the material is *before* the transformation. A **eutectic** change starts from a *liquid*: liquid becomes two solids. A **eutectoid** change starts from a *solid*: one solid becomes two different solids. The suffix *-oid* means "like" — a eutectoid reaction is eutectic-like but happens entirely in the solid state. In the iron-carbon system the eutectoid sits at 0.83% carbon and 723 °C and belongs to steel; the eutectic sits at 4.3%

carbon and 1,147 °C and belongs to cast iron. A steel never reaches its eutectic composition, so an engineer working on airframes deals with the eutectoid and only the eutectoid.

What the Carbon Content Does to the Room-Temperature Structure

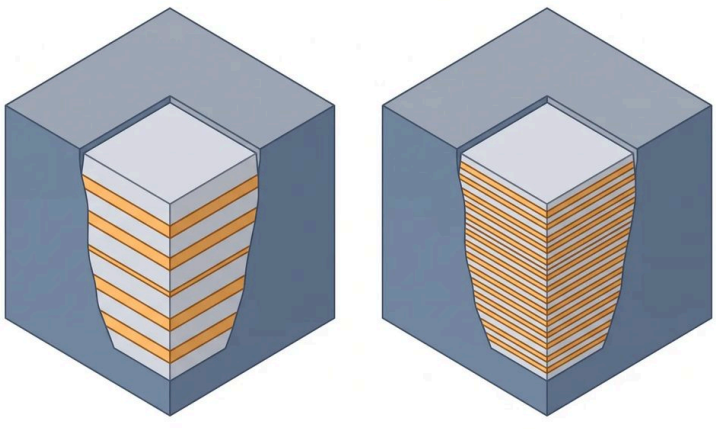
Cool a steel slowly from the austenitic range and the structure it ends up with depends entirely on how much carbon it started with, measured against that eutectoid composition of 0.83%.

Class	Carbon	Slow-cooled structure	Consequence
Hypo-eutectoid (below the eutectoid)	under 0.83%	Grains of free ferrite, with pearlite filling the space between them. The lower the carbon, the more free ferrite	Soft, ductile, tough, weldable. Nearly every structural and fastener steel lives here
Eutectoid	0.83%	100% pearlite and nothing else	The maximum strength obtainable in the slow-cooled condition, with usable toughness
Hyper-eutectoid (above the eutectoid)	0.83% to about 2.0%	Pearlite with a network of free cementite along the grain boundaries	Very hard and wear resistant, but the brittle carbide network makes it weak in shock. Tool and bearing steels

The reason 0.83% carbon is the composition that produces a wholly pearlitic structure is now visible. Below it there is more iron than the carbon can pair with, and the surplus iron is left over as free ferrite. Above it there is more carbon than the iron can absorb into the pearlite, and the surplus appears as free cementite. Only at the eutectoid composition does everything pair up.

Why Fast Cooling Produces Something Different

Every transformation described so far assumes cooling slowly enough for atoms to move. Carbon atoms escaping from austenite have to *diffuse* through the lattice to reach the places where cementite is forming, and diffusion takes time. Take that time away and the transformation cannot happen as written.



Slow cooling makes coarse pearlite plates; faster cooling makes finer, closer and harder ones.

- **Cool slowly (furnace).** The carbon has all the time it needs. It forms coarse pearlite with thick, widely spaced plates. The structure is soft and very ductile.
- **Cool moderately (still air).** The carbon has less time, so it travels shorter distances before it is trapped. The pearlite plates are finer and more closely spaced. Finer pearlite is **harder and stronger** than coarse pearlite of the same composition, because the closely spaced hard plates obstruct deformation more effectively.
- **Cool fast enough (quench).** Once the cooling is faster than a certain rate — the **critical cooling rate** for that steel — the carbon cannot diffuse at all. The face-centred cubic austenite still has to become body-centred cubic because that is what iron does at low temperature, but it must do so with all its carbon still inside it. It shears over into a distorted, stretched body-centred cell called **body-centred tetragonal**, and that structure is **martensite**.

Martensite is hard for a specific and satisfying reason. The trapped carbon atoms wedge the lattice open, and a lattice under that much internal strain resists any further slipping. Hardness is resistance to permanent deformation, and permanent deformation in a metal happens by planes of atoms sliding over one another, so a lattice that cannot slip is a hard one. The same strain is why martensite is brittle and why it is left full of internal stress: nothing about the structure has relaxed.

Two consequences follow directly, and both are examinable:

- **The hardness of martensite depends on the carbon content, not on the alloying.** More carbon means more lattice distortion means more hardness. Hardness climbs steeply with carbon up to roughly 0.6% and then flattens off, so past that point extra carbon buys wear resistance and brittleness rather than more hardness.
- **A steel with too little carbon cannot be usefully quench hardened.** A mild steel of 0.15% carbon forms a martensite so lightly strained that it is barely harder than the pearlite it replaced. This is the reason a low-carbon component that needs a hard surface has to have carbon *put into* that surface first — the process covered later under case hardening.

✈ **Three facts, one sentence each, that answer most iron-carbon questions:** steel becomes austenite on heating and that austenite becomes pearlite again on *slow* cooling. The hardness available in a plain carbon steel comes from the hard constituents its carbon allows to form — iron carbide (cementite) when the carbon has time to diffuse, and a strained martensite lattice when it does not — so it tracks the carbon content. And retained austenite left over after quenching is *softer* than martensite, so a steel is never harder because it contains more austenite — that is a common distractor and it is backwards.

Transformation Is a Race Against Time

It helps to picture the choice between pearlite, bainite and martensite as a race. On cooling, the austenite is unstable and will transform into *something*; which product it gets depends on how much time it is given at each temperature on the way down.

- Held just below the lower critical point, where diffusion is easy but the driving force is small, austenite forms **coarse pearlite** slowly.

- Held a couple of hundred degrees lower, where the driving force is large and diffusion still just possible, it forms **fine pearlite** quickly. This is the fastest region of all, and it is why a quench has to get the part *past* this temperature band without lingering.
- Lower again, where diffusion is sluggish, it forms **bainite** — a fine mixture that is hard and unusually tough.
- Below a temperature called the **martensite start** temperature, no diffusion is possible and the remaining austenite shears over into **martensite** instantly, without any time dependence at all. Transformation continues only as the temperature keeps falling and stops at the **martensite finish** temperature.

Two practical results come out of this picture. First, alloying elements slow diffusion down, which delays the formation of pearlite and lets a slower, gentler quench still reach martensite — the property called **hardenability**, dealt with fully in the heat treatment section. Second, the martensite finish temperature is not necessarily above room temperature, so a quench does not necessarily transform everything — the untransformed remainder is called **retained austenite**, and what it does to a component is dealt with in the heat treatment section along with the treatment used to remove it.

Properties of Ferrous Materials

Property	Description
Tensile strength	Resistance to being pulled apart — steel has very high tensile strength, typically 400–2000 MPa depending on alloy and treatment
Hardness	Resistance to surface indentation or scratching — can be greatly increased by heat treatment and alloying
Ductility	Ability to be drawn into wire or deformed without fracturing — low-carbon steels are very ductile, high-carbon steels less so
Malleability	Ability to be hammered or rolled into shape without cracking
Toughness	Ability to absorb energy and deform plastically before fracturing — a combination of strength and ductility
Elasticity	Ability to return to original shape after load is removed — steel has a well-defined elastic limit
Fatigue strength	Resistance to failure under repeated cyclic loading — critical in aircraft structures
Corrosion resistance	Plain carbon steels rust readily; stainless steels and surface treatments improve corrosion resistance
Density	Approximately 7,850 kg/m³ — about 2.8 times heavier than aluminium
Stiffness (Young's modulus)	Resistance to elastic deflection under load, quite separate from strength — close to 200 GPa for practically every steel, and essentially unaffected by alloying or heat treatment
Yield strength	The stress at which permanent deformation begins; the practical design limit, since a part that has yielded is out of tolerance even though it has not broken
Brittleness	The opposite of toughness — fracture with little or no plastic deformation and little energy absorbed. Rises with carbon content, with hardness, and as temperature falls

Property	Description
Resilience	The energy a material can absorb and give back *within* the elastic range — the property a spring is chosen for, and not the same thing as toughness
Hardenability	The *depth* to which a steel will harden on quenching, as distinct from the hardness value reached. Set mainly by the alloying elements, not by carbon
Wear resistance	Resistance to material loss by rubbing, rolling or abrasive contact; broadly tracks surface hardness, which is why bearing and gear surfaces are case hardened
Creep resistance	Alone among the properties here this one is *time*-dependent: a hot, steadily loaded part can be within every limit on the day it is fitted and outside them years later. It governs exhaust, turbine and hot-zone fastener life
Machinability	The ease with which the material can be cut, and the surface finish and tool life obtained. Best in the medium-hardness range; both very soft and very hard steels machine poorly
Weldability	The ability to be joined by fusion without cracking or forming a brittle heat-affected zone. Falls sharply as carbon and alloy content rise

Stress and Strain

Most of the entries in the table above are mechanical properties, and a mechanical property is defined in terms of stress and strain, so the two definitions have to be exact.

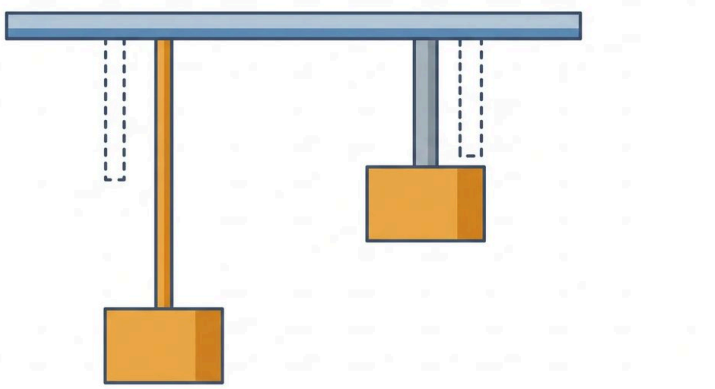
- **Stress** is force divided by the cross-sectional area carrying it. In SI units it is newtons per square millimetre, which is the same as the megapascal. Stress is what the *material* feels; two rods carrying the same load are at different stresses if their diameters differ.
- **Strain** is the change in length divided by the original length. It is a ratio, so it has no units, and it is often quoted as a percentage.

Three kinds of direct stress appear on an aircraft — tensile (pulling apart), compressive (pushing together) and shear (sliding one plane over the next) — together with the combinations engineers call bending and torsion. Steel behaves almost identically in tension and compression up to the yield point, which is not true of every material: cast iron, for example, is several times stronger in compression than in tension, and that asymmetry is why it is used for housings but never for a tie rod.

Elasticity and Stiffness Are Not the Same Property

This distinction causes more confusion than any other in the property list, and getting it clear pays for itself repeatedly.

Elasticity is the *ability* to return to the original shape when the load is removed. **Stiffness**, measured by **Young's modulus**, is *how far the material moves* while it is loaded. A rubber band is highly elastic and extremely un-stiff. A steel bar is elastic over a much smaller range but is enormously stiffer within it.



Under the same load, a rubber band stretches a long way; a steel bar, far stiffer, barely moves.

Within the elastic range, stress and strain are proportional. That proportionality is **Hooke's law**, and the constant of proportionality is Young's modulus:

Hooke's law and Young's modulus:

- stress = Young's modulus × strain
- Young's modulus E = stress / strain
- extension = (load × original length) / (cross-sectional area × E)
- For all steels, E is close to **200 GPa** — about 200,000 N/mm²

Now the point that matters, and it is one of the most useful single facts in this module: **Young's modulus for steel is essentially the same whatever the alloy and whatever the heat treatment.** Mild steel, 4130, 4340 hardened to its maximum, and annealed stainless all sit within a few per cent of 200 GPa. Austenitic stainless grades sit slightly lower, but nowhere near enough to change the conclusion. Heat treatment changes how much load the steel can take before it deforms permanently. It does not change how much it deflects while still elastic.

Worked example — why a stronger steel does not deflect less:

- A steel tie rod is 10 mm in diameter and 1.0 m long, and carries 20 kN in tension.
- Cross-sectional area = *pi*/4 × 10² = 78.54 mm²
- Stress = 20,000 N / 78.54 mm² = **254.6 N/mm²** (254.6 MPa)
- Strain = stress / E = 254.6 / 200,000 = **0.001273**
- Extension = strain × original length = 0.001273 × 1,000 mm = **1.27 mm**
- Now replace the rod with one of exactly the same size in a steel heat treated to *twice* the tensile strength. The stress is unchanged at 254.6 MPa, E is unchanged at 200 GPa, so the extension is still **1.27 mm**. Nothing has been gained.
- To halve the extension, the *area* must be doubled: 157.1 mm², which is a diameter of **14.1 mm**. Geometry is the only lever available.

The maintenance consequence is direct. If a control run feels spongy, or a bracket flexes more than the drawing allows, fitting a part in a stronger steel will not help. Only a larger section, a shorter unsupported length, or an extra support will. Conversely, if a part is *yielding* — taking a permanent set — then strength is exactly what is short, and heat treatment or a stronger grade is the right answer.

Hardness, and What It Really Tells You

Hardness is resistance to being permanently indented. What makes it so useful is that it is measured on the finished part, in minutes, without destroying anything, and for steels it correlates closely with tensile strength. Both properties depend on the same underlying thing: how strongly the crystal structure resists planes of atoms sliding over one another.

Hardness to tensile strength, for steels:

- UTS in MPa is approximately **3.45 × the Brinell hardness number**
- So 200 HB corresponds to roughly 690 MPa, and 400 HB to roughly 1,380 MPa
- The relationship holds for steels roughly between 125 and 500 HB. It does **not** hold for cold-worked austenitic stainless, for cast iron, or for any non-ferrous metal

This correlation is why a hardness check is the standard way of confirming that a heat treatment has been carried out correctly. It is also why an unexpectedly *low* hardness reading on a highly stressed steel part is treated so seriously: it means the part is not at the strength the design assumed.

Ductility and Malleability

Both describe the ability to deform permanently without cracking, but they describe it under opposite kinds of load, and the exam distinguishes them.

- **Ductility** is deformation under *tension* — the property that lets a metal be drawn out into wire. It is measured by percentage elongation and by reduction of area in a tensile test.
- **Malleability** is deformation under *compression* — the property that lets a metal be hammered, rolled or pressed into shape without cracking.

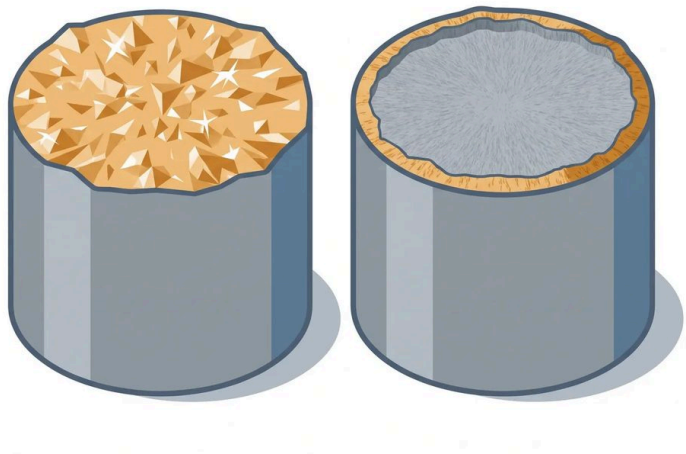
The two usually go together in steels but they are not identical: lead is highly malleable and only moderately ductile. Both fall as carbon content rises and as hardness rises, which is why annealing — the treatment that produces the softest condition — is the one that makes the metal easiest to work.

Toughness, Brittleness and the Trade-Off That Governs Steel Selection

Toughness is the ability to absorb **energy** before fracturing. That is a different question from "how much load can it take", and the difference is easiest to see by noticing that a fully hardened, untempered steel is very strong and very hard yet has almost no toughness at all: it takes a large load, but a sharp knock shatters it. Energy absorbed is load multiplied by the distance moved, so a material that fractures without deforming absorbs almost none, however high the load was.

Brittleness is simply the absence of toughness. A brittle fracture happens suddenly, at right angles to the load, with a bright crystalline appearance and no measurable stretching. A ductile fracture is preceded by

visible necking and leaves a dull, fibrous, cup-and-cone surface.



A brittle fracture face is bright and crystalline; a ductile fracture face is dull and fibrous.

The governing rule for steel selection follows: **strength and toughness generally trade against one another.** Every treatment that raises hardness and tensile strength lowers the energy the steel will absorb before it breaks, and vice versa. The engineer's job is not to maximise either but to select the point on that curve the application needs. There is one important exception to the trade-off, covered in the heat treatment section: refining the grain size raises strength and toughness together.

⚠ **Strength is not toughness, and the strongest steel is rarely the right steel.** Landing gear steels are deliberately *not* heat treated to their maximum attainable hardness. A gear leg that shatters on a heavy landing is far worse than one that bends: a bent leg is detectable, contained and repairable, while a brittle fracture releases all its stored energy at once. The same reasoning explains why specifications set a hardness **range** with an upper limit as well as a lower one, and why a part found *harder* than the drawing allows is just as unserviceable as one found too soft.

Resilience — the Property a Spring Needs

Resilience is the energy a material stores and returns while remaining *elastic*. On a stress-strain diagram it is the area under the elastic part of the curve only, whereas toughness is the area under the whole curve to fracture. A spring steel is chosen for high resilience: a high yield strength combined with the ordinary steel value of Young's modulus means it can be deflected a long way, store a lot of energy, and give all of it back. Spring behaviour itself, including rate and the effect of coil geometry, belongs with the springs material later in this module.

Fatigue and Creep in Service

Two of the properties in the table above describe failure under conditions where the load never approaches the material's static strength, and between them they account for the majority of metallic failures on aircraft.

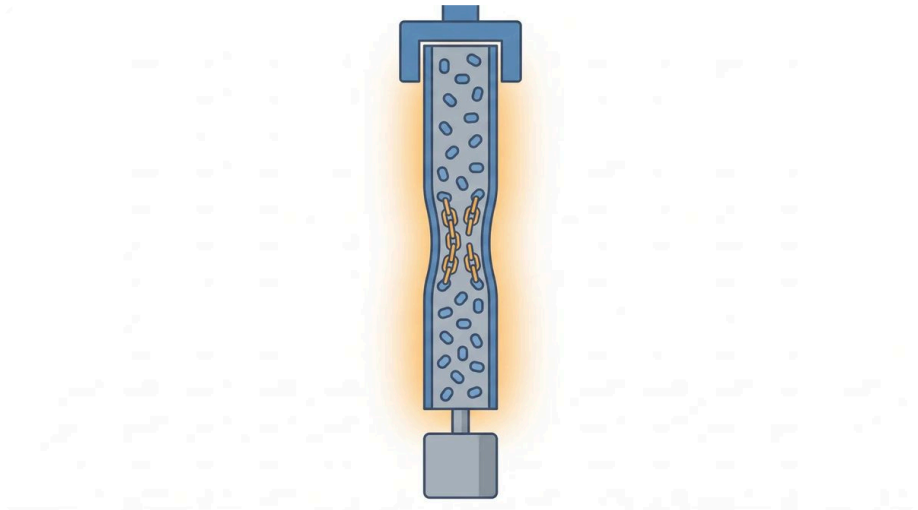
Fatigue is progressive cracking under repeated or alternating load. The defining and slightly alarming feature is that it occurs at stresses well *below* the ultimate tensile strength — often below the yield strength — so a

part can fail after enough cycles without ever having been overloaded in the ordinary sense. It is not a single-overload failure and it is not an impact failure; it is a reduction in load-carrying capacity caused by the cycling itself. This is why aircraft components carry life limits expressed in flight cycles and hours rather than in load.

Creep is slow, continuous, permanent stretching under a *steady* load at elevated temperature. Time and temperature do the work that stress alone could not. It becomes a design consideration in steels above roughly four tenths of the absolute melting temperature, which for steel means the hot end of an engine installation: turbine discs and blades, shafts, exhaust components and the fasteners holding them. Creep proceeds in three distinct stages:

Stage	What is happening	Rate of strain	Engineering response
Primary	Initial rapid extension that then slows as the material work hardens	Decreasing	Accepted; allowed for in the design clearances
Secondary	Steady state — work hardening and thermal recovery are in balance. The longest stage by far	Constant and minimum	This is the design regime. The permitted life is calculated from this rate
Tertiary	Internal voids link up and the section necks down, so the true stress climbs and accelerates the process further	Rapidly increasing	Irreversible and self-accelerating. Rupture follows quickly

⚠ **Tertiary creep is not a monitoring case, it is a removal case.** Once a component has entered the third stage of creep the strain rate is accelerating and nothing can arrest it — the damage is distributed voids inside the metal, not a single crack. Crack-arrest measures such as stop drilling are meaningless against it, and condition monitoring only records the approach of a failure that is already certain. **The component is replaced immediately.** Creep damage cannot be heat treated out, and it cannot be repaired.



In tertiary creep, internal voids link up and the hot, loaded section necks down toward rupture.

How Treatments Move the Properties

The single most useful thing to hold in your head is the *direction* each process moves each property. Each row below states its starting condition, because a claim such as "annealing softens the steel" only means something relative to what it started as.

Change, and from what starting state	Hardness and tensile strength	Ductility	Toughness	Internal (residual) stress	Grain
More carbon (comparing two steels given the same treatment)	Higher	Lower	Lower	No direct effect	No direct effect
Quench hardening, from the normalised condition	Much higher	Much lower	Much lower	Much higher	Unchanged in size; structure becomes martensitic
Tempering, from the as-quenched condition	Slightly lower	Higher	Much higher	Much lower	Unchanged
Full annealing, from any condition	Lowest obtainable	Highest obtainable	Moderate — limited by the coarse structure produced	Lowest obtainable	Coarse
Normalising, from a coarse as-forged or as-welded condition	Moderately higher than annealed	Good	Higher than annealed	Much lower than as-forged	Refined and uniform
Cold working, from the annealed condition	Higher	Lower	Lower	Higher	Distorted and elongated in the working direction
Grain refinement at constant composition	Higher	Little change	Higher	No direct effect	Finer — the one change that improves strength and toughness together
Case hardening, from a low-carbon base	Much higher at the surface only	Lower at the surface; core unchanged	Core unchanged — that is the point of the process	Surface left in useful compression	Unchanged in the core

Physical Properties and What They Cost You in Practice

Physical property	Typical value for steel	Practical consequence
Melting point	about 1,400 to 1,540 °C depending on carbon content	Comfortably above any airframe temperature, but relevant to welding, brazing and fire resistance requirements
Thermal conductivity	about 50 W/m·K for carbon steel; about 15 for austenitic stainless	Heat generated at a cutting edge stays at the cutting edge in stainless. This is a direct cause of the drilling difficulties covered later
Coefficient of linear expansion	about 12 microns/m/°C for carbon steel; about 17 for austenitic stainless	A steel part fitted into an aluminium structure sees its clearance change with temperature, because the aluminium expands about twice as much
Specific heat capacity	about 480 J/kg·K — roughly half that of aluminium	Together with the conductivity above, it governs how long a part takes to heat through in a furnace and how steeply the temperature varies through a thick section during a quench
Electrical resistivity	roughly ten times that of copper for carbon steel, and several times higher again for austenitic stainless	Steel is a poor conductor, so a steel structure is a poor bonding path and steel is never used as a primary electrical conductor
Magnetic permeability	High for ferrite and martensite; effectively that of air for austenite	Enables magnetic particle inspection and the magnet identification check; forces ferrous items to be kept out of compass-safe zones and demands demagnetisation after magnetic inspection

The expansion figures deserve one extra sentence, because they explain a family of problems that look unrelated. Steel expands roughly **half** as much as aluminium alloy for the same temperature rise. Where the two are bolted together over a long run — a steel fitting on an aluminium spar, or a steel cable strung along an aluminium fuselage — the aluminium moves further, and the joint or the run has to accommodate the difference. Corrosion resistance is the remaining property whose consequences reach furthest, and it is dealt with in full with the corrosion material later in this module; here it is enough to say that plain carbon and low-alloy steels have essentially none of it and depend entirely on a protective scheme, while the stainless grades carry their own.

Types of Steel

Plain Carbon Steel

Plain carbon steel contains iron and carbon with only small amounts of other elements (manganese, silicon, sulphur, phosphorus). It is classified by carbon content:

Type	Carbon %	Properties	Aircraft Applications
Low carbon (mild steel)	0.05–0.30%	Soft, ductile, easily welded; low strength	General fittings, non-structural brackets, wire, welded tube structures (some light aircraft fuselages)
Medium carbon	0.30–0.50%	Harder, stronger; responds to heat treatment	Bolts, studs, axles, forgings
High carbon	0.50–1.50%	Very hard after heat treatment; brittle; holds sharp edge	Springs, cutting tools, wire rope; not common in primary structures
Dead mild (ultra-low carbon)	below 0.15%	Extremely ductile; folds and deep-draws without cracking; cannot be usefully quench hardened because there is too little carbon to form a hard martensite	Pressed and formed sheet parts, shims, clips, and the base stock for components that will be carburised
Free-cutting (resulphurised)	0.08–0.45%, with sulphur deliberately raised	Machines cleanly and breaks its own chips, but the sulphide stringers act as internal notches and cut ductility and fatigue strength across the grain	Shop tooling and non-structural items only — specifically excluded from highly stressed aircraft parts

✈ **Where the boundaries fall is a convention, not a physical step.** Carbon content varies continuously and the properties change continuously with it, so the lines drawn between "low", "medium" and "high" are book-keeping. Different references place the medium-to-high division at 0.5% or at 0.6%, and both descriptions cover the same continuum. In an examination, read all the options first: they are normally written to a single scheme and only one band will be offered for the class asked about. What genuinely does *not* vary is what the carbon is doing — more carbon means more of the hard constituent available, so more attainable hardness and strength and less ductility, weldability and toughness.

What Each Carbon Range Can and Cannot Do

The classification is only useful if you know what it buys you, so take one grade from each band and follow it through:

- **A 0.20% carbon steel** can be cold bent around a tight radius, welded with no preheat and no post-weld treatment, and formed on a folder. Quench it and it will barely change hardness — there is not enough carbon to distort the lattice. If a part in this material needs a hard surface, carbon has to be diffused into the surface first.
- **A 0.45% carbon steel** responds properly to quenching and tempering, reaching a useful hardness through the section on small parts. It can still be welded, but a preheat and a controlled cooling rate are needed or the heat-affected zone hardens and cracks. This is the band that bolts, studs, axles and shafts come from.
- **A 0.95% carbon steel** hardens to a file-hard condition and holds a cutting edge, but it snaps rather than bends, it cannot practically be welded, and it is used only where hardness is the whole point — springs, cutting tools, and wire that must resist wear.

Notice that the direction of every property is the same as carbon rises: attainable hardness, tensile strength and wear resistance all go **up**; ductility, malleability, toughness, weldability and formability all go **down**. There is no carbon content that is good at everything, which is exactly why alloying elements exist — they let a designer move one property without paying the full price in the others.

⚠ **Never substitute a carbon grade on the basis of strength alone.** Two steels with the same tensile strength can behave completely differently under a sudden load, at low temperature or when welded. A part called out in a low-carbon weldable grade and replaced with a stronger medium-carbon one may crack in the weld or fail in a brittle manner in service. Material substitution is a design change and needs approved data — it is never a maintenance decision.

Alloy Steels

Alloy steels have deliberate additions of other elements to improve specific properties. Common alloying elements in aircraft steels:

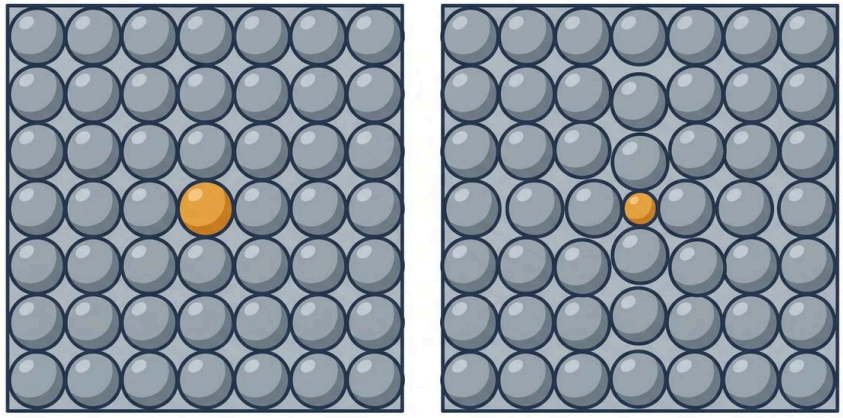
Element	Effect
Chromium (Cr)	Increases hardness, wear resistance, and corrosion resistance; key element in stainless steel (≥10.5% Cr)
Nickel (Ni)	Increases toughness and impact resistance; improves low-temperature properties; corrosion resistance
Molybdenum (Mo)	Increases strength at elevated temperatures; reduces temper brittleness; improves hardenability
Vanadium (V)	Refines grain structure; increases strength, toughness, and wear resistance
Tungsten (W)	Increases hardness at high temperatures; used in high-speed tool steels
Manganese (Mn)	Increases strength and hardness; improves hardenability; counteracts sulphur brittleness
Silicon (Si)	Improves strength and elasticity; used in spring steels
Cobalt (Co)	Raises hot hardness and is the element that lets a tool steel keep its edge at the highest cutting temperatures; also a constituent of maraging steel. Unusually among the alloying elements it does *not* improve hardenability
Boron (B)	In quantities of a few thousandths of one per cent it produces a large increase in hardenability, allowing a leaner and cheaper steel to harden through a given section
Aluminium (Al)	A deoxidiser and grain refiner in ordinary steels; in nitriding grades it is the essential element, because it forms the extremely hard aluminium nitrides that give a nitrided case its hardness
Titanium (Ti) and Niobium (Nb)	Strong carbide formers used to "stabilise" stainless steel: they lock up the carbon so that chromium carbides cannot form at the grain boundaries during welding, which is what preserves corrosion resistance in a weld zone
Sulphur (S)	Normally an unwanted impurity that causes cracking during hot working. Added deliberately only in free-cutting steels to improve machinability, at the cost of ductility and fatigue strength across the grain

Element	Effect
Phosphorus (P)	An impurity that raises strength slightly but embrittles the steel, particularly at low temperature. Aerospace specifications set a tight maximum on it
Copper (Cu)	Gives a modest improvement in atmospheric corrosion resistance and, in the precipitation-hardening stainless grades, forms the fine particles that produce the strengthening
Lead (Pb)	Added only to improve machinability in general engineering steels; not used in aircraft structural grades

How an Alloying Element Actually Works

The table above lists what each element does. Understanding *how* it does it turns a list into something you can reason with, and it explains several results that otherwise have to be memorised.

An added element has to go somewhere in the iron crystal, and there are only two places available:



Same-size atoms replace iron atoms; much smaller atoms squeeze into the gaps and distort the lattice.

- **Substitutional.** The atom is roughly the same size as an iron atom, so it simply takes an iron atom's place in the lattice. Chromium, nickel, manganese, molybdenum, silicon, tungsten, vanadium and cobalt all behave this way. Because the fit is reasonably good, the lattice is only mildly distorted, and useful effects need percentage-level additions.
- **Interstitial.** The atom is much smaller and squeezes into the gaps *between* the iron atoms. Carbon, nitrogen and boron do this. The fit is poor, so the lattice is heavily distorted, and a very small amount produces a large effect. This is the reason a tenth of one per cent of carbon changes a steel more than one per cent of nickel does, and the reason carbon and nitrogen are the elements used for surface hardening.

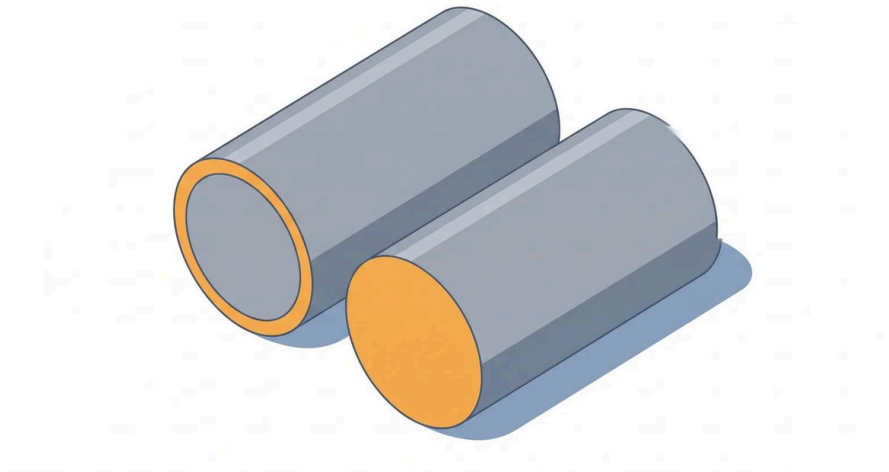
From there, alloying elements do three distinct jobs:

- **They strengthen the solid solution.** Any foreign atom distorts the lattice around it and makes slip harder. This is a modest, general effect.
- **They form carbides.** Chromium, molybdenum, vanadium, tungsten, titanium and niobium bond with carbon more readily than iron does, producing carbides that are harder and far more stable at temperature

than plain iron carbide. Hard carbides raise hardness and wear resistance directly; stable carbides are what let a tool steel keep its edge when it is running hot, and what give the alloy steels their resistance to softening and to creep.

- **They change when and how the steel transforms.** This is the most important effect and the least obvious. Alloying elements slow down the diffusion that pearlite formation depends on. Delay the pearlite and a slower quench still reaches martensite — which is **hardenability**, the depth of hardening. Molybdenum, chromium, manganese, nickel and boron all raise it.

✈ **Hardenability is not hardness, and confusing the two costs marks.** *Hardness* is how hard the steel gets, and it is set almost entirely by the **carbon** content. *Hardenability* is how **deep** the hardening penetrates, and it is set almost entirely by the **alloying elements**. Two steels can contain identical carbon and reach identical surface hardness, yet one is hard only in a thin skin while the other is hard right through a thick section. Worked case: a plain 0.40% carbon steel and 4340 both contain 0.40% carbon and both reach roughly the same maximum hardness, but the plain carbon grade needs a violent water quench and even then hardens only a shallow rim of a large bar, while 4340 hardens through a much heavier section in a gentle oil quench. Since a gentler quench means less distortion and far less risk of quench cracking, high hardenability is worth paying for even when the final hardness is the same.



Same carbon, same surface hardness: plain carbon steel hardens a shallow rim, 4340 hardens right through.

Austenite Stabilisers and Ferrite Stabilisers

One further effect explains most of what you need to know about stainless steel, so it is worth stating on its own. Some elements widen the temperature range over which austenite is stable, and some narrow it.

- **Austenite stabilisers** — nickel, manganese, carbon, nitrogen. Adding these **lowers** the temperature at which austenite gives way to ferrite. Add enough and austenite survives all the way down to room temperature.
- **Ferrite stabilisers** — chromium, molybdenum, silicon, titanium, niobium, aluminium. Adding these **raises** that temperature, and enough of them removes the austenite field altogether so the steel is ferritic at every temperature up to melting.

Now work three real cases through, and the whole stainless family falls out of one mechanism:

- **17% chromium, very low carbon, no nickel.** Chromium is a ferrite stabiliser and there is nothing to oppose it, so the steel stays body-centred cubic ferrite from room temperature upwards. It never becomes austenite on heating, so it can never be quenched to martensite. Result: **magnetic, and not hardenable by heat treatment** — only by cold work. This is the *ferritic* family.
- **18% chromium plus 8% nickel.** The nickel is a strong enough austenite stabiliser to overcome the chromium and hold the face-centred cubic structure down to room temperature. Result: **non-magnetic** (austenite has no ferromagnetism), **not hardenable by heat treatment** (there is no transformation left to exploit — it is already austenite, so quenching it changes nothing), very ductile, very formable, and strengthened only by cold work. This is the *austenitic* family, and it is the 18-8 composition.
- **12 to 17% chromium with the carbon raised to 0.15% or more.** Carbon is an austenite stabiliser, and enough of it re-opens the austenite field at high temperature. The steel can therefore be heated to austenite and quenched to martensite exactly like an ordinary carbon steel. Result: **magnetic and hardenable by quench and temper**, with corrosion resistance lower than the austenitic grades because much of the chromium is tied up as carbide. This is the *martensitic* family.

Two more consequences of the austenitic structure are worth carrying with you. Austenitic stainless expands about half as much again as carbon steel for the same temperature rise, and conducts heat only about a third as well, so a welded austenitic assembly distorts far more than an equivalent carbon-steel one. And austenite work hardens very rapidly, which is why machining it demands a particular technique — covered later under working and handling.

Important Aircraft Steel Alloys

Designation	Composition	Properties	Typical Aircraft Use
SAE 4130 (chromoly)	Cr-Mo steel (0.30% C, 1% Cr, 0.2% Mo)	Excellent strength-to-weight ratio; good weldability; responds well to heat treatment	Engine mounts, fuselage tubing (light aircraft), landing gear, structural fittings
SAE 4340	Ni-Cr-Mo steel (0.40% C, 0.8% Cr, 1.8% Ni, 0.25% Mo)	Very high strength (up to 1800 MPa heat-treated); excellent fatigue strength and toughness	Landing gear components, crankshafts, connecting rods, high-stress structural fittings
Stainless steel (18-8)	18% Cr, 8% Ni (austenitic type 304/321)	Excellent corrosion resistance; non-magnetic; good high-temperature properties	Exhaust systems, firewalls, high-temperature zones, fasteners
Maraging steel	18% Ni, plus Co, Mo, Ti	Ultra-high strength (up to 2400 MPa); tough; good fatigue resistance	Landing gear, high-performance structural parts
SAE 4140	Cr-Mo steel (0.40% C, 1% Cr, 0.2% Mo)	The higher-carbon partner to 4130: a greater attainable strength through a heavier section, but noticeably harder to weld	Machined fittings, shafts, axles, pins, torque links, and heavy-duty tooling

Designation	Composition	Properties	Typical Aircraft Use
300M (modified 4340)	Ni-Cr-Mo with about 1.6% silicon and vanadium added (0.42% C)	Ultra-high strength, around 1,900–2,000 MPa, with better toughness at that strength than 4340; very notch-sensitive and demands an intact protective finish	Main landing gear cylinders, pistons and axles on large transport aircraft
SAE 8620	Ni-Cr-Mo case-hardening steel (0.20% C)	Low-carbon core stays tough while the carburised surface hardens; good hardenability for its cost	Gears, splines, cam followers, bushes and other wear surfaces
SAE 9310	Ni-Cr-Mo carburising steel with higher nickel (0.10% C)	Premium aerospace gear steel; exceptional core toughness under a hard case, and normally supplied vacuum melted	Engine and gearbox gears, accessory drive gears, splined shafts
SAE 52100	High-carbon chromium bearing steel (about 1% C, 1.45% Cr)	Through-hardens to a very high, uniform hardness; made to extreme cleanliness standards because a single inclusion becomes a spall	Ball and roller bearing races and rolling elements
Nitriding steel (e.g. En41B, the 1.5% Cr-Al-Mo type)	About 0.4% C with roughly 1.5% Cr, 1% Al and 0.25% Mo	The aluminium and chromium form extremely hard nitrides, giving a surface hardness no carburised case can match, with almost no distortion	Crankshaft and camshaft journals, cylinder liners, gears and spindles needing dimensional stability
17-4 PH stainless	About 17% Cr, 4% Ni, 4% Cu with niobium	Precipitation-hardening stainless: solution treated soft, then aged to high strength with very little dimensional change and good corrosion resistance	Highly loaded fittings, actuator components, shafts and valve parts in wet or hot zones
Stainless steel (321)	18-8 austenitic stabilised with titanium	Titanium locks up the carbon so chromium carbides cannot form during welding, preserving corrosion resistance in and beside the weld	Welded exhaust systems, hot-air and bleed ducting, firewall assemblies
Martensitic stainless (410 / 431)	12–17% Cr with the carbon raised enough to allow transformation	Magnetic and hardenable by quench and temper; corrosion resistance lower than the austenitic grades because chromium is tied up as carbide	Compressor blades, fasteners, valve parts, shafts, knife-edge seals and cutting edges
A286 (iron-base superalloy)	About 25% Ni, 15% Cr and 2% Ti, balance iron	Age hardened like the PH grades but retains useful strength to around 700 °C; non-magnetic and corrosion resistant	High-temperature bolts and studs, turbine and hot-section hardware, exhaust clamps

Designation	Composition	Properties	Typical Aircraft Use
Spring steel (silicon-manganese or chrome-vanadium)	About 0.6% C with 2% Si, or about 0.5% C with Cr and V	Hardened and tempered to a very high yield strength, giving a large elastic range and the ability to return fully to shape after deflection	Valve springs, control-system return springs, retaining clips, circlips, tab washers
Low-carbon aircraft tube (e.g. SAE 1025)	Plain carbon steel, about 0.25% C	Soft, ductile and very easily welded; low strength, so sections are larger than the equivalent in 4130	Non-structural tube, secondary structure and brackets on light aircraft

The Stainless Steel Families

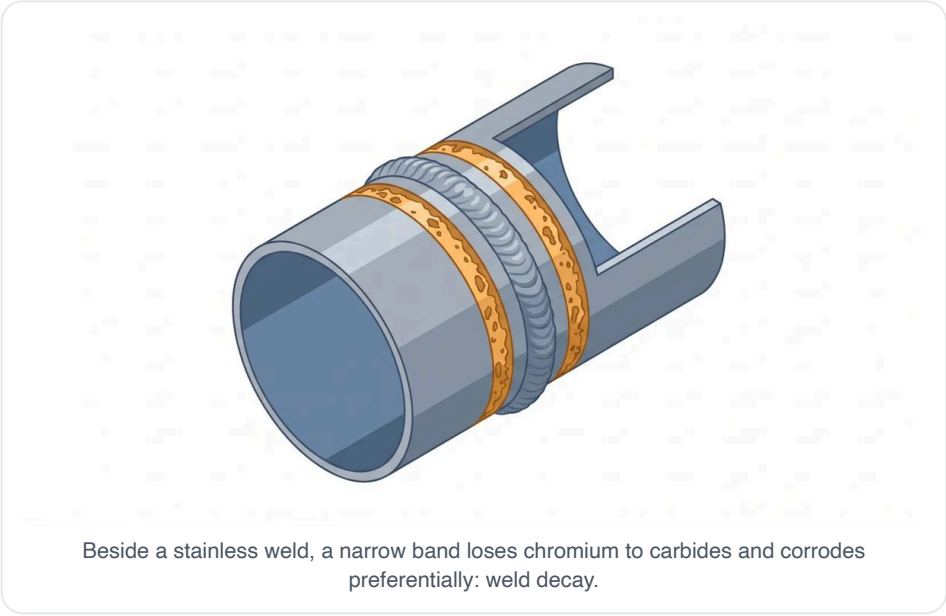
"Stainless" is not one material. The defining requirement is a minimum of about **10.5% chromium**, which reacts with oxygen to form an invisible, tightly adherent chromium-oxide film a few atoms thick. That film is *passive* — it blocks further attack — and it is *self-repairing*, re-forming within seconds if it is scratched, provided oxygen is available. Nickel is frequently added as well, but nickel is not what makes a steel stainless; chromium is.

Beyond that minimum the family splits according to the crystal structure the composition produces, and structure decides everything else:

Family	Structure	Magnetic?	Strengthened by	Corrosion resistance	Typical aircraft use
Austenitic (300 series — 304, 316, 321, 347)	Face-centred cubic austenite, held down to room temperature by nickel	No	Cold work only — heat treatment cannot harden it	Best of the families; 316 adds molybdenum for pitting resistance	Firewalls, exhaust systems, hot-air ducting, clamps, safety wire, some fasteners
Ferritic (400 series — 430)	Body-centred cubic ferrite at all temperatures; 11–30% Cr with low carbon	Yes	Cold work only — it never becomes austenite, so it cannot be quench hardened	Moderate	Trim, non-structural panelling, low-stress hot items
Martensitic (410, 416, 420, 431, 440C)	Quenched to martensite from austenite, made possible by the higher carbon	Yes	Quench and temper, exactly like a carbon steel	Lowest of the families — chromium is tied up as carbide	Compressor blades, fasteners, valve parts, bearings, cutting edges
Precipitation hardening (17-4 PH, 15-5 PH, 17-7 PH)	Martensitic or semi-austenitic, strengthened by fine particles rather than by carbon	Usually yes	Solution treatment followed by ageing	Good — close to the austenitic grades	Highly loaded fittings, actuator rods, shafts, structural parts in wet zones

Family	Structure	Magnetic?	Strengthened by	Corrosion resistance	Typical aircraft use
Duplex	Roughly half austenite and half ferrite by design	Yes	Cold work; the mixed structure gives high strength as supplied	Very good, especially against stress corrosion in chlorides	Rare in airframes; met mainly in marine and process plant

⚠ **Weld decay, and why an exhaust repair uses 321 and not 304.** Hold an ordinary 18-8 austenitic stainless anywhere in the approximate range 425 to 815 °C — which is exactly what a welding torch does to the metal each side of the bead — and chromium combines with carbon to form chromium carbides along the grain boundaries. The chromium in those carbides is no longer available to maintain the passive film, so a narrow band beside the weld is left *locally* short of chromium and corrodes preferentially. This is called **sensitisation**, and the corrosion it produces is **weld decay**. There are three cures and only three: use a low-carbon grade (304L, 316L) so there is little carbon to react; use a *stabilised* grade in which titanium (321) or niobium (347) has already claimed the carbon; or re-solution-treat the whole assembly afterwards, which is impractical on an installed part. Aircraft exhaust and hot-duct work is welded in 321 for precisely this reason.



Beside a stainless weld, a narrow band loses chromium to carbides and corrodes preferentially: weld decay.

Tool Steels and High-Speed Steel

Tool steels are not airframe materials, but an engineer uses them every day in drills, reamers, taps, punches and dies, and the exam treats them as part of the ferrous map. They are high-carbon steels, usually with heavy alloy additions, heat treated to very high hardness and correspondingly high wear resistance. The price is brittleness: a tool steel is chosen on the understanding that it will chip rather than bend.

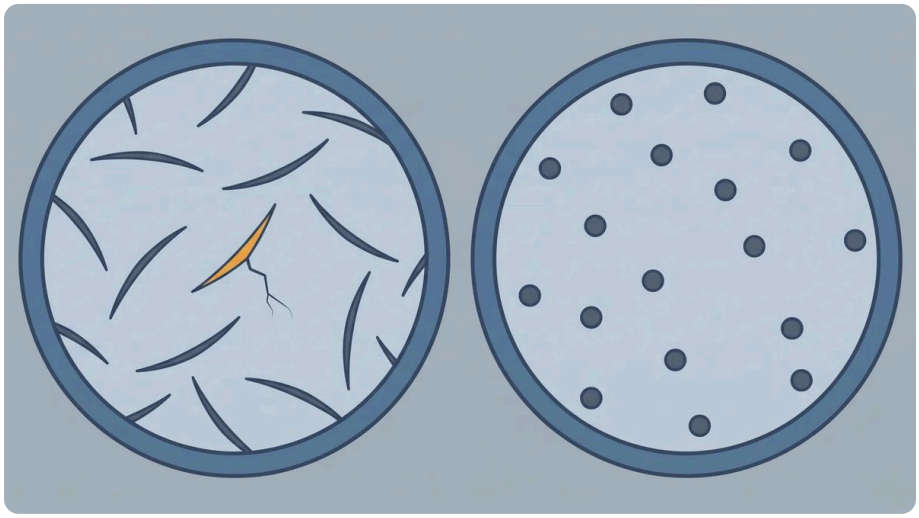
High-speed steel deserves separate treatment because of one property. An ordinary carbon-steel cutting tool loses its hardness as soon as friction heats the edge much above the temperature at which it was tempered — the edge simply tempers itself further and goes soft. High-speed steel does not. It keeps its hardness when the edge is running at a dull red heat, a property called **red hardness** or **hot hardness**, and that is what

allows it to cut at high speed. The element principally responsible is **tungsten**, which forms carbides so stable that they do not dissolve or coarsen at cutting temperatures. The classic composition is quoted as 18-4-1 — about 18% tungsten, 4% chromium and 1% vanadium — and molybdenum can substitute for much of the tungsten in the cheaper grades.

Where the very hardest and hottest cutting is required, **cobalt** is added on top. Cobalt raises hot hardness further, and cobalt-bearing tool and high-speed steels are among the hardest ferrous materials in normal use — a Brinell hardness in the region of 600 to 700 is typical of them, against roughly 100 to 175 for a soft annealed low-carbon steel. Note that this is one of the few practical uses of the Brinell test at the top of its range; a hardened tool steel is normally checked on the Rockwell C scale instead.

Cast Iron

Cast iron carries so much carbon — 2.0% to about 4.5% — that the iron cannot dissolve it all under any conditions. The surplus separates out, and the *shape* it takes decides the properties completely.



Grey iron's graphite flakes act as internal cracks; ductile iron's spheres concentrate stress far less.

Type	Form the carbon takes	Character
Grey cast iron	Graphite flakes throughout the structure	The commonest type. The flakes behave as thousands of internal cracks, so tensile strength is low and ductility is essentially zero, but compressive strength is three to four times the tensile figure. Excellent vibration damping and outstanding machinability
White cast iron	Carbon locked up as cementite, produced by rapid cooling	Extremely hard, extremely brittle, and effectively unmachinable. Used as a wear surface or as the starting stock for malleable iron
Malleable iron	Graphite in irregular rosettes, produced by a long anneal of white iron	Some genuine ductility and shock resistance, at the cost of a lengthy heat treatment

Type	Form the carbon takes	Character
Spheroidal graphite (ductile or nodular) iron	Graphite as discrete spheres, produced by treating the melt with magnesium or cerium	A sphere concentrates stress far less than a flake, so this iron is genuinely tough and ductile while keeping castability. The best of the family mechanically

The reason cast iron is described simply as **heavy and brittle** in an examination is that the question is about grey iron, and grey iron is neither tough nor malleable. It fractures without warning under tension or shock. It has essentially no place in airframe structure; where it appears at all it is in piston-engine ancillaries, valve guides, some accessory housings, and ground support equipment, always in compression and never carrying a shock load.

Wrought Iron

Wrought iron is almost pure iron containing fine threads of slag left over from its manufacture. Those slag stringers give it a fibrous grain that makes it tough, easy to forge-weld and notably resistant to corrosion. It was the structural metal of early aviation and it is obsolete: any airframe carrying it is a historic aircraft, and it has been entirely displaced by mild steel, which is stronger, more consistent and far cheaper to make.

Maraging Steel — Strength Without Carbon

Maraging steel is worth understanding in its own right because it strengthens by a route that does not depend on carbon, and because that route is the same one the aluminium alloys use.

The name is a contraction of *martensitic* and *ageing*. The carbon content is deliberately kept extremely low — below about 0.03% — so the martensite that forms when the steel is cooled from about 820 °C is not the hard, brittle, carbon-distorted martensite described earlier. It is a **soft, tough iron-nickel martensite**, and the steel in that condition can be machined, drilled and formed conventionally. The strength comes afterwards, from two heat-treatment steps that are defined the same way for every alloy that uses them:

- **Solution treatment** — heating to a temperature at which the strengthening constituents dissolve fully into solid solution, then cooling fast enough to hold them there in supersaturated solution. The alloy is comparatively soft and workable immediately afterwards.
- **Precipitation hardening** (also called **age hardening**) — holding the supersaturated alloy at room temperature (*natural* ageing) or at a moderately elevated temperature (*artificial* ageing) so that the dissolved constituents come back out as extremely fine, closely spaced particles distributed through the grains. Those particles obstruct the movement of dislocations, so hardness and strength rise while ductility falls somewhat. Held too long or too hot, the particles coarsen and spread apart, and the strength falls again — that is **over-ageing**.

In maraging steel the ageing is carried out at around 480 °C for a few hours, and the precipitates are intermetallic compounds of nickel with molybdenum and titanium. The practical advantages are considerable: an ultra-high tensile strength is reached with better toughness than a carbon-hardened steel of comparable strength, and because the ageing involves no quench there is almost no distortion, so parts can be finish machined *before* the strengthening treatment. The disadvantages are cost and the fact that maraging steels

have poor corrosion resistance and depend entirely on a protective scheme. These two treatments are the principal strengthening route for the light alloys and are covered fully with the non-ferrous materials.

Bearing Steels

Rolling-contact bearings make demands nothing else does: the contact patch between a ball and a race is tiny, so the local stress is enormous, and it is applied and released millions of times. The material has to be very hard where the rolling contact takes place, dimensionally stable, and above all *clean*, because a single non-metallic inclusion just beneath the surface becomes the origin of a spall, which is why aircraft bearing steels are made to premium melting standards. Two families of steel meet those demands, and they meet them in opposite ways. The commoner of the two is the *through-hardening* family: high-carbon steels — around 1% — with a chromium addition, hardened uniformly right through the rolling element rather than only at the surface. The standard grade in that family is **SAE 52100** and its equivalents — En31 in the older British system, 100Cr6 in the German — all of them roughly 1% carbon with about 1.5% chromium. The second family goes the other way: a low-carbon steel is *carburised*, so that a hard case is produced over a core that stays tough. That is also where nickel belongs in a bearing steel. It is not an addition to the through-hardening grades, in which it is present only as a residual; it is a deliberate addition in the carburising bearing steels such as SAE 4320 and 9310, whose carbon is around 0.1 to 0.2%, and its job there is to keep the core tough under the hardened case. Bearing types, loads and construction are covered with the bearings material later in this module.

Identification of Ferrous Materials

Aircraft steels are identified by several methods:

- **Designation systems:** SAE/AISI (USA), BS (British), DIN (German), or specification numbers (e.g. AMS, MIL-S)
- **Colour coding:** Bars and sheets may be colour-coded on the end or edge per specification standards
- **Spark test:** Grinding the steel on a wheel produces characteristic spark patterns — carbon content affects spark length, branching, and colour. High carbon produces more branching and bursting sparks
- **Magnetic test:** Most ferrous materials are magnetic (attracted to a magnet). Exception: austenitic stainless steels (18-8 type) are non-magnetic
- **Markings:** Part numbers stamped or etched on components; material certificates and traceability documentation

The Order in Which Evidence Counts

The five methods listed above are not equal in weight, and treating them as though they were is the mistake that gets an engineer into trouble. They divide sharply into two classes:

- **Documentary evidence** — the release paperwork, the material certificate, the heat or cast number, the part number on the item. This is what actually establishes what a piece of metal *is*, because it links the physical item to a record of its composition, its melting route, its heat treatment and its test results.

- **Physical tests** — spark, magnet, file, hardness, chemical spot, portable analysis. These are *confirmatory*. They can tell you that a piece of metal is not what it was supposed to be, which is genuinely useful, but no combination of them can establish that it *is* the specified material.

The reason is worth stating plainly. Two bars can have identical chemical analysis, identical hardness and identical spark patterns and still be different materials for aviation purposes, because one was vacuum remelted and heat treated to a certificate and the other was not. Composition is only part of a specification. This is why **an unidentified piece of steel is scrap**: it cannot be made airworthy by testing it, because the tests available cannot recover the history that was lost.

Designation Systems in Detail

The SAE/AISI four-digit system is the one an engineer meets most often on aircraft drawings, and it is genuinely readable once the scheme is known. It encodes composition, not the manufacturer and not any physical property:

- The **first digit** gives the major alloying class.
- The **second digit** gives, approximately, the percentage of the principal alloying element in that class.
- The **last two digits** give the carbon content in **hundredths of one per cent**.

First digit	Class	Common series
1	Plain carbon steels	10xx plain; 11xx and 12xx free cutting; 15xx higher manganese
2	Nickel steels	23xx, 25xx
3	Nickel-chromium steels	31xx, 33xx
4	Molybdenum-bearing steels	40xx molybdenum; **41xx chromium-molybdenum**; **43xx nickel-chromium-molybdenum**; 46xx and 48xx nickel-molybdenum
5	Chromium steels	51xx; 52xxx high-carbon high-chromium bearing steels
6	Chromium-vanadium steels	61xx
7	Tungsten-chromium steels	72xx — largely obsolete
8	Nickel-chromium-molybdenum, in lower proportions than the 43xx series	86xx, 87xx, 88xx
9	Silicon-manganese steels and further Ni-Cr-Mo grades	92xx silicon-manganese; 93xx, 94xx, 98xx

Worked decodes — read each one right to left:

- **4130** — *41* is chromium-molybdenum, *30* is 0.30% carbon. A weldable, heat-treatable structural steel.
- **4340** — *43* is nickel-chromium-molybdenum, *40* is 0.40% carbon. More carbon than 4130 and a richer alloy content, so a higher attainable strength through a heavier section.
- **1020** — *10* is plain carbon, *20* is 0.20% carbon. A mild steel.
- **1095** — *10* is plain carbon, *95* is 0.95% carbon. A high-carbon spring and tool grade.
- **6150** — *61* is chromium-vanadium, *50* is 0.50% carbon. A spring steel.
- **8620** — *86* is a leaner nickel-chromium-molybdenum, *20* is 0.20% carbon. A case-hardening gear steel.
- **9260** — *92* is silicon-manganese, *60* is 0.60% carbon. A spring steel.
- **52100** — five digits rather than four, because the carbon is at or above 1.00% and needs three digits to express. The leading *5* is the chromium class and the final *100* gives 1.00% carbon. This is the high-carbon chromium bearing steel.

The stainless steels use a different, three-digit AISI scheme in which the number is a series identifier rather than a code to be decoded:

- **200 series** — austenitic chromium-nickel-manganese, with manganese substituting for some of the nickel.
- **300 series** — austenitic chromium-nickel. 304 is the basic 18-8; 316 adds molybdenum for pitting resistance; 321 and 347 are the titanium- and niobium-stabilised welding grades.
- **400 series** — straight chromium. Some are ferritic (430) and some martensitic (410, 420, 440C); the carbon content decides which.
- A suffix **L** means a reduced carbon maximum for welding; **H** means a raised carbon content for high-temperature strength.

Alongside those you will meet specification numbers rather than grade names, and the distinction matters:

System	Issued by	What it controls
SAE / AISI	SAE International	Chemical composition only
AMS (Aerospace Material Specification)	SAE International	Composition *plus* the product form, the melting practice, the heat-treated condition, the required mechanical properties, the testing and the acceptance criteria. This is what an aerospace drawing calls out
MIL-S / MIL-DTL	US Department of Defense	Comparable coverage to AMS, on military programmes
BS, EN, DIN, W.Nr	British, European and German standards bodies	European equivalents; EN 10027 uses a descriptive form in which the elements and their percentages appear in the name itself
UNS	SAE and ASTM jointly	A single unified number for every metal and alloy, used to cross-reference between the systems above

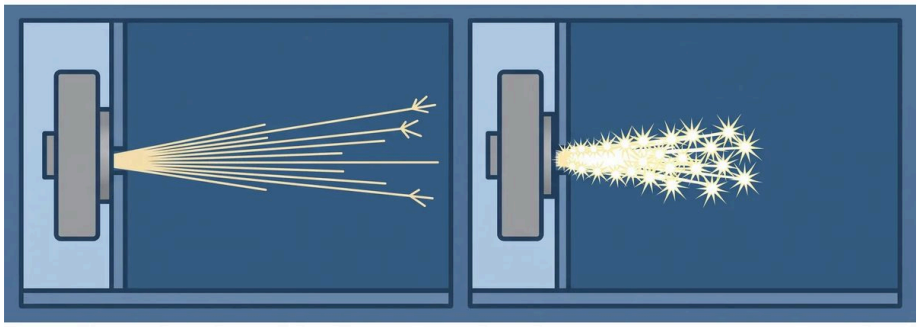
✈ **"4340" and an AMS number are not the same statement.** Saying a part is 4340 states its chemistry. Saying it is made to a particular AMS specification additionally states how the steel was melted, what form it was supplied in, what condition it was heat treated to, what strength it must demonstrate and how that was to be proved. Two bars can both be honestly described as 4340 and be entirely unequal for an aircraft part. When a drawing calls out a specification, the specification is the requirement — the grade name inside it is only part of it.

Colour Coding — Read the Scheme, Never Assume It

Coloured paint or lacquer on the end or the edge of a bar, sheet or tube is a stock-control convenience. The important thing to understand about it is that **there is no single universal colour scheme**. The meaning of a colour is defined by the specification, the standard or the supplier that applied it, and it varies between them. A colour is therefore a pointer to the paperwork, not a substitute for it, and a bar whose colour code cannot be matched to a document is an unidentified bar. Colour marks are also the first thing lost when a bar is cut, cleaned or handled, which is exactly why the heat or cast number is stamped or etched rather than painted.

The Spark Test in Practice

Grinding a steel throws off particles that are heated to incandescence and then burn in air. The carbon in them oxidises explosively, and each explosion splits the streak — so the *amount of branching and bursting* is a direct visual read-out of carbon content, while the colour and length of the stream are influenced by the alloying elements. More carbon means more, and more repeated, bursts.



Low-carbon steel throws long, straight sparks; high-carbon steel throws short, bushy streams full of bursts.

Material	Stream	Colour	Bursts
Wrought iron	Long, straight streaks that widen at the far end into leaf or spear shapes	Straw to white	None
Low-carbon (mild) steel	Long and straight, with a few forks appearing near the end of the stream	Bright straw-yellow	Very few

Material	Stream	Colour	Bursts
Medium-carbon steel	Shorter, with forking along much of its length and small bursts	Bright yellow	Moderate in number
High-carbon steel	Short, dense and bushy, with many repeated star-shaped bursts starting close to the wheel	Bright white-yellow	Numerous and repeating
Grey cast iron	Short stream that stays close to the wheel	Red near the wheel, straw further out	Many small, fine bursts
Austenitic stainless steel	Short and only moderately forked	Straw with a reddish tint	Few
High-speed steel	Short and interrupted, with the streak often ending in a spearhead or bulb	Dark red	None or very few
Nickel and high-nickel alloys	Short, ending in a characteristic wedge or tongue	Orange	None

Three rules make the test worth doing and its results defensible:

- **Always compare against a known sample.** The test is comparative. Grind a piece of certified material of the suspected grade on the same wheel, at the same pressure, immediately before or after the unknown, and compare the two streams directly. An absolute judgement from memory is unreliable.
- **Control the conditions.** Use a clean, sharp wheel and light, steady pressure — heavy pressure lengthens and brightens the stream and masks the differences. View the stream against a dark background in subdued light; in bright sunlight the bursts are simply invisible.
- **Observe the safety precautions.** Eye protection, the correct guard and rest settings, and awareness that grinding sparks are an ignition source. Never spark test a material of unknown composition where magnesium may be present, and never grind a component that is to remain in service — take the sample from scrap stock or from a part already condemned.

The Magnet Test and Its Traps

Most ferrous metals are attracted to a magnet, which makes a pocket magnet the fastest first check available. The response comes from ferrite and martensite, both of which are ferromagnetic; austenite is not. Used carefully the test is genuinely informative, but it has three specific traps.

- **A magnetic response does not prove the metal is ferrous.** Nickel and cobalt are ferromagnetic in their own right, and so are some of their alloys. A magnet cannot by itself distinguish a steel from a nickel alloy.
- **A weak magnetic response does not prove an austenitic part is the wrong material.** Cold working transforms some austenite into martensite, so a heavily formed, drawn or machined 18-8 component — a bent bracket, a swaged fitting, a cut thread — is often weakly magnetic at the worked areas while the parent material beside them is not. The same applies to welds in austenitic stainless, whose filler frequently contains a deliberate small proportion of ferrite to prevent hot cracking.
- **The response disappears with heat.** Iron loses its ferromagnetism at about 770 °C, well below its melting point, so a hot part gives no response at all until it has cooled.

The clean, reliable conclusion the test does support is the one worth memorising: **austenitic stainless steels of the 18-8 type are non-magnetic in the annealed condition, whereas ferritic and martensitic stainless steels are magnetic.** A magnet is the quickest way to separate a 300-series grade from a 400-series one.

File Test, Hardness Comparison and Chemical Spot Tests

A sharp file drawn across an edge tells you a surprising amount. It bites readily into an annealed or normalised steel, cuts with more effort on a hardened and tempered one, and simply skates over a fully hardened surface without removing metal. Because it is quick and needs no equipment it is the classic way of confirming that a case-hardening or hardening operation has actually taken effect. Like the spark test it is comparative and it removes metal, so it is done on a non-critical area or on scrap.

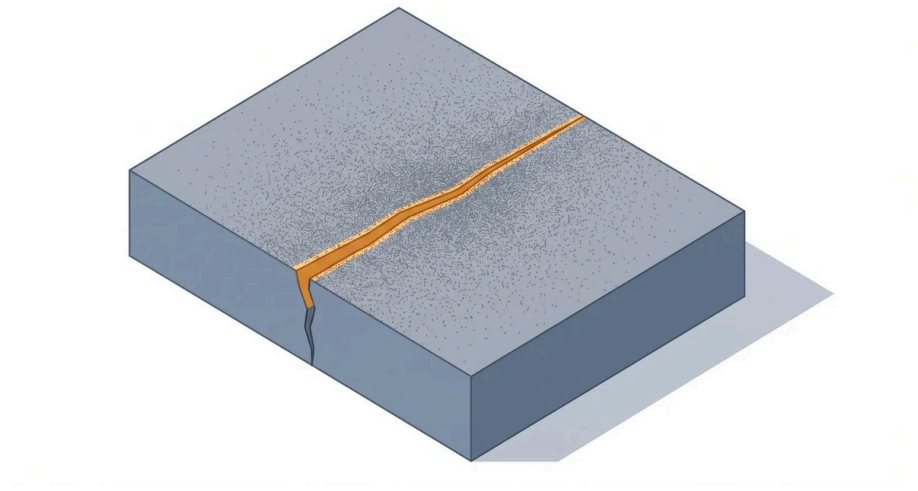
Chemical spot tests apply a drop of reagent and read the colour produced. They can distinguish, for instance, a nickel-bearing stainless from a plain chromium one. They are useful, but they are chemically specific, they mark the surface, and their use on an aircraft part requires approved data.

Instrumental Identification

Method	How it works	What it can do	What it cannot do
Portable X-ray fluorescence	The surface is irradiated and each element re-emits X-rays at its own characteristic energy	Reads chromium, nickel, molybdenum, manganese and the other alloying metals in seconds, without marking the part	**It cannot read carbon** — the element is too light to fluoresce usefully. So it separates 4130 from 4340, but it cannot separate 1020 from 1095, or 410 from 420, because those differ only in carbon
Optical emission spectrometry (arc or spark)	A controlled electrical discharge vaporises a trace of the surface and the light emitted is split into its spectrum	Reads carbon as well as the alloying elements, giving a full analysis	Leaves a small burn mark, so it needs approved data before it is used on a part that is to remain in service
Laboratory combustion analysis	A machined sample is burned in oxygen and the carbon dioxide produced is measured	The reference method for carbon and sulphur	Destructive — it consumes a sample, so it applies to stock rather than to an installed part

The Magnetic Property Is Also an Inspection Tool

The same ferromagnetism that makes the magnet test work is what makes **magnetic particle inspection** possible. The part is magnetised and fine iron particles, dry or suspended in a liquid and often fluorescent, are applied. A crack or inclusion interrupts the magnetic field and forces some of it out of the surface; that leakage field holds the particles and draws a visible line along the defect. The technique finds surface and near-surface discontinuities, and it is fast, sensitive and cheap.



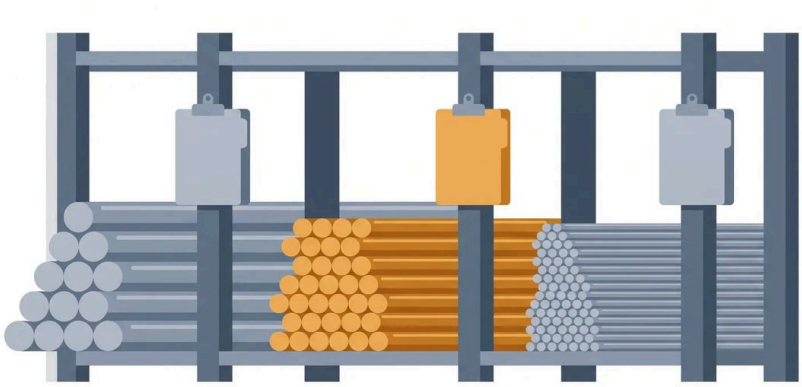
A crack forces a leakage field out of the surface, which gathers iron particles into a visible line.

Its limitation follows directly from its principle: **it works only on ferromagnetic materials.** Ferritic and martensitic steels can be inspected this way; austenitic stainless, aluminium, titanium, magnesium and composites cannot. The non-magnetic *metals* go instead to dye penetrant, eddy current, ultrasonic or radiographic inspection. Composites do not follow them onto the first two of those. Eddy current needs a good conductor: a glass or aramid laminate is not one at all, and carbon fibre conducts far too poorly and too directionally for a probe to read. Dye penetrant soaks into a porous laminate and contaminates it for any bonded repair afterwards. A damaged laminate is examined ultrasonically or radiographically, by tap testing, or — where the laminate is translucent — by stripping the paint and shining a strong light through it. A part must also be demagnetised (degaussed) after inspection, or residual magnetism will attract ferrous debris in service and can disturb nearby instruments. The methods themselves belong with the maintenance practices material.

Traceability — What the Paperwork Actually Records

Every batch of aerospace steel is made from a single melt, and that melt carries a **heat number** or **cast number** that follows the material from the furnace to the finished part. The certificate issued against that number records the chemical analysis, the melting practice, the product form and size, the heat-treated condition, and the results of the mechanical tests carried out on samples taken from that same melt.

Three practical rules follow, and they are the reason a stores procedure looks the way it does:



Different grades and heats are racked separately, each kept together with its own release paperwork.

- **Segregate.** Different grades, and different heats of the same grade, are stored separately and never allowed to become mixed. Once two bars of similar appearance are in the same rack without markings, both are unidentified.
- **Re-identify on cutting.** When a bar is cut, the offcut loses whatever identification was on the removed end. The heat number and specification must be transferred to the remaining piece before it leaves the saw.
- **Keep the release documentation with the material.** The certificate and the release document are what allow the finished part to be certified; without them the material cannot be used, however good it may be.

Heat Treatment of Steel

Heat treatment is the controlled heating and cooling of metals to alter their mechanical properties without changing the shape. It is one of the most important processes in aircraft steel manufacture and maintenance.

Hardening (Quenching)

The steel is heated to above its **upper critical temperature** (typically 750–900°C depending on alloy) until the crystal structure transforms to **austenite**. It is then rapidly cooled (**quenched**) in oil, water, or air. The rapid cooling traps the carbon atoms in the crystal lattice, forming **martensite** — an extremely hard but brittle structure.

Result: Maximum hardness and strength, but very low ductility and toughness. The steel is too brittle for most applications in this state — it must be tempered.

Tempering

After hardening, the steel is reheated to a moderate temperature (150–650°C) and held for a period, then cooled. Tempering **reduces brittleness** while retaining most of the hardness gained from quenching. The higher the tempering temperature, the softer and tougher (but less hard) the steel becomes.

Result: The desired balance between hardness and toughness. Different applications require different tempering temperatures — springs are tempered at around 300–400°C; cutting tools at 150–250°C.

Annealing

The steel is heated to above its critical temperature and then cooled **very slowly** (typically in the furnace itself). This produces the **softest possible** condition with maximum ductility.

Purpose: To soften the steel for machining, forming, or cold working; to relieve internal stresses; to refine the grain structure.

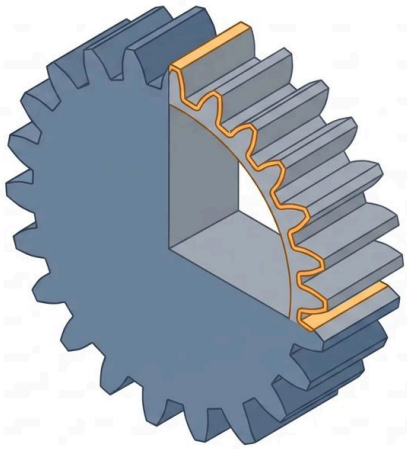
Normalising

Similar to annealing, but the steel is cooled in **still air** (faster than furnace cooling but slower than quenching). This produces a finer, more uniform grain structure than annealing.

Purpose: To refine grain structure after hot working (forging, welding); to produce a uniform structure; to improve machinability.

Case Hardening

A process that hardens only the **outer surface** of the steel while leaving the core soft and tough. This is ideal for components that need a wear-resistant surface but must withstand shock and fatigue loads (like gears and bearing races). Methods include:



Case hardening: a thin, hard, wear-resistant layer over a soft, tough core, as on a gear's teeth.

- **Carburising:** The steel is heated in a carbon-rich environment (solid, liquid, or gas). Carbon diffuses into the surface, increasing the surface carbon content. The part is then quenched to harden the carbon-rich surface layer.
- **Nitriding:** The steel is heated in an ammonia atmosphere at around 500°C. Nitrogen diffuses into the surface, forming extremely hard nitride compounds. No quenching is needed — the surface is hard as-treated. Excellent for fatigue resistance.
- **Induction hardening:** The surface is rapidly heated by electromagnetic induction and then quenched. Only the heated surface layer hardens.

The Anatomy of a Heat Treatment Cycle

Every treatment described above — and every one that follows — is built from the same four steps, and a specification controls all four. Understanding what each step is for makes the individual processes much easier to keep apart.

- **Heating rate.** Steel expands as it heats, and it does so unevenly if the outside heats much faster than the inside. A thick or complex part is therefore heated slowly, or in stages with a holding period part-way up, to keep the temperature difference through the section small. Rushing this step distorts or cracks the part before the treatment has even begun.
- **Soak temperature.** This is the step that decides which structure you are aiming at, and it is set relative to the steel's critical points, not to a universal number. A hardening or normalising temperature is above the upper critical point; a tempering or stress-relieving temperature is below the lower critical point.
- **Soak time.** Time at temperature is needed for two separate reasons: the centre of the part has to reach the same temperature as the surface, and the transformation itself has to complete. A common workshop guide is about one hour for every 25 mm of the ruling section — the thickest part of the component — but the approved process specification governs. Too short a soak leaves the core untransformed; too long a soak coarsens the grain, and coarse grain lowers toughness.
- **Cooling rate.** This is the step that decides what the austenite turns into, and therefore what the properties are. Everything from a furnace cool to a brine quench is a choice of cooling rate.

The Furnace Atmosphere Matters As Much As the Temperature

Steel heated in air above roughly 700 °C does two unwanted things at once, and both degrade the surface that will eventually carry the highest stress.

- **Scaling.** Iron oxidises to a loose, flaky black scale. The part loses dimension, the surface finish is destroyed, and the scale has to be removed afterwards.
- **Decarburisation.** Oxygen and water vapour react with the carbon in the surface layer and remove it as carbon monoxide. The surface is left with less carbon than the specification calls for, so after quenching it is *softer* than the core rather than harder, and its fatigue strength is reduced. This is a serious defect on a spring, a gear tooth or a fastener thread, because those are exactly the features where the stress peaks at the surface.

The remedies are to heat in a controlled protective atmosphere, in a vacuum furnace, or in a molten salt bath that excludes air from the surface; or, where none of those is available, to leave a machining allowance and remove the affected layer afterwards. A decarburised layer cannot be put right by re-heat-treating — the carbon has gone.

Hardening in Detail

Two decisions define a hardening operation: what temperature to austenitise at, and how fast to cool.

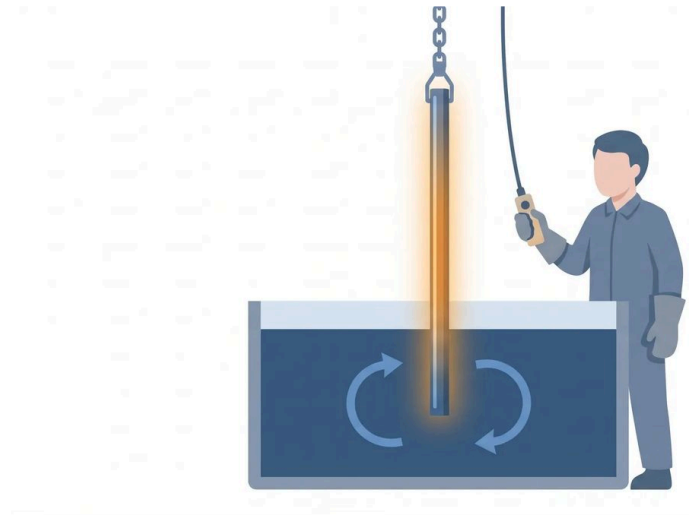
The austenitising temperature is normally set a modest amount — typically some tens of degrees — above the upper critical point for the grade. High enough to be certain the transformation is complete through the section; not so high that the grain coarsens or the surface scales and decarburises. There is one important exception. For a *hyper-eutectoid* steel — one carrying more than 0.83% carbon, such as a bearing or tool

steel — the part is deliberately heated to above the lower critical point but *below* the upper, so that the existing hard cementite particles are not dissolved. Keeping them undissolved gives a harder, more wear-resistant result and avoids the excess retained austenite that dissolving them would produce.

The quench medium is chosen to cool the part faster than its critical cooling rate, and no faster than necessary. Overshooting costs distortion and risks cracking.

Quench medium	Relative severity	Behaviour	Where it is used
Brine (salt solution)	Most severe — roughly twice water on the standard comparative scale	The dissolved salt disrupts the insulating vapour film, giving very rapid and unusually even cooling	Plain carbon steels of low hardenability in simple shapes. Highest distortion and cracking risk of all
Water	Severe — the reference value of 1.0 on that scale	Fast, but a clinging vapour blanket makes the cooling uneven, which is what produces distortion	Plain carbon steels; simple, symmetrical sections
Polymer solution	Adjustable between water and oil	The polymer concentration sets the severity, so one bath can be tuned to the steel	A controllable substitute for water or oil, and non-flammable
Oil	Moderate — roughly a quarter to a third of water	Slower and much more even; there are fast, medium and slow quenching oils, and warming the oil speeds it up by thinning it	The great majority of alloy steels, including 4130 and 4340. The standard aerospace quench
Molten salt	Moderate, but held at a chosen temperature rather than cold	Cools rapidly to the bath temperature and then stops, which is what makes martempering and austempering possible	Distortion-critical and thin-section parts
Gas (nitrogen or helium under pressure)	Mild, and adjustable with pressure and flow	Cools cleanly with no residue and no fire risk, inside a vacuum furnace	High-hardenability tool, die and aerospace steels
Still air	Mildest — roughly one fiftieth of water	Slow and gentle	Normalising, and the genuinely air-hardening steels whose hardenability is high enough to reach martensite in air

Two operating details change the result more than students expect. **Agitation** — moving the part through the quenchant or circulating the quenchant past it — breaks up the vapour blanket and can raise the effective severity of a given medium substantially, as well as making the cooling far more uniform. And the **direction of entry** matters: a long slender part is quenched vertically, end-on, so that it cools symmetrically, because entering it flat causes one side to cool first and bows it permanently.



A long, slender part is quenched vertically, end-on, so that it cools symmetrically instead of bowing.

⚠ **Quench cracking, and why a quenched part is never left overnight.** Different parts of a section cool and transform at different moments, and the resulting internal stresses can exceed the strength of the freshly formed martensite — which, before tempering, is the most brittle condition the steel will ever be in. Cracks start at sharp corners, at abrupt changes of section, at the edges of holes and at tool marks, and they may appear *hours after the quench* rather than during it. Three defences work: choose a steel with enough hardenability that a milder quench will do; design out sharp corners and abrupt section changes; and **temper immediately after quenching** rather than leaving the part standing in the as-quenched condition.

Two further points about the quench close the topic:

- **The transformation to martensite does not finish at room temperature in every steel.** It begins at a temperature called the martensite start and continues only while the temperature keeps falling. In high-carbon and heavily alloyed steels the finish temperature lies below room temperature, so a proportion of **retained austenite** survives the quench. Retained austenite is soft, and it may transform later in service or in storage, changing the part's dimensions. Where that matters — bearing races, gauges, precision spindles — a **sub-zero treatment** immediately after quenching drives the transformation to completion before tempering.
- **Hardenability decides how much of the section actually hardens.** The hardness reached at the surface is set by the carbon content; the depth to which that hardness penetrates is set by the alloy content. A large plain-carbon part quenched in water can be file-hard on the outside and completely untransformed a few millimetres in. The standard measurement of hardenability is the Jominy end-quench test, described with the other material tests later in this note.

Interrupted Quenching — Martempering and Austempering

Both of these use a molten salt bath held at a chosen temperature, and both exist to reduce the internal stresses an ordinary quench generates.

- **Martempering.** The part is quenched into a bath held just above the martensite start temperature and held there until the whole section has reached the bath temperature. Only then is it removed and allowed to cool

in air, so the martensite forms slowly and, crucially, at the *same time throughout the section* rather than surface-first. The result is the same martensitic structure with far less distortion and a much lower risk of cracking. It still requires tempering afterwards.

- **Austempering.** The part is quenched into a bath held *above* the martensite start temperature, in the band where austenite transforms to **bainite** — commonly in the region of 260 to 400 °C — and held there long enough for that transformation to go to completion. The difference from martempering is what the hold is for: martempering holds only until the section has equalised in temperature and then forms its martensite on the way out, while austempering holds until there is no austenite left to become martensite. Bainite is hard and unusually tough for its hardness, and because no martensite is formed the process produces very little distortion and *needs no subsequent tempering*. Its limitation is section thickness: the part must be thin enough for the whole section to reach the bath temperature before transformation begins.

Tempering in Detail

Tempering allows carbon to come out of the strained martensite lattice as fine carbide particles. As it does so the lattice relaxes, so the internal stress falls, ductility and toughness rise, and hardness falls. The amount of each depends almost entirely on the **temperature** chosen; time has a much weaker effect. As a rule, in plain carbon and ordinary low-alloy steels a higher tempering temperature gives a softer, tougher result. Two exceptions to that rule are set out below and both matter in practice: **secondary hardening**, where in steels carrying molybdenum, vanadium, tungsten or chromium the hardness stops falling and can rise again; and **temper embrittlement**, where certain steels lose impact toughness in a particular band with no change in hardness to warn you.

Approximate tempering range	What is happening in the steel	Result and typical application
150–250 °C	Carbon segregates and a very fine transition carbide precipitates. The lattice strain falls sharply while the carbide is still extremely fine	A large reduction in internal stress for only a small loss of hardness. Cutting tools, dies and other parts where hardness is everything
250–400 °C	Retained austenite transforms, and the transition carbide converts into cementite	Hardness drops noticeably while resilience reaches its best. Springs are tempered in this band
400–650 °C	The cementite particles coarsen and become rounded, and the ferrite matrix recovers	Hardness falls further and toughness climbs strongly. Highly stressed shafts, gears, landing-gear parts and structural fittings
500–600 °C, in steels containing molybdenum, vanadium, tungsten or chromium	Fine, stable alloy carbides precipitate for the first time	**Secondary hardening** — hardness stops falling and can rise again. This is why high-speed steel is tempered in this range rather than at a low temperature

Two refinements are worth knowing because they explain requirements that otherwise look arbitrary:

- **Double and triple tempering.** In high-alloy and tool steels the first temper transforms the retained austenite into *fresh, untempered* martensite, which is brittle. A second temper is therefore run to temper

that new martensite, and on some grades a third. The number of tempers is specified, and skipping one leaves brittle material in the part.

- **Temper embrittlement.** Certain steels lose impact toughness if they are tempered in, or cooled slowly through, particular temperature bands, without any change in hardness to warn you. In nickel-chromium steels the effect is associated with slow cooling through roughly the 375 to 575 °C region and with trace impurities gathering at the grain boundaries. It is countered by adding **molybdenum** — which is precisely the reason molybdenum appears in the alloying list as reducing temper brittleness — and, where the steel is susceptible, by cooling briskly from the tempering temperature rather than letting the part cool slowly in the furnace. The general rule that tempering ends in a gentle cool still holds for ordinary carbon and low-alloy steels; this is a specific alloy exception, and the process specification says which applies.

Temper Colours — a Workshop Guide and Nothing More

A clean, bright steel surface grows a thin oxide film as it is heated, and the film's thickness — and therefore its interference colour — changes progressively with temperature. This gives a traditional and genuinely useful way of judging tempering temperature by eye at the bench.

Approximate temperature	Oxide colour on bright steel	Condition produced
about 220 °C	Pale straw	Almost all the quenched hardness retained; cutting edges
about 240 °C	Dark straw	Slightly tougher; drills, taps, reamers
about 260 °C	Brown	Punches, shear blades
about 280 °C	Purple	Impact tools such as cold chisels and screwdrivers
about 300 °C	Dark blue	Springs and other resilient parts
about 320 °C and above	Light blue, then grey	Distinctly softer and much tougher

⚠ **Temper colours are not an approved process control.** They appear only on a clean, bright, bare surface, so they are invisible on a scaled, oiled, plated or painted part. They indicate the peak temperature the *surface* reached and say nothing at all about whether the core reached it, or for how long. And the colours can be produced on a surface by frictional heating — from grinding, from a seized bearing, from a brake — which is why blue or straw discolouration found on an installed component is treated as evidence of *overheating in service* and is investigated, not accepted. Aircraft parts are tempered in calibrated furnaces with a recorded temperature chart.

The Annealing Family

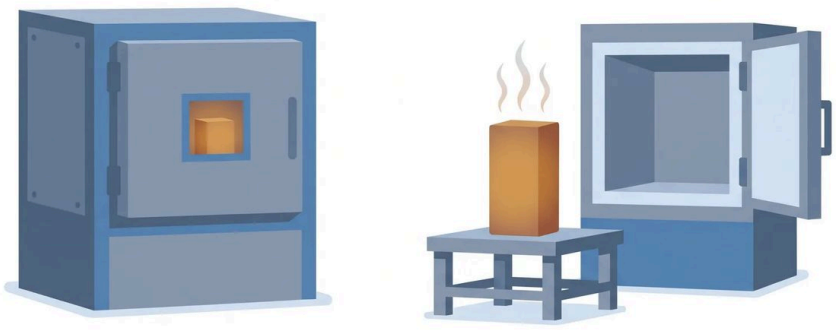
Annealing is not one process. Four distinct treatments carry the name, and they differ in the temperature reached and therefore in what they can achieve.

Treatment	Temperature	Cooling	What it achieves
Full annealing	Above the upper critical point	Very slowly, in the furnace	The softest and most ductile condition obtainable, with all internal stress relieved and the previous structure completely replaced. The grain ends up coarse, which is the price
Process (sub-critical) annealing	Below the lower critical point, typically in the 550–650 °C region	Air or furnace	Recrystallises a cold-worked structure and restores ductility so that further cold working can continue, without any phase change. Used between drawing or pressing operations
Spheroidise annealing	A long soak just below the lower critical point, or cycling gently around it	Very slowly	Rounds the cementite plates into discrete spheres. This is the softest and most machinable condition a high-carbon steel can be put in, and it is the standard condition for supplying tool and bearing steel for machining
Stress relieving	Below the lower critical point, typically 550–650 °C for carbon and low-alloy steels	Slowly, to avoid introducing fresh thermal stress	Reduces residual stress without changing the structure, the hardness or the strength. Applied after welding, heavy machining or forming

Residual stress is the stress locked into a component after manufacture and all heat treatment, with no external load applied at all — as distinct from working or applied stress, which exists only while the part is actually carrying a load. It comes from uneven cooling, from welding, from machining and grinding, and from cold forming. It matters because it adds to the service stress: a part carrying 300 MPa of locked-in tension reaches its yield point at a lower applied load than the drawing assumes, it distorts unpredictably when metal is machined off and releases the balance, and tensile residual stress at a surface makes stress-corrosion cracking and fatigue cracking far more likely. Stress relieving exists to remove it.

Annealing Against Normalising — Choosing Between Them

Both start from above the upper critical temperature and both relieve stress; the only difference in the process is the cooling rate, and every difference in the result follows from that one thing.



Full annealing cools the part in the switched-off furnace; normalising stands it in still air.

Aspect	Full annealing	Normalising
Where the part cools	Inside the furnace, with the furnace switched off	Out of the furnace, standing in still air at room temperature
Resulting pearlite	Coarse — thick, widely spaced plates	Fine — thin, closely spaced plates
Grain size	Coarse	Finer and more uniform
Hardness and strength	The lowest obtainable	Moderately higher than annealed
Ductility	The highest obtainable	Good, but below the annealed value
Impact toughness	Moderate — limited by the coarse grain	Better than annealed, because of the finer grain
Furnace time and cost	Long — the furnace is occupied throughout the cool	Short — the furnace is released as soon as the part comes out
Typically specified when	The part must be as soft and workable as possible for heavy machining or cold forming, or must be fully softened before re-hardening	The part must be returned to a uniform, fine-grained, stress-free condition after forging, welding, casting or rolling — the standard preparation before hardening

✈ **Why grain refinement is worth so much:** nearly every treatment in this section trades one property against another, but refining the grain raises strength *and* toughness at the same time. The mechanism is that permanent deformation spreads by slip along crystal planes, and a grain boundary blocks it — so more boundaries per unit volume means more obstruction, which is strength. At the same time a fine-grained structure has more boundaries to divert and blunt a running crack, which is toughness. This is why normalising is the standard treatment after forging and welding, and why a component that has been overheated to the point of grain coarsening is unserviceable even though nothing about it looks wrong.

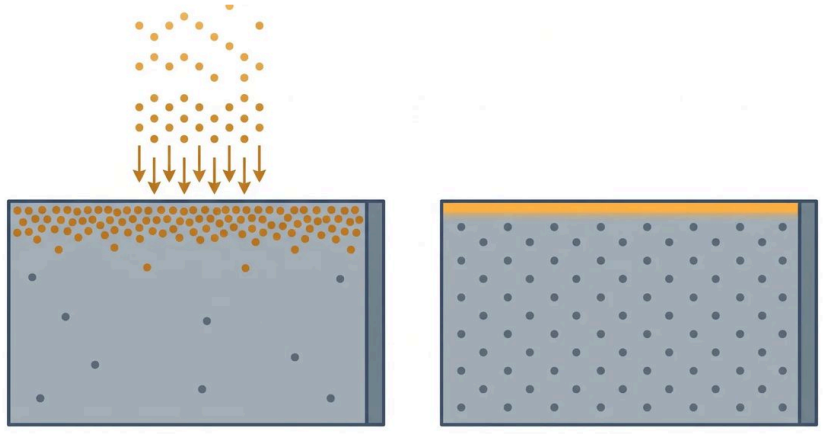
Case Hardening — Comparing the Processes

The object of every one of these processes is the same — a hard case over a tough core — but they reach it by two quite different principles, and confusing the two is a common error.

- **Carburising, carbonitriding and nitriding change the composition of the surface.** They diffuse an element into it that was not there before. They are the answer when the base steel does not contain enough carbon to harden.
- **Induction and flame hardening change only the temperature of the surface.** The composition is untouched; they simply austenitise and quench a thin layer of a steel that already contains enough carbon to respond. They are the answer when the base steel *does* contain enough carbon but you want only part of it hard.

Process	Temperature	Medium	Diffuses	Quench needed?	Typical case depth	Base steel required	Distortion
Pack carburising	about 900–950 °C	Solid charcoal and an energiser in a sealed box	Carbon	Yes	Deep — up to a few millimetres over a long cycle	Low carbon, roughly 0.1–0.25%	High — the whole part is soaked at temperature for many hours
Gas carburising	about 900–950 °C	A hydrocarbon gas atmosphere in a sealed furnace	Carbon	Yes	Controllable, from a few tenths of a millimetre upwards	Low carbon	Moderate, and far more controllable than pack
Liquid (salt bath) carburising	about 800–900 °C	Molten cyanide-bearing salts	Carbon, with some nitrogen	Yes	Shallow, and quickly produced	Low carbon	Moderate. The salts are extremely toxic and their use is tightly controlled
Carbonitriding	about 800–870 °C	Carburising gas with ammonia added	Carbon and nitrogen together	Yes, but a milder one	Shallow	Low carbon	Lower than carburising, because both the temperature and the quench are gentler

Process	Temperature	Medium	Diffuses	Quench needed?	Typical case depth	Base steel required	Distortion
Nitriding	about 500 °C — well below the lower critical point	Ammonia gas, a salt bath, or a plasma	Nitrogen only	**No** — the case is hard as treated	Thin, a few tenths of a millimetre	Alloy steels containing aluminium, chromium, molybdenum or vanadium	Minimal — no phase change and no quench
Induction hardening	Locally above the upper critical point	None — an induction coil heats the surface electrically	Nothing; it uses the carbon already present	Yes, usually an integral water spray	Set by the depth heated, and controllable	Must already contain roughly 0.35% carbon or more	Low and localised; can be applied to a finished part
Flame hardening	Locally above the upper critical point	None — an oxy-fuel torch heats the surface	Nothing; it uses the carbon already present	Yes, a following water quench	Generally deeper than induction and less precisely controlled	Must already contain roughly 0.35% carbon or more	Low and localised



Diffusion processes add carbon to the surface; induction and flame hardening only heat the surface.

Four practical points govern how case hardening is used on real components:

- **Case depth grows with the square root of time, not in proportion to it.** Diffusion is slow and gets slower as the concentration gradient flattens, so doubling the depth of a carburised case takes roughly *four* times as long at the same temperature. Deep cases are therefore expensive, and this is one reason deep-case components are specified only where the contact stress genuinely demands it.
- **Areas that must stay soft are masked.** A thin copper plating, or a proprietary stop-off paste, prevents carbon reaching the surface underneath. It is used where a feature will be machined after treatment, where

a thread must not be made brittle, or where a hole is to be drilled.

- **Carburising coarsens the grain, so a refining treatment usually follows.** Hours at 900 °C or more coarsens the grain of both the case and the core. The classic remedy is a two-stage reheat and quench: first from above the *core*'s critical point to refine the core, then from above the *case*'s critical point — which is lower, because the case is now high-carbon — to refine the case, followed by a low-temperature temper. Modern fine-grained steels often permit a direct quench from the carburising temperature instead.
- **Nitrided parts are hardened and tempered *before* nitriding.** The tempering temperature is chosen to be above the nitriding temperature, so that the core does not soften during the long nitriding cycle. This ordering is not optional; reversing it destroys the core properties. It also follows that a nitrided surface cannot be machined afterwards except by grinding, and any brittle white compound layer that forms is removed if the specification calls for it.

Nitriding earns a further note because of what it does to fatigue strength. The nitrides formed in the surface occupy more volume than the metal they replaced, so the case is left in a state of **compression**. An applied tensile stress at that surface must first cancel the built-in compression before the metal sees any net tension at all, which delays the initiation of a fatigue crack and raises the fatigue strength markedly. It is the same principle that makes shot peening effective.

Finally, the depth of a case is specified in one of two ways and they are not interchangeable. The **total case depth** is the full depth over which the composition has been altered. The **effective case depth** is the depth down to a stated hardness — commonly taken as 550 HV for a carburised case — and it is the figure that matters structurally, because it is the depth of genuinely hard material. The specification states which is required and what the value is.

Hydrogen Embrittlement — a Heat Treatment That Undoes Damage

High-strength steels — broadly, those heat treated above about 1,000 MPa tensile strength, though process specifications set the threshold in their own terms and some state it as a hardness — will absorb atomic hydrogen from acid pickling, acid cleaning, electroplating and some corrosion reactions. The hydrogen diffuses through the steel and collects where the internal stress is highest, and there it causes a delayed, brittle fracture at a load well below the design strength. The failures characteristically appear *after* the part has been in service for some time, which makes them particularly dangerous, and landing-gear steels are the classic case.

The remedy is a low-temperature bake carried out within a short specified period after the process that introduced the hydrogen, at a temperature commonly in the region of 190 °C for a number of hours set by the process specification. The bake drives the hydrogen back out before it has done its damage. Two things are worth being clear about: the bake must be done **promptly**, because once cracking has initiated it cannot be reversed; and no mechanical treatment — peening, polishing, stress relieving at the wrong temperature — substitutes for it.

Dimensional Change, and What You Must Never Do to a Heat-Treated Part

Martensite occupies slightly more volume than the austenite it came from, so a part *grows* very slightly on hardening. Combined with the distortion that uneven cooling produces, this is why highly accurate

components are rough machined, heat treated, and then finish ground to size — and why a machining allowance is left on the drawing for exactly that purpose.

⚠ **Local heating destroys a heat treatment, and it does so invisibly.** Welding a hardened or tempered steel part, flame straightening it, or grinding it aggressively enough to discolour the surface all re-heat a zone of the metal above its tempering temperature, and in the worst case above its critical point followed by a self-quench into fresh untempered martensite. The part looks unchanged and its dimensions are unchanged, but a band of it now has entirely different properties from the ones the design assumed. **None of these operations may be carried out on a heat-treated aircraft part outside an approved procedure**, and the same reasoning is why a component recovered from an overheat or a fire is inspected and hardness checked rather than assumed serviceable.

Heat treatment summary:

- **Hardening** = heat above critical temp → rapid cool (quench) → very hard, brittle
- **Tempering** = reheat hardened steel to moderate temp → improves toughness, reduces brittleness
- **Annealing** = heat above critical temp → very slow cool → softest, most ductile
- **Normalising** = heat above critical temp → cool in still air → refined, uniform grain
- **Case hardening** = harden surface only → hard outside, tough inside

Working and Handling Ferrous Materials

Heat treatment is not the only thing that changes a steel's properties. Every mechanical operation an engineer performs on it — bending, drawing, machining, grinding, welding — alters the metal as well, and several of those changes are permanent and invisible. This section covers what happens to steel when you work it, and the handling rules that follow.

Cold Working and Strain Hardening

Deform a metal permanently at room temperature and it becomes harder and stronger. This is **cold working**, and the strengthening it produces is called **strain hardening** or **work hardening**.

The mechanism is worth knowing because it explains every consequence. Permanent deformation happens by the movement of line defects called dislocations through the crystal lattice. As they move they multiply, and they run into one another and become entangled. A tangled dislocation is a dislocation that cannot move, so the more the metal is deformed the harder it becomes to deform it further.

Every property that moves at all does so in a definite direction, and those directions are opposite to the ones annealing produces. The third column below is the part worth learning, because it is what lets you predict the answer for a property this table does not list:

Property	Effect of cold working	Why it follows from the dislocation mechanism
Yield strength	**Increases** — markedly, and this is the largest effect of all	Yield is by definition the stress at which dislocations start moving in bulk, so obstructing them raises it directly and immediately

Property	Effect of cold working	Why it follows from the dislocation mechanism
Tensile strength	**Increases**, but by less than the yield strength does	The material is being pushed up its own strain-hardening curve, so it starts closer to the top of it. The yield and tensile figures converge
Hardness	**Increases**	Hardness measures resistance to permanent indentation, which is the same dislocation movement measured a different way
Ductility and malleability	**Decrease**	Part of the total deformation the metal can ever undergo has already been spent, so less remains before fracture
Toughness	**Decreases**	Toughness is the energy absorbed before fracture, and with less remaining deformation available there is less distance over which to absorb it
Internal (residual) stress	**Increases**	Different grains deform by different amounts and end up locked against one another. This is why a process anneal, not just a rest, is needed to relieve it
Young's modulus (stiffness)	**Unchanged**	Stiffness comes from the strength of the atomic bonds themselves, and tangling dislocations does not alter those. Nothing an engineer does to a steel changes its stiffness
Grain shape	Distorted and elongated in the direction of working	The grains are physically dragged out with the metal, which is why a cold-worked material has different properties along and across the working direction

Check the direction against a concrete case. A mild-steel rod is drawn through a series of progressively smaller dies to make wire. Each pass deforms it permanently. The finished wire has a tensile strength several times that of the rod it started as, and where the rod could be bent double, the wire snaps if it is bent too sharply. Strength up, ductility down — and it was the drawing, not any heat treatment, that did it.

The same mechanism explains a feature of the tensile test that puzzles students. Once mild steel has passed its yield point it continues to accept *more* load, not less, all the way up to the ultimate tensile strength. It does so because the plastic deformation is strain hardening it as it goes: the material is getting stronger faster than the specimen is getting thinner. Only when that balance tips does the specimen begin to neck, and necking — the local thinning near the end of the test — is the consequence, not the cause.

Cold working is used deliberately wherever it helps:

- **Wire drawing, cold rolling and cold heading** produce a finished item that is stronger than the stock it was made from.
- **Thread rolling** forms a thread by displacing metal rather than cutting it. The result is stronger in fatigue than a cut thread, both because the grain flow follows the thread form instead of being severed by it and because the root is left in residual compression.
- **Shot peening** bombards a surface with hard shot, deforming a thin layer and leaving it in **compression**. Since a fatigue crack must start under tension, a surface that begins in compression has to have that compression cancelled before it sees any net tension at all, which is why peening raises fatigue life.
- **Austenitic stainless steel is strengthened by cold work and by nothing else**, since heat treatment cannot harden it.

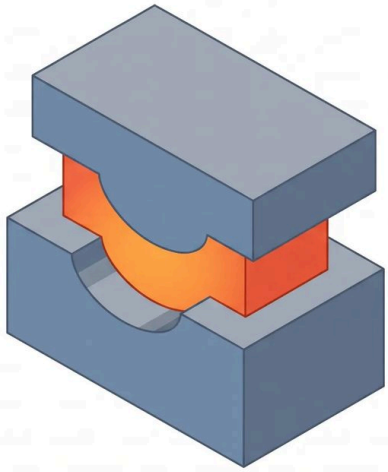
It also has to be controlled. A heavily cold-worked part has spent much of its ductility and is closer to cracking, which is why sheet-metal work has minimum bend radii, why a formed part must not be straightened and re-bent, and why a **process anneal** is inserted between severe forming operations to recrystallise the structure and restore the ductility before work continues.

Hot Working and Why Forgings Beat Billets

Work the same metal above its recrystallisation temperature and the picture changes completely. New, strain-free grains form as fast as the deformation distorts the old ones, so no strain hardening accumulates. Enormous shape changes become possible with modest force, and the metal never runs out of ductility.

The important by-product is **grain flow**. Hot working stretches the grains, and the inclusions and segregation between them, into continuous fibres that follow the shape being formed. A forging die shapes those fibres so that they run *along* the part, around its corners and through its fillets — that is, along the load path.

Now compare two identical-looking parts. One is closed-die forged to near its final shape and then finish machined. The other is machined out of a solid rolled bar. The forged part has continuous grain flow following its contours. The machined part has the bar's straight flow lines cut clean through, so their ends are exposed at every curved surface, and each exposed end is a potential crack origin. The two have the same chemistry, the same hardness and the same tensile test result — and markedly different fatigue lives. This is why landing gear components, engine mount fittings and highly stressed brackets are forgings, and it is one more reason a part may not be locally reshaped or built up in a way the designer did not intend.



Closed-die forging: hot metal is pressed between two shaped dies to near the part's final shape.

Hot working leaves a coarse grain and a scaled, decarburised surface, so it is normally followed by normalising and by removal of the affected surface layer.

Machining Steel

Cutting speed and tool life are governed principally by the hardness of the workpiece, and the relationship is simple: harder material means a lower cutting speed. A fully hardened steel cannot be turned or milled at all with conventional tooling and must be ground.

Machinability itself peaks in the middle of the hardness range and falls away at both ends, which is not obvious. Very soft, very ductile low-carbon steel is difficult to machine well because the chip tears rather than shearing cleanly and builds up on the tool edge, leaving a poor finish. Very hard steel wears the tool out. A normalised medium-carbon steel — or a spheroidise-annealed high-carbon one — machines best.

✈ **Drilling austenitic stainless steel — the technique and the reason for it.** Austenitic stainless has two awkward properties in combination: it *work hardens very rapidly*, and it conducts heat only about a third as well as carbon steel, so the heat generated at the cutting edge stays at the cutting edge instead of being carried away into the work. The two together mean that a drill which rubs instead of cutting will glaze and harden the surface immediately beneath it, and will then be unable to cut the surface it has just created. The technique that works is a **slow cutting speed with a firm, continuous feed**, using a drill ground to a **more obtuse point angle of about 120 degrees**, which reduces the cutting load and helps the lips bite in before the surface hardens. A sharper point and a fast speed — the combination that suits softer, free-cutting materials — is precisely wrong here: it generates the heat that hardens the metal ahead of the drill. Add a rigid set-up, a genuinely sharp drill, plenty of cutting fluid, and **never let the drill dwell in the hole while it is turning**.

Grinding Damage — the Defect You Cannot See

Grinding is the normal way of finishing a hardened steel part to size, and it is also the operation most capable of ruining one. A grinding wheel puts a great deal of energy into a very thin surface layer. If the wheel is blunt or loaded, the cut is too heavy, or the coolant is inadequate, that layer gets hot enough to change its own heat treatment:

- **Over-tempering (grinding burn).** The surface is reheated above its tempering temperature and becomes locally soft. It usually shows as a straw or blue discolouration, and it means the surface no longer has the hardness or the wear resistance the design assumed.
- **Re-hardening and grinding cracks.** If the layer gets hot enough to re-austenitise, the cold mass of metal underneath quenches it the instant the wheel passes, forming a thin skin of *untempered martensite*. That skin is brittle and in tension, and it cracks — typically in a fine network running at right angles to the grinding direction. Grinding cracks are a well-known cause of fatigue failure in hardened components.

The damage may be invisible to the naked eye, which is why hardened, ground aerospace parts are commonly subject to a nital etch inspection: a controlled acid etch that reveals over-tempered and re-hardened areas as distinct dark and light patches. Prevention is a sharp, correctly dressed, open-structure wheel, light cuts, and a generous flow of coolant delivered where the wheel meets the work.

Welding Ferrous Materials

Welding is a local heat treatment that nobody controls. The weld metal itself is a small casting; beside it is a **heat-affected zone** in which the parent metal has been heated to every temperature between melting and ambient, and then cooled at whatever rate the surrounding cold metal dictated — which in a thick or restrained joint is fast enough to count as a quench.

The consequence is that the heat-affected zone of a hardenable steel can end up as hard, brittle, untempered martensite, full of residual stress and vulnerable to cracking. How likely that is depends on how hardenable

the steel is, and hardenability is estimated for welding purposes by the **carbon equivalent**, which converts the alloy content into an equivalent amount of carbon:

Carbon equivalent (the widely used form):

- CE = C + (Mn / 6) + ((Cr + Mo + V) / 5) + ((Ni + Cu) / 15), with each element as its percentage
- For a mild steel with 0.20% C and 0.45% Mn: CE = 0.20 + 0.075 = **0.28**
- For 4130 with 0.30% C, 0.50% Mn, 0.95% Cr and 0.20% Mo: CE = 0.30 + 0.083 + 0.230 = **0.61**

A higher carbon equivalent means a more hardenable heat-affected zone and therefore a greater need for preheat, for controlled interpass temperature, for low-hydrogen consumables and for post-weld heat treatment. The two figures above show why a mild-steel bracket is welded without ceremony while a 4130 joint needs a procedure. But the carbon equivalent is only one input: joint thickness, restraint, heat input, the hydrogen content of the consumables and the cooling rate all bear on the outcome, which is why thin-wall 4130 tube is routinely welded under one set of conditions and a heavy 4130 section is not. **The approved welding procedure is what governs**, and it is written for the material, the thickness and the joint concerned.

Three further points cover most of what an engineer meets:

- **Welding a heat-treated part destroys its heat treatment locally.** There is no way around this. A part supplied hardened and tempered cannot be welded and left in that condition; either the whole assembly is re-heat-treated afterwards, or the repair is not permitted. This is why welded 4130 tube structures are welded first and normalised or stress relieved afterwards, and why the repair scheme — not the welder — decides what may be welded.
- **Austenitic stainless welds differently from carbon steel.** It expands about half as much again and conducts heat about a third as well, so the same weld puts far more distortion into it; fixturing and a controlled welding sequence matter more. And the heat-affected zone is at risk of sensitisation, which is why stabilised grades are used for welded assemblies.
- **Dissimilar joints need thought.** Joining a stainless to a carbon steel, or a high-carbon to a low-carbon steel, produces a weld metal of intermediate composition whose properties are not those of either parent. The filler is specified for that reason and is not a free choice.

Forming, Bending and Springback

Three material properties govern what can be done to sheet and tube:

- **Ductility** sets the minimum bend radius. The outside of a bend is in tension, and the tighter the radius the greater the strain there. Exceed the material's ductility and the outer fibre cracks. Bend radii are therefore quoted in terms of material thickness, and the governing figure comes from the structural repair manual or the drawing.
- **Grain direction** matters because rolled sheet is not the same in both directions. A bend made *across* the rolling direction — so that the bend line runs at right angles to the grain — tolerates a tighter radius than one made along it. This is why sheet is marked with its rolling direction and why nesting a part on a sheet is not arbitrary.

- **Yield strength** sets the **springback**. Only the plastic part of a bend is permanent; the elastic part recovers when the tool is released. Since Young's modulus is the same for every steel, the amount of stored elastic strain depends on the yield strength — so a *stronger* steel springs back *more*. Check the direction against a case: annealed mild steel bent to 90 degrees comes off the tool at very nearly 90 degrees, while a spring steel bent over the same tool opens out noticeably, and the tool has to over-bend it to compensate.

Marking Out

Accurate marking out on bright steel needs a background that a scribed line will show against clearly. The traditional and still-standard preparation is a swabbed coat of **copper sulphate** solution, which reacts with the clean steel surface to deposit an extremely thin, adherent film of copper. A scribed line cuts through that dull copper film to bright metal beneath and is easy to see, and it does not rub off the way a chalk or crayon mark does.



A scribed line cuts through the thin copper film to the bright steel beneath, so it is easy to see.

Two related products are often confused with it. **Engineer's blue** is a non-drying paste used to check the contact between two mating surfaces, not to mark out. A **wax crayon** gives a quick temporary mark on hot or oily surfaces but is far too coarse for accurate work.

⚠ **Copper sulphate is for steel only.** It must never be allowed to reach aluminium or magnesium alloys. On those metals it deposits metallic copper onto the surface, and the resulting couple drives rapid and severe corrosion that continues long after the solution has been wiped off. Keep the marking-out area, the swabs and the rags separated from light-alloy work. The corrosion mechanism itself is covered with the corrosion material later in this module. Remember too that a scribed line is a stress raiser in its own right: on a highly stressed part, scribing is confined to material that will subsequently be removed.

Protecting a Steel Surface

Plain carbon and low-alloy steels have essentially no natural corrosion resistance and depend entirely on an applied protective scheme; unprotected steel in a damp environment rusts, and the rust does not stop the attack the way an aluminium oxide film does. One treatment belongs here rather than with the corrosion material because it is specific to ferrous metals: **phosphating**. The steel is immersed in a solution of

phosphoric acid and metal phosphates, which reacts with the surface to grow an integral crystalline phosphate layer. It is neither a paint nor a plating — it is a chemical conversion of the steel itself — and it does two jobs: it inhibits corrosion in its own right, and its crystalline texture provides an excellent mechanical key for the paint scheme applied over it. Note that other acid baths are used for other purposes; a nitric or sulphuric acid solution is a pickling or cleaning bath, not a phosphating one. The full range of protective treatments, the plating systems and the corrosion mechanisms themselves are covered with the corrosion material.

Handling and Storage

- **Keep grades and heats segregated**, and transfer the identification to the remainder whenever stock is cut.
- **Protect bare machined surfaces.** A finish-machined high-strength steel part with no protective finish will begin to rust in hours in a humid store. Oil, vapour-inhibiting paper or an approved temporary coating is applied immediately.
- **Avoid contact with dissimilar metals** in storage and in assembly, and keep steel swarf and steel wool away from aluminium structure.
- **Treat mechanical damage seriously.** A nick, a scratch or a tool mark on a high-strength steel part is a stress raiser and a fatigue origin, and the higher the strength of the steel the more sensitive it is to one. Landing-gear components are protected in transit for this reason and not merely to keep them looking clean.

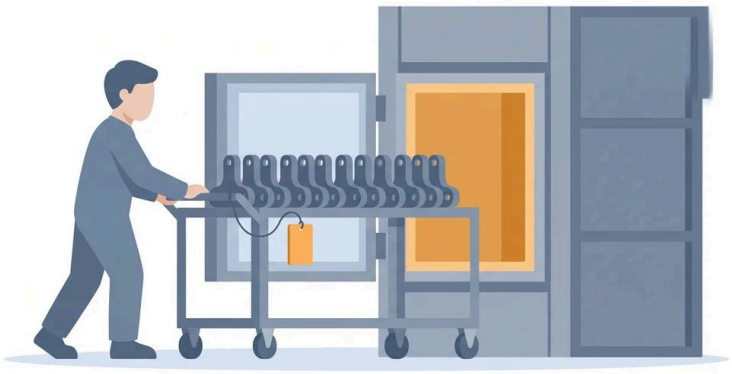
Testing of Ferrous Materials

Material testing answers two quite different questions, and it is worth being clear which one is being asked before choosing a test.

- **What are this material's properties?** This is the development and qualification question, answered once for a grade and condition and then written into a specification.
- **Is this particular batch, or this particular part, what it is supposed to be?** This is the acceptance and verification question, and it is the one an engineer meets in service. It is usually answered by a small number of quick, targeted checks rather than by a full characterisation.

Tests also divide by what they cost you:

- **Destructive tests** — tensile, impact, fatigue, bend, metallographic examination — consume the specimen. They are therefore *never* carried out on the part itself. Instead a **test coupon**, cut from the same bar or forging, from the same heat of steel, and passed through the same heat-treatment furnace load, travels with the batch and is tested in its place. The coupon is the evidence, and this is the practical reason heat and cast numbers have to be tracked so carefully.
- **Non-destructive tests** — the whole range of crack-detection methods, and hardness testing, which leaves only a very small indentation — can be applied to the part itself. Hardness is the workhorse for exactly that reason.



A coupon from the same heat rides through the same furnace load as the parts and is tested in their place.

One point applies to every test in this section: **a result only means anything if the test is standardised.** The specimen dimensions, the notch geometry, the indenter, the applied load, the rate of loading, the dwell time and the temperature are all laid down, because changing any of them changes the number. A hardness figure quoted without its scale, or an elongation figure quoted without its gauge length, is not a measurement — it is a number without units.

The four families that follow each answer a different question. Hardness asks how well the surface resists being pushed in, and by proxy how strong the material is and whether it was heat treated correctly. The tensile test asks how much load the material carries and how much it stretches first. Fatigue testing asks how long it survives loads that are repeated. Impact testing asks how much energy it absorbs when the load arrives suddenly. A material can pass any one of them and fail another, which is why all four appear in a specification.

Hardness Testing

- **Brinell test:** A hardened steel or tungsten carbide ball is pressed into the surface under a known load. The diameter of the indentation is measured and converted to a Brinell Hardness Number (BHN). Suitable for large, rough surfaces.
- **Rockwell test:** A diamond cone (Rockwell C scale) or steel ball (Rockwell B scale) is pressed into the surface. Hardness is read directly from the machine dial based on depth of penetration. Fast and widely used.
- **Vickers test:** A diamond pyramid indenter is pressed into the surface. The diagonal of the resulting square indentation is measured. Very accurate; suitable for thin materials and surface-hardened layers.

All three of the methods above work on the same principle: press a hard indenter into the surface under a controlled load, and measure how big a mark it leaves. A large mark means a soft material and a small one means a hard material. What separates the three is the shape of the indenter, the size of the load, and — the most important difference in practice — whether the indentation is measured *optically* after the load is removed or *electrically* as a depth while the load is still applied.

The Brinell Test in Detail

A ball indenter is pressed into the surface under a specified load, held there for a specified time, and removed. The *diameter* of the round indentation left behind is measured with a calibrated measuring microscope, and the hardness number is the applied load divided by the curved surface area of the impression.

Brinell hardness number:

- BHN = 2P / [pi × D × (D - sqrt(D² - d²))]
- P = applied load in kgf, D = ball diameter in mm, d = indentation diameter in mm
- The result has units of kgf/mm², but it is quoted as a bare number
- Load and ball diameter must be kept in a fixed ratio — P/D² is **30** for steel and cast iron, 10 for copper alloys, 5 for aluminium alloys — so that results taken with different ball sizes are comparable
- The standard combination for steel is therefore a **10 mm ball at 3,000 kgf**

Worked example — a Brinell test on a steel forging:

- 10 mm ball, 3,000 kgf load, indentation measured as **d = 4.0 mm**
- D² - d² = 100 - 16 = 84, so sqrt(84) = **9.165**
- D - 9.165 = 10 - 9.165 = **0.835**
- pi × D × 0.835 = 3.1416 × 10 × 0.835 = **26.23 mm²**
- BHN = (2 × 3,000) / 26.23 = 6,000 / 26.23 = **229**
- Applying the hardness-to-tensile correlation given in the properties section, 3.45 × 229 gives an expected tensile strength of about **790 MPa** — a figure that can be checked against the tensile test example later in this section.

The load is held for a specified **dwell time** so that the material finishes deforming and a stable, repeatable impression is produced before the load is released. The figure conventionally quoted for both the Brinell and the Vickers tests is about **15 seconds**. Too short a dwell gives an undersized impression and an inconsistent, falsely high reading.

Strengths. The indentation is large, so it averages over a substantial volume of material. That makes Brinell the right test for materials that are not uniform on a small scale — castings, forgings, coarse-grained stock — where a small indenter might land entirely inside one constituent and read the wrong thing. It also tolerates a comparatively rough surface.

Limitations. The large indentation is a defect on a finished part and can only be placed where it does no harm. The test is unsuitable for thin material or for a hardened case, because the indenter reaches through the layer being tested into the softer material beneath. And there is an upper limit: above roughly 450 on the scale a hardened *steel* ball begins to flatten and reads low, which is why a tungsten carbide ball is used for harder materials — and even then the test is not used at the very top of the range.

The Rockwell Test in Detail

The Rockwell test measures a *depth*, and it does so in two stages, which is what makes it fast and free of optical measurement.

- A small **minor load** — conventionally 10 kgf — is applied first. This seats the indenter, beds down any surface roughness and any dirt or scale, and establishes the zero datum. Because the datum is set after this seating load, surface imperfections are largely removed from the result.
- The **major load** is then applied and removed. The depth the indenter has moved *beyond* the datum, measured with the minor load still applied, is the permanent penetration, and the machine converts it directly into a hardness number on a dial or display.

How the Rockwell number is formed:

- The permanent increase in depth is counted in units of **0.002 mm**. Call that number of units *e*
- Diamond-cone scales (including HRC): hardness = **100 - e**
- Ball scales (including HRB): hardness = **130 - e**
- Both are subtractions, so a *deeper* indentation gives a *lower* hardness number — which is the correct direction, since deeper means softer

Scale	Indenter	Total (major) load	Used for
HRC	Diamond cone of 120 degrees included angle, with a small spherical tip (the "brale")	150 kgf	Hardened and tempered steels — the standard aerospace scale for heat-treated parts
HRB	1.588 mm (1/16 inch) hardened steel ball	100 kgf	Softer steels, annealed and normalised material, copper and aluminium alloys
HRA	Diamond cone	60 kgf	Very hard and thin material, cemented carbides, thin case-hardened layers
Superficial scales (15N, 30N, 45N and the corresponding T scales)	Diamond cone or ball	15, 30 or 45 kgf, with a 3 kgf minor load	Thin sheet, shallow cases, nitrided surfaces and coatings, where a normal load would punch through the layer being measured

Strengths. It is fast — a reading takes seconds; it needs no optical measurement and no operator judgement; the indentation is small; and it works over the whole range from soft to fully hardened simply by changing scale. This is why it is the test found in nearly every workshop and why heat-treatment acceptance is normally written in Rockwell.

Limitations. The part must be properly supported on the anvil, because any flexing of the specimen is measured as penetration and reads falsely soft. The surface must be reasonably clean and flat. Readings on different scales are not directly comparable. And with a thin specimen the indenter can deform the anvil face beneath — the classic "anvil effect" — which again reads falsely soft.

The Vickers Test in Detail

A diamond indenter ground as a square-based pyramid with 136 degrees between opposite faces is pressed into the surface. It leaves a square impression, and the two *diagonals* are measured optically and averaged.

Vickers hardness number:

- HV = 1.854 × F / d²
- F = load in kgf, d = mean of the two diagonals in mm
- The 1.854 comes from the geometry of the 136-degree pyramid, so it never changes

The property that makes Vickers special follows from that geometry. Because a pyramid is *geometrically similar* at every size, a bigger load simply makes a bigger impression of exactly the same shape, and the calculated hardness comes out the same. So **the Vickers number is independent of the load**, and a single continuous scale covers everything from soft aluminium to the hardest tool steel. Brinell and Rockwell both need the scale to be changed as the material changes; Vickers does not.

Strengths. One scale for all materials; very small, precisely defined indentations that can be placed on thin sheet, on tiny features and, most usefully, in a line running inward from a surface to survey a case-hardening depth; and the highest accuracy of the three.

Limitations. The surface must be prepared to a fine, often polished, finish, because the diagonals of a small impression cannot be measured on a rough surface. The measurement is optical and therefore slower, and it depends on the operator setting the measuring lines correctly. This is a laboratory test rather than a shop-floor one.

Comparing the Three

	Brinell	**Rockwell**	**Vickers**
Indenter	Hardened steel or tungsten carbide **ball**	**Diamond cone** (C scale) or **steel ball** (B scale)	**Diamond pyramid**, 136 degrees
Quantity measured	**Diameter** of the impression	**Depth** of penetration	**Diagonal** of the impression
How it is read	Optically, then calculated or read from a table	Directly from the machine — no calculation	Optically, then calculated
Indentation size	Large	Small	Very small
Number of scales	One, with the load matched to the material	Many, and they are not interchangeable	**One, for all materials**
Best suited to	Castings, forgings, coarse or non-uniform material, rougher surfaces	Routine shop and acceptance testing across the whole hardness range	Thin sections, small parts, case-depth surveys, laboratory work
Main drawback	Large mark; upper limit on hardness; thin sections unsuitable	Needs solid support; scales not comparable with each other	Needs a prepared surface and careful optical measurement

⚠ **Hardness scales convert only approximately, and only within limits.** Published conversion tables between Brinell, Rockwell and Vickers exist and are useful, but they are empirical relationships derived for particular material groups — principally plain and low-alloy steels — and they are not exact. Converting a reading and then quoting the converted value as though it were measured is a real source of dispute. If a specification calls for a Rockwell C figure, measure Rockwell C. Note also that a hardness figure is meaningless without its scale, so a bare number on a job card should never be accepted.

Portable and In-Situ Hardness Testing

Bench machines require the part to be brought to them and supported rigidly, which rules out anything installed. Two families of portable instrument exist for in-situ work, and both carry conditions:

- **Rebound (Leeb) testers** fire a small impact body at the surface and compare its rebound velocity with its impact velocity. A hard surface returns more of the energy. The distractor in a Rockwell question that mentions "the rebound height of a ball" describes this principle, not the Rockwell one, and the classic instrument of this kind is the Shore scleroscope.
- **Ultrasonic contact impedance testers** press a Vickers diamond into the surface under a fixed load and detect the change in the resonant frequency of the probe rod, which depends on the contact area.

Both require the part to be massive enough and thick enough not to move or ring under the probe; on a thin panel or a light unsupported component they read low, and a correction or a fixture is required. Both need a good surface finish. They are comparative field tools, and where an acceptance figure is in dispute the bench test is the reference.

Typical Hardness Values

Some idea of the range is useful, both for sanity-checking a reading and for recognising when a part has not been heat treated as intended.

Material and condition	Approximate Brinell hardness
Annealed low-carbon steel	about 100 to 175
Normalised medium-carbon steel	roughly 170 to 220
Alloy steel, hardened and tempered for structural use	roughly 300 to 450
Cobalt-bearing tool and high-speed steel	about **600 to 700**
Carburised or nitrided case	Beyond the useful Brinell range — measured in Vickers or on a superficial Rockwell scale

Tensile Testing

A standard test specimen is pulled in a tensile testing machine until it breaks. The test measures:

- **Ultimate tensile strength (UTS):** Maximum stress the material can withstand
- **Yield strength:** Stress at which permanent deformation begins
- **Elongation:** Percentage increase in length at fracture (measure of ductility)

- **Reduction of area:** Percentage decrease in cross-section at the fracture point

The Tensile Specimen

A tensile test measures a material property, so the answer must not depend on the size of the piece tested. That requires the specimen to be standardised, and in one respect the standardisation is not obvious. The **gauge length** — the marked length over which extension is measured — is made *proportional* to the cross-section, because a short specimen concentrates all of its necking inside a short gauge length and so reports a higher percentage elongation than a long one cut from identical material. The usual proportional gauge length works out at about **five diameters** for a round specimen.

The consequence for the engineer is a rule with no exceptions: **an elongation figure is meaningless without the gauge length it was measured over.** Two certificates quoting 18% and 22% elongation may describe the same steel measured to different standards.

The specimen has enlarged, threaded or shouldered ends so that it grips securely and always fails in the reduced parallel portion, and generous blend radii where the section changes so that the failure is not triggered by a stress concentration at the shoulder. It is pulled at a controlled rate — pull too fast and the measured strength comes out high — while an extensometer records the extension across the gauge length, until it fractures.

Reading the Stress-Strain Curve

Plot stress against strain and the whole mechanical character of the steel appears in one line. Take it region by region.

Region	What is happening in the metal	What is read off
Straight-line portion	Atomic bonds stretch elastically. Everything is recoverable	The slope is **Young's modulus**. The top of the straight line is the **limit of proportionality**
Elastic limit	Slightly beyond the straight line; strain is still fully recoverable, but no longer proportional	The last point from which the specimen returns to its original length
Yield	Dislocations break free and move in large numbers; deformation becomes permanent	**Yield strength** — the stress at which permanent deformation begins
Strain-hardening region	Dislocations multiply and tangle, so the material strengthens faster than the specimen thins	The rising portion; the material accepts increasing load
Maximum	Strengthening can no longer keep pace with the reduction in area	**Ultimate tensile strength** — the maximum stress the material withstands
Necking	Deformation localises into one narrow band, which thins rapidly	The curve falls, because engineering stress is calculated on the *original* area while the real area is shrinking fast
Fracture	The neck separates	Elongation and reduction of area, both measures of **ductility**

Two features of the yield region are worth separating, because they distinguish the steels an engineer meets.

- **Mild steel has a genuine, visible yield point** — in fact two. The load rises to an *upper* yield point, then drops abruptly to a *lower* yield point at which the specimen extends appreciably with no increase in load at all. The cause is that interstitial carbon and nitrogen atoms have gathered around the dislocations and pinned them; once the stress is high enough to tear the dislocations free, they move much more easily, so less stress is needed to keep them going. This is a low-carbon-steel phenomenon and it produces the visible surface bands seen on mild steel pressings.
- **Heat-treated alloy steels have no distinct yield point.** The curve simply bends over smoothly, and there is no single point that can be identified as "yield". For these the specification quotes a **proof stress** instead: the stress that leaves a defined permanent strain — commonly 0.1% or 0.2% — after unloading. It is found by drawing a line parallel to the elastic portion, offset along the strain axis by that amount, and reading where it cuts the curve. The 0.2% proof stress is what "yield strength" means on a certificate for an alloy steel.

Worked example — a complete tensile test result:

- Round specimen, original diameter **12.5 mm**, gauge length **62.5 mm** (five diameters)
- Original area A = pi/4 × 12.5² = **122.72 mm²**
- Load at the 0.2% proof point: **72.0 kN** → 72,000 / 122.72 = **587 MPa proof stress**
- Maximum load reached: **96.0 kN** → 96,000 / 122.72 = **782 MPa ultimate tensile strength**
- Gauge length after fracture, pieces fitted together: **76.0 mm** → elongation = (76.0 - 62.5) / 62.5 = 13.5 / 62.5 = **21.6% on a 62.5 mm gauge length**
- Diameter at the neck after fracture: **8.9 mm** → final area = pi/4 × 8.9² = 62.20 mm², so reduction of area = (122.72 - 62.20) / 122.72 = **49.3%**
- **Cross-check:** the ratio of proof stress to ultimate strength is 587 / 782 = 0.75, which is normal for a hardened and tempered alloy steel. And the Brinell example earlier in this section predicted about 790 MPa from a hardness of 229 — within roughly one per cent of the 782 MPa measured here. The two independent tests agree, which is exactly what makes a hardness check an acceptable proxy.

What the Numbers Are Used For

- **Proof or yield stress** is the practical design limit for a structural part. A component that has yielded is out of tolerance and is unserviceable, even though it has not broken — which is exactly why a bolt stretched beyond its elastic limit is scrapped rather than reused. It deforms *permanently*; it does not spring back, and being stretched does not make it more ductile.
- **Ultimate tensile strength** sets the ultimate load capability and is the figure the hardness correlation predicts.
- **Elongation and reduction of area** are the ductility measures, and they are the numbers that tell you whether the material will give warning before it fails. A steel with high strength and 2% elongation is a very different proposition from one with the same strength and 15%. Reduction of area is the more sensitive of the two, because it measures what happened at the neck rather than averaging over the whole gauge length.

Heat treatment moves the whole curve in a predictable way. Annealed steel gives a low, long curve: low strength, large elongation. Quenched and lightly tempered steel gives a high, short one: high strength, very little elongation, and a fracture with no necking at all. Increasing the tempering temperature moves the curve progressively from the second shape towards the first. The tensile test therefore verifies a heat treatment directly, and does it more completely than a hardness check because it reports the ductility as well.

Finally, the broken specimen is itself evidence. A ductile steel necks visibly and breaks with a fibrous, grey **cup-and-cone** fracture, the cone of one half fitting the cup of the other. A brittle one shows no necking, breaks square across, and leaves a bright, faceted, crystalline surface. The appearance of a fracture face is the first thing an investigator looks at, and the same distinction applies to a service failure as to a test piece.

Fatigue Testing

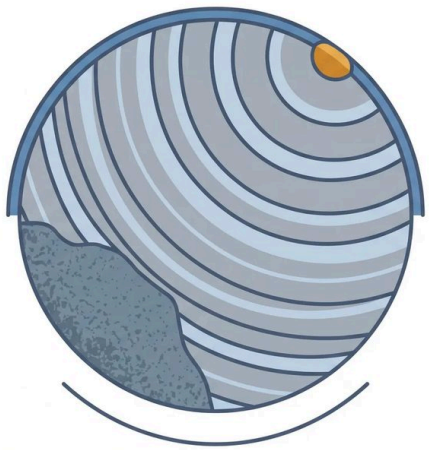
A specimen is subjected to repeated cyclic loading (bending, torsion, or axial) at a specified stress level until it fails. The number of cycles to failure is recorded. Testing at various stress levels produces an **S-N curve** (stress vs number of cycles). Steel typically has a definite **fatigue limit** — a stress level below which it can endure an infinite number of cycles without failing.

How a Fatigue Failure Actually Develops

Fatigue is not a gradual weakening of the whole part. It is a single crack, growing a little on each load cycle, in three distinct phases:

- **Initiation.** A crack starts at the point of highest local stress, which is almost always at a *surface* and almost always at a stress raiser — a fillet, a hole, a thread root, a keyway, a tool mark, a corrosion pit, an inclusion just below the surface. This phase can occupy the great majority of the component's life and produces nothing detectable.
- **Propagation.** The crack extends a minute distance on each cycle, opening under the tensile part of the cycle and closing again. The two faces rub as they open and close, so the fracture surface produced in this phase is smooth and often burnished.
- **Final fracture.** Eventually the remaining uncracked section is too small to carry the peak load, and it fails in one cycle as a straightforward overload. This last stage is instantaneous.

The fracture face records all of this and is the most useful single piece of evidence in a failure investigation:



A fatigue fracture face: beach marks curve away from the origin; a rough final overload zone fills the rest.

Feature	What it looks like	What it tells you
Origin	A single point, often at a corner, a hole edge or a change of section, from which the markings radiate	Where the crack started, and therefore what caused it
Beach marks (progression marks)	Smooth, curved bands like the tide lines on a beach, curving away from the origin	Interruptions or changes in the loading — each band is a period of different service. Visible to the naked eye
Striations	Extremely fine parallel lines, visible only under an electron microscope	One line per load cycle — they allow the crack growth rate to be measured directly
Ratchet marks	Small steps running back towards the surface at the origin end	Several cracks started at slightly different points and merged — a sign of a high stress or a severe stress raiser
Final overload zone	A rough, dull, fibrous or crystalline area quite unlike the smooth fatigue region	The part that broke in one go. Its *size relative to the smooth region* indicates the stress level: a **small** final zone means the crack grew a long way before the part could no longer carry the load, so the stress was low; a **large** final zone means the load was high and only a short crack was needed

The S-N Curve, the Endurance Limit and What Modifies It

Testing many specimens at different stress levels and plotting stress against cycles to failure produces the S-N curve. For steel the curve falls steeply at first and then flattens into a horizontal line — the "knee" — typically somewhere between one million and ten million cycles. The stress at that flat portion is the **fatigue limit**, also called the endurance limit. Below it the steel survives indefinitely; above it the life is finite and falls sharply as the stress rises. The stress level, not the proof stress, is what governs which of those two regimes a part is in.

The useful rule of thumb is that for most steels the fatigue limit in fully reversed loading lies at roughly **40% to 60% of the static ultimate tensile strength**, and it is commonly taken as one half. Two qualifications belong with it. Above a tensile strength of roughly 1,400 MPa the fatigue limit stops rising in proportion — making a steel stronger past that point buys very little extra fatigue performance while making it much more sensitive to notches. And the flat knee is characteristic of ferrous materials tested in *air*: in a corrosive environment it disappears entirely and the curve keeps falling, so a steel exposed to corrosion has no true endurance limit at all. The behaviour of the light alloys, which do not show the knee in the first place, is covered with the non-ferrous materials.

The S-N data themselves come from small, carefully polished laboratory specimens, and a real component never matches one. The published figure has to be reduced to allow for the differences:

Influence	Direction of the effect on fatigue strength	Why
Rougher surface finish	**Reduces** it	Every machining mark is a small notch at the surface, and the surface is where cracks start
Notches, holes, fillets, thread roots, keyways	**Reduces** it, often severely	They multiply the local stress well above the nominal value
Larger component size	**Reduces** it	A bigger part contains more material at high stress, so it is more likely to contain a critical defect
Corrosion, or a corrosive environment	**Reduces** it, and removes the endurance limit	Pits act as notches and the environment keeps attacking the freshly exposed crack tip
Fretting at a clamped joint	**Reduces** it	Micro-movement damages the surface and generates debris, creating an origin in an area under high clamp-up
Decarburisation of the surface	**Reduces** it	The surface layer is left softer and weaker than the material beneath it, exactly where the stress peaks
Tensile residual stress at the surface	**Reduces** it	It adds directly to the applied tensile stress
Shot peening, nitriding, cold rolling of fillets and threads	**Increases** it	All three leave the surface in *compression*, which the applied tension must first cancel before the metal sees any net tension
A polished surface, and clean vacuum-remelted steel	**Increases** it	Fewer surface notches and fewer internal inclusions means fewer places for a crack to start

Worked example — from a laboratory figure to a component allowable:

- A steel has an ultimate tensile strength of **1,000 MPa**.
- Taking the fatigue limit as one half of the UTS, the polished-specimen figure is 0.5 × 1,000 = **500 MPa**.
- Suppose the surface-finish and size corrections for this particular machined component work out at **0.75** and **0.85**. (These are illustrative — the real values are read from design data for the finish, the size and the material, and they are not universal constants.)
- Component fatigue limit = 500 × 0.75 × 0.85 = **319 MPa**.
- That is under **a third** of the tensile strength — and it takes no account of a notch, of corrosion or of a mean tensile stress, each of which would reduce it further. This is why a repeatedly loaded part is designed to a stress that looks absurdly conservative beside the material's UTS, and why fatigue, not static strength, sizes most aircraft structure.

How the Test Is Run

- **Rotating bending.** The classic machine, and the one the original S-N work was done on. A round specimen is loaded in bending and rotated, so every point on its surface passes from maximum tension to maximum compression once per revolution. This gives fully reversed loading — equal tension and compression, zero mean stress — which is the reference condition most published data are quoted for.
- **Axial (push-pull).** The specimen is pulled and pushed along its axis, which allows any combination of mean and alternating stress to be applied. This matters because a *tensile* mean stress lowers the fatigue strength while a *compressive* mean stress raises it, and most real components carry a mean load rather than a purely alternating one.
- **Torsional and resonance machines** apply the loading in other modes, and resonance machines run at high frequency to accumulate very large cycle counts in reasonable time.

Two features of fatigue testing catch people out. First, the **scatter is large**: two identical specimens at the same stress can differ in life by a factor of several, because the outcome depends on where a crack happens to start. Fatigue data are therefore statistical and require many specimens, and design values are set at a low probability of failure rather than at the average. Second, a specimen that survives a pre-agreed number of cycles is stopped and recorded as a **run-out** rather than being run to destruction, since no test can literally demonstrate infinite life.

In service, all of this appears as the life limits and inspection intervals on the maintenance schedule. A component whose failure would be critical is either given a *safe life* — retired after a set number of cycles, well short of the demonstrated fatigue life, whether or not anything has been found wrong with it — or is designed to be *damage tolerant*, so that a crack can be found by inspection and the structure will carry its load until it is. Heat treatment cannot restore fatigue life that has already been consumed, and no maintenance action resets the count; this is why life-limited parts are tracked individually by serial number.

Impact Testing

- **Charpy test:** A notched specimen is placed as a beam and struck by a heavy pendulum. The energy absorbed in breaking the specimen (measured in joules) indicates toughness. Low values indicate brittle material.

- **Izod test:** Similar to Charpy but the specimen is held as a vertical cantilever. Less commonly used than Charpy in aviation.

Charpy and Izod Compared

Both tests swing a calibrated pendulum hammer into a notched bar and measure the energy the bar absorbs in breaking. The energy is found from the difference in the pendulum's height before and after the blow: whatever it fails to carry through to the far side of its swing went into fracturing the specimen. The machine reads it directly in joules. The differences between the two tests are in how the specimen is held and where it is struck.

	Charpy	**Izod**
Specimen	Square bar 10 mm × 10 mm × 55 mm long	Square bar 10 mm × 10 mm × 75 mm long
How it is held	Laid horizontally and **simply supported at both ends**, like a beam	Clamped vertically in a vice, **as a cantilever**
Notch position	At mid-span, on the face **away from** the striker	Level with the top of the vice jaws, facing **towards** the striker
Where the hammer lands	At mid-span, on the face **opposite** the notch	Above the notch, on the **same** side as the notch
Striking energy of the standard machine	Commonly around **300 J**	Approximately **163 J**
Practical position	The international standard for structural steels, and the easier of the two to run at controlled sub-zero temperatures because the specimen is simply laid on the anvil and struck within seconds of leaving its cooling bath	Long used in Britain; less common in aviation work, and awkward for low-temperature testing because clamping the specimen in a vice takes time during which it warms

The **notch is not incidental** — it is the whole point of the test. An unnotched ductile bar would simply bend around the hammer. The notch forces the fracture to start at a known place and, more importantly, creates the sharp, three-dimensionally constrained stress state in which even a normally ductile steel can behave in a brittle manner. What the test measures is therefore not raw toughness but *notch toughness*: how the material behaves when there is already a defect in it, which is the realistic case. The notch geometry is standardised precisely because a sharper notch gives a lower result on the same material.

A result may be reported in three complementary ways, and a specification often calls for more than one:

- **Absorbed energy**, in joules. The primary figure. A low value means a brittle material.
- **Lateral expansion** — how much the specimen bulged sideways at the compression face before breaking. It measures how much plastic deformation actually took place.
- **Fracture appearance** — the percentage of the broken face that is dull and fibrous (a ductile, energy-absorbing failure) as opposed to bright and crystalline (a brittle one). It is judged against reference charts and is the most immediate of the three.

The Ductile-to-Brittle Transition

Run the same impact test on the same steel at a series of temperatures and the result is not a constant. It falls as the temperature falls, and over a fairly narrow band it falls steeply, from a high "upper shelf" value to a low "lower shelf" one. The temperature at the middle of that drop is the **ductile-to-brittle transition temperature**.

Direction check with a concrete case: an ordinary ferritic structural steel that absorbs a large amount of energy and breaks with a fibrous, heavily deformed face at room temperature may, at forty degrees below zero, absorb only a small fraction of that energy and break flat and crystalline with no deformation at all. Nothing about the steel has changed except its temperature. **Colder means more brittle**, without exception, for this class of material.

Whether a metal shows the transition at all depends on its crystal structure, and this is one of the most useful practical consequences of the phases covered earlier in this note:

- **Body-centred cubic metals show a transition.** That covers ferrite and martensite, and therefore plain carbon steels, low-alloy steels, and the ferritic and martensitic stainless grades.
- **Face-centred cubic metals do not.** Austenitic stainless steels keep their toughness all the way down to cryogenic temperatures, which is one reason they are chosen for very cold service as well as very hot.

What moves the transition temperature is worth knowing, because it is what a specification is controlling when it calls for a Charpy value at a stated temperature. Raising the carbon content, coarsening the grain, allowing phosphorus to remain, sharpening the notch, increasing the section thickness and increasing the rate of loading all push the transition to a *higher* temperature — that is, they make the steel brittle in conditions it would otherwise have survived. Adding nickel and refining the grain push it *lower*. This is why nickel steels are chosen for low-temperature duty, and it is a further reason grain refinement is valued.

The aviation relevance is direct. Airframe structure is cold-soaked at cruise altitude, ground equipment and aircraft operate in arctic conditions, and components exposed to ram air or in contact with cold fuel run well below ambient. A steel that is perfectly tough in the hangar can be at the bottom of its transition curve in service, and a designer specifies impact energy *at the lowest temperature the part will see*, not at room temperature.

Other Tests an Engineer Meets

The Jominy End-Quench Test

This is the standard measurement of **hardenability** — the depth to which a steel hardens, as opposed to the hardness it reaches. A standard bar, 25 mm in diameter and 100 mm long, is heated to its austenitising temperature and then mounted vertically in a fixture where a jet of water of specified size, flow rate and temperature plays on the *lower end only*. The bottom of the bar is therefore quenched violently while the top cools only in air, and every intermediate cooling rate exists somewhere along its length.

Once cold, two flats are ground along the bar and Rockwell C hardness is measured at intervals from the quenched end. Plotting hardness against distance gives the steel's hardenability curve, and it separates the two variables cleanly:

- The hardness **at the quenched end** is set by the **carbon content**, because that end has certainly cooled fast enough to form martensite.
- The rate at which the hardness **falls off with distance** is set by the **alloy content**. Two steels of the same carbon content start at the same hardness at the quenched end; the more hardenable of the two falls away more slowly, and that shallower slope is exactly what "high hardenability" means.

Bend Testing

A strip or a welded coupon is bent around a former of specified radius through a specified angle, and it passes if no crack opens on the outside of the bend. It is a simple, direct measure of ductility in the form that matters for sheet work, and it is a standard part of welding procedure qualification, where it demonstrates that the weld and its heat-affected zone are no less ductile than the parent metal.

Metallographic Examination

A sample is cut, mounted, ground, polished and then etched with a suitable reagent, and examined under a microscope. It is the only method that shows directly what the structure *is*, and it answers questions no mechanical test can:

- Is the structure the one the heat treatment was supposed to produce — pearlite, tempered martensite, bainite?
- How coarse is the grain? Grain size is reported as a standard number, and the convention catches people out: **the higher the number, the finer the grain**, because each step up doubles the number of grains counted in a given area.
- How deep is the case, and how deep is any decarburised layer?
- How clean is the steel — how many non-metallic inclusions, and of what type?
- A coarse *macro*-etch of a whole section, examined by eye rather than under a microscope, reveals grain flow in a forging, and with it forging laps, bursts and flow lines that have been machined through.
- An etch applied to the surface of a finished ground part reveals grinding damage as patches of differing shade, as described earlier under grinding: an over-tempered area is attacked more readily and comes up **dark**, while a re-hardened skin of untempered martensite resists the etch and comes up **white**.

Creep and Stress-Rupture Testing

A specimen is held at a constant load and a constant elevated temperature, and its extension is recorded against time. A *creep test* runs long enough to establish the steady, secondary-stage strain rate, which is the number used to calculate how long a part will last before it stretches beyond its allowable limit. A *stress-rupture test* simply runs to failure and records the time taken. Both are needed to set the temperature and stress limits on hot-section components, and both take a very long time, which is why the data are extrapolated using established time-temperature relationships rather than measured directly at service durations.

Testing the Part Rather Than the Material

Two categories of test apply to a finished component instead of a specimen. A **proof load test** applies a specified load, higher than the working load but below the design yield, and the part passes if it shows no permanent set and no damage. It is used on shackles, lifting equipment and certain fittings, and because it consumes a little fatigue life it is done only where the design data permit it. And the whole family of **non-destructive testing** methods — penetrant, magnetic particle, eddy current, ultrasonic and radiographic inspection — look for cracks and internal defects in the actual part without harming it. Which of them can be used on a given item depends on the material: magnetic particle inspection needs a ferromagnetic material, eddy current needs an electrical conductor, and penetrant needs only a non-porous surface but finds only surface-breaking defects. The methods themselves are covered with the maintenance practices material.

Choosing the Right Test

The question being asked	The test that answers it	What you get
Was this part heat treated to the correct condition?	Hardness, normally Rockwell C	A number to compare against the range on the drawing, obtained in seconds without damaging the part
How deep and how hard is the case?	A Vickers hardness traverse across a sectioned coupon	The hardness profile, and from it the effective and total case depth
How strong is the material, and how much warning does it give?	Tensile test	Proof stress, ultimate tensile strength, elongation and reduction of area
Will it survive repeated loading, and for how long?	Fatigue test	The S-N curve and the fatigue limit
Will it take a sudden blow, and at what temperature does it stop doing so?	Charpy or Izod impact test at a series of temperatures	Absorbed energy and the ductile-to-brittle transition temperature
How deeply will this grade harden in a given quench?	Jominy end-quench test	The hardenability curve
What is this material?	Optical emission spectrometry, or portable X-ray fluorescence for the alloying elements alone	A chemical analysis
Is the internal structure right, and is there decarburisation, grinding damage or coarse grain?	Metallographic examination and surface etching	The microstructure, the grain size, and the depth of any affected layer
Will it stretch under a steady load at temperature?	Creep and stress-rupture testing	The secondary creep rate and the time to rupture
Is there a crack in this particular part?	The appropriate non-destructive method for the material	The presence, position and extent of the defect

✈ **Aviation context:** All aircraft structural steels must meet rigorous specification requirements. Material certificates (mill test reports) accompany every batch of steel used in aviation, documenting chemical composition, mechanical test results, and heat treatment. Full traceability from raw material to finished aircraft part is required by EASA Part 21 and Part 145 regulations.

Module 7 — Maintenance Practices

7.1 — Safety Precautions — Aircraft and Workshop

Safety is the overriding priority in all aircraft maintenance activities. The maintenance environment contains numerous hazards — high-voltage electricity, compressed gases, flammable fluids, toxic chemicals, heavy components, and moving machinery. Every maintenance engineer must understand these hazards, know how to prevent accidents, and be able to take immediate remedial action when incidents occur. All EASA Part 66 licence categories require **Level 3 (detailed knowledge)** of this section.

General Safety Principles

Safe working practices are governed by a hierarchy of controls:

Priority	Control Type	Example
1 (Highest)	**Elimination**	Remove the hazard entirely (e.g. substitute a toxic solvent with a non-toxic one)
2	**Engineering controls**	Ventilation systems, machine guards, interlocks
3	**Administrative controls**	Procedures, training, warning signs, permits to work
4 (Lowest)	**Personal Protective Equipment (PPE)**	Safety glasses, gloves, hearing protection, steel-toe boots

Why the Hierarchy Is Ordered That Way

The order in the table is not a ranking of how much protection each control gives on a good day — it is a ranking of how *reliably* each one keeps working on a bad day. Elimination and substitution remove the hazard from the workplace permanently and for everyone, including the apprentice who was not at the briefing and the contractor who arrives at 02:00. Engineering controls work without anyone deciding to use them: an extraction hood pulls fume away whether or not the operator remembers it is switched on, and an interlock stops a machine whether or not the guard was opened deliberately. Administrative controls — procedures, training, signs, permits — only work if a person reads them, understands them, remembers them under time pressure and chooses to follow them. Personal protective equipment sits at the bottom because it protects one person, only while it is worn, only if it was correctly selected for that specific hazard, and only if it fits and is in good condition.

There is a second reason PPE ranks last that matters more than the first: PPE **fails without warning**. A respirator whose cartridge is saturated goes on looking exactly like a respirator that is still filtering, and the wearer's first indication may be the smell of solvent — by which time they have already been exposed. A glove

permeated by a solvent looks undamaged from the outside while the chemical is in contact with the skin underneath and is being held there. An extraction system that stops working, by contrast, is usually obvious, and an interlocked guard that fails normally fails to the safe side by stopping the machine.



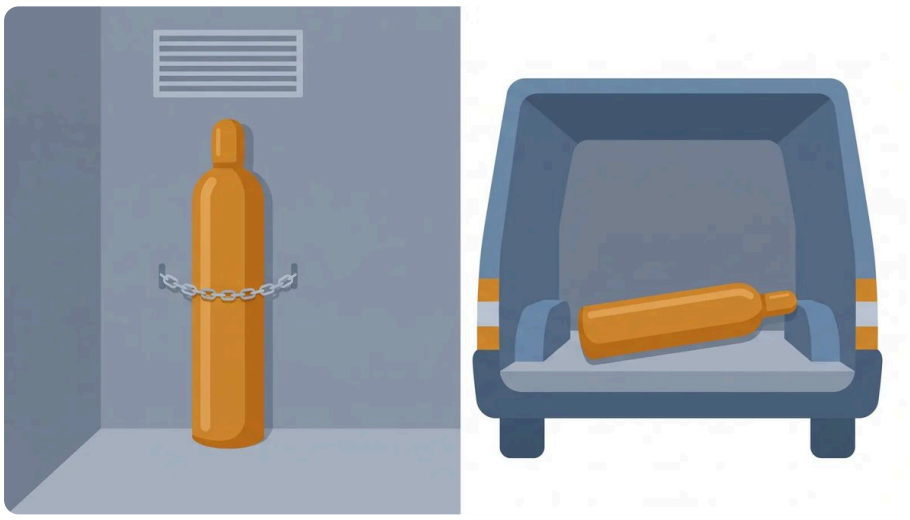
A saturated cartridge looks exactly like one still filtering; the first sign may be the solvent smell.

In practice the tiers are combined rather than chosen between. A chromate-primer spraying task illustrates all four at once: buy the chromate-free primer where the specification allows it (elimination by substitution); spray in a booth with downdraught extraction (engineering); restrict the area, control who is in it and set a re-entry time (administrative); and put the sprayer in an air-fed hood and a coverall (PPE). Removing any one of those does not simply reduce protection a little — it moves the whole task's safety onto the tiers below it, which are the less reliable ones. That is also why "the operator was wearing a mask" is never, on its own, an adequate answer to an exposure question.

Whatever controls are applied, some **residual risk** always remains. The purpose of the assessment is to reduce risk to a level that is acceptable and then to make sure everyone involved knows what is left and how to react to it — not to pretend the hazard has gone.

Hazard, Risk and Risk Assessment

The two words are not interchangeable. A **hazard** is anything with the potential to cause harm — a 200 bar nitrogen cylinder, a drum of MEK, a 3 metre drop from a wing. **Risk** is the combination of how likely that harm is and how severe it would be if it happened. The nitrogen cylinder is an extremely serious hazard, but a cylinder chained upright in a ventilated store with its valve cap fitted presents a low risk; the same cylinder lying loose in the back of a van presents a high one. Controls do not change the hazard, they change the risk.



The same nitrogen cylinder: chained upright in a ventilated store is low risk; loose in a van is high risk.

A risk assessment is a structured, recorded process with five recognisable steps: identify the hazards; decide who might be harmed and how (including people who are not doing the work — the cleaner, the visitor, the crew on the flight deck above you); evaluate the risk and decide on controls, working down the hierarchy from the top rather than reaching straight for PPE; record the significant findings and communicate them to everyone affected; and review, both periodically and whenever something changes.

⚠ **Common exam trap:** risk assessment applies to **every** task and process carried out, not only to those involving hazardous chemicals, and not only to work above some trigger height. Restricting assessment to the obviously dangerous jobs is exactly how the routine task that injures somebody escapes control — most maintenance injuries happen during work nobody thought was dangerous.

Assessments come in two forms and both are needed. A **generic** assessment covers a repeated activity in a standard environment — wheel change, battery servicing, panel removal — and lives in the organisation's procedures. A **task-specific** assessment covers the job in front of you today, with its actual aircraft, its actual weather, its actual access equipment and whoever else is working nearby. On top of both sits **dynamic** assessment: the continuous judgement an engineer makes as conditions change, and the authority to stop when they do. Reviews are triggered by new equipment, a change of process or product, a change of location, an accident or near miss, or simply the passage of time.

The Safe System of Work

The output of an assessment is a safe system of work: the agreed sequence of steps, isolations, controls, equipment and people that makes a specific job safe. It is what turns "be careful" into something auditable. A complete one names the competent person in charge, the isolations required and who applies them, the access equipment, the PPE, the atmospheric or electrical tests to be passed before work starts, the

communication arrangements, the emergency and rescue plan, and how the system is formally stood down at the end so the aircraft can be returned to service.

Two habits carry most of the practical value. The first is the pre-task brief, in which the person leading the job walks the team through the hazards, the controls and the stop conditions before anyone picks up a tool — the point being that everyone hears the same plan at the same time, including the parts that concern somebody else's task on the same aircraft. The second is the discipline that safety, of yourself and of everyone in the immediate area, ranks ahead of finding the right tool and ahead of the aircraft's departure time. That is not a slogan: an injured engineer cannot complete the task, occupies the two colleagues who go to help, and frequently creates a second casualty. Schedule pressure is a recognised precursor to maintenance error, which is precisely why the priority order has to be settled before the pressure arrives rather than during it.

Permits to Work

A permit to work is a formal written authorisation, issued by a competent person, to carry out a defined task, in a defined place, within a defined period, subject to stated precautions. It is an administrative control — the third tier of the hierarchy — and its value is that it forces the precautions to be checked and signed for before work starts rather than assumed. Its essential features are the same whatever the hazard:

- **A single point of control.** One issuing authority, one copy displayed at the worksite, one register. Two teams cannot each hold a permit for conflicting work on the same system.
- **A defined scope and validity period.** The permit expires; it does not roll over to the next shift by default, because the conditions it certified may not have.
- **Cross-referenced isolations.** The permit lists the isolations and lock-off points, tying it to the locks and tags physically fitted to the equipment.
- **Formal hand-back.** The worker signs that the job is finished or suspended and the area is safe; the issuer signs that the system has been accepted back and the isolations may be removed. An unreturned permit is treated as work still in progress.

In a maintenance organisation the three permits an engineer meets most often are hot work (any spark-, flame- or heat-producing operation near an aircraft), confined space entry (fuel tanks above all), and isolation of a hazardous energy source before work on a system. Each is dealt with in its own section of this note.

Safety Signs: Shape and Colour Carry the Meaning

European safety signs are standardised so that their meaning can be read at a distance, in poor light, and by someone who does not speak the local language: the *shape and colour* give the category, and the pictogram gives the detail. Text is an optional supplement, never the primary message.

Category	Shape and Colour	Meaning	Hangar Example
Prohibition	Round, white background, red border with a red diagonal bar, black pictogram	You must not do this	No smoking; no naked flames; do not operate
Mandatory	Round, blue background, white pictogram	You must do this	Eye protection must be worn; hearing protection zone
Warning	Triangular, yellow background, black border and pictogram	Beware, hazard present	Flammable material; overhead load; laser; toxic
Safe condition	Square or rectangular, green background, white pictogram	The safe way, or safety equipment	Emergency exit; assembly point; first aid; eyewash
Fire equipment	Square or rectangular, red background, white pictogram	Location of fire-fighting equipment	Extinguisher point; fire alarm call point; hose reel

Note that green carries one meaning in the scheme and red carries two, in quite different senses: green is always "safety" (escape, rescue, first aid) and red is either a prohibition or the location of fire equipment. A red sign therefore never means "danger" in this system — danger is the yellow triangle. Getting that backwards matters on the hangar floor, where the green running-man sign is the one you look for in smoke.

Personal Protective Equipment: Selecting It and Knowing Its Limits

PPE is only protective if it is matched to the specific hazard. The categories below are the ones an aircraft engineer uses daily, with the selection point that most often gets missed.

- **Eye protection.** Three different hazards need three different products. Impact from grinding, drilling and chiselling needs impact-rated spectacles or, better, a face shield worn over them, because a face shield alone does not seal at the sides. Chemical splash needs sealed goggles — open-sided spectacles will not stop a splash of hydraulic fluid running down the frame into the eye. Radiation from welding or from a laser needs the correct filter shade. European impact protection is graded, in increasing order of impact energy, as S, F, B and A, and the grade is marked on both the lens and the frame — both must carry the rating, since a lens that holds is useless in a frame that lets go.
- **Hand protection.** A glove is chosen against the specific chemical from the manufacturer's compatibility data, not from what is in the box on the bench. Every glove material has a measurable *breakthrough time* for each chemical, after which the chemical is passing through the intact glove; the general-purpose nitrile glove that is excellent against mineral oil is a poor choice against phosphate-ester hydraulic fluid, for which butyl rubber is the usual recommendation. Gloves are also a hazard in their own right near rotating machinery: they are the mechanism by which a hand is drawn into a drill or a lathe, so gloves come *off* at those machines.
- **Foot protection.** Safety footwear protects against three distinct things: a dropped component (toe cap), standing on a drill bit or a swarf spike (penetration-resistant midsole), and electrical contact or static

(insulating, antistatic or conductive soles, which are three different specifications — conductive footwear intended to drain static is not electrical insulation and must never be relied on as such). The requirement applies on the ramp exactly as it does in the hangar.

- **Body protection.** Coveralls that fit and fasten, with nothing loose to catch: no ties, no lanyards, no dangling identity cards near machinery or an engine intake. Jewellery, watches and rings come off — a ring is both a conductor and, caught on a projection during a fall from a stand, the classic degloving injury.
- **Respiratory protection.** A disposable filtering facepiece is graded FFP1, FFP2 or FFP3 in increasing order of protection; a cartridge respirator must have the right filter class for the substance and the cartridge has a limited service life once opened. Any tight-fitting mask needs a *face-fit test* on the individual wearing it, and any beard growth along the sealing line defeats the seal completely — for a wearer who cannot be clean shaven the answer is a loose-fitting powered or air-fed hood, not a tighter strap. For the most hazardous work (isocyanate paint spraying, disturbing asbestos-containing material) an air-fed system is the norm rather than a filter at all.
- **Hearing protection.** Covered with the noise hazard later in this note; the selection point is that more attenuation is not automatically better.

All PPE has to be inspected before use, kept clean, stored where it cannot be damaged or contaminated, and replaced on condition. A pair of goggles kept in a tool drawer with the swarf is not eye protection, and a filtering facepiece stuffed in a pocket between shifts is contaminated on the inside — the side that goes against the face.

Competence, Supervision and the Authority to Stop

Part-145 requires the organisation to provide facilities and a working environment appropriate to the work being carried out, and to ensure that the environment does not impair the effectiveness of personnel — temperature, lighting, dust, contamination and noise are named as things that must be controlled, not tolerated. That is a regulatory hook the engineer can use: work that cannot be done to standard in the conditions available is work that should be stopped and rescheduled, and doing so is compliance rather than obstruction.

The same logic applies to the individual. Nobody carries out a task for which they are not trained, authorised and equipped, and every person on the job has the authority to stop it if a control is missing, a condition has changed or something simply does not look right. An organisation in which stopping the job is treated as a failure will not hear about the near misses either, and the two problems have the same root.

⚡ Electrical Safety

Electricity is one of the most dangerous hazards in aircraft maintenance. Both the aircraft's own electrical systems (typically 28 V DC and 115 V AC at 400 Hz) and workshop mains supplies (230 V AC / 50 Hz in

Europe) can cause fatal electric shock.

How Electric Current Injures

It is **current**, not voltage, that injures. Voltage matters only because it is what drives current through the body's resistance, which is why the table below is written in milliamps. Four factors decide the outcome of a shock: how much current flows, for how long, along which path, and at what frequency.

- **Path.** Hand-to-hand and hand-to-opposite-foot are the dangerous paths because they take the current across the chest and through the heart. A path confined to one hand, such as finger to finger, may cause a severe local burn without reaching the heart at all. This is the reason for the old rule of keeping one hand in a pocket when working near a live circuit that cannot be isolated, and for never standing on damp concrete or a steel dock while doing it.
- **Duration.** The heart is only vulnerable to fibrillation during a short window in each beat, so a shock lasting less than one cardiac cycle is far less likely to trigger it than one lasting several seconds. This is exactly what a residual current device exploits: it cannot prevent the shock, it can only cut it short.
- **Frequency.** Mains-frequency AC at 50 or 60 Hz is close to the worst case for the heart, because the current reverses often enough to keep re-stimulating the muscle through its vulnerable window. When a shock lasts longer than one heartbeat, direct current needs a considerably higher current than AC to fibrillate the heart; letting go of a gripped conductor is also less difficult with DC — which does not make DC safe. The a.c. threshold values in IEC TS 60479-1 are given only for 15 to 100 Hz, so no lower hazard can be assumed at the aircraft's 400 Hz: 115 V AC 400 Hz is treated as fully lethal and given exactly the same respect as the workshop mains.

The specific injuries follow from the mechanism. The **let-go** effect happens because the flexor muscles that close the hand are stronger than the extensors that open it, so a current large enough to tetanise both clamps the hand shut on the conductor and the victim cannot let go — the shock then continues for as long as the supply does. That is also why a rescuer must never simply grab the casualty. **Respiratory arrest** can occur without fibrillation if the current paralyses the diaphragm and intercostal muscles. **Ventricular fibrillation** is the usual cause of death: the heart's coordinated contraction breaks down into a useless quiver and circulation stops, and it will not restart on its own. **Burns** occur at the entry and exit points, where current density is highest, and along the path through the body — and the internal damage is routinely far worse than the small mark on the skin suggests.

Body Resistance: Why 230 V Kills and 28 V Usually Does Not

Almost all of the body's electrical resistance is in the outer layer of dry skin. Internally, from one hand to the other, the body is a bag of salt water measuring only a few hundred ohms. Dry, intact, callused skin can measure tens or hundreds of kilohms; skin that is damp with sweat, wet with coolant or cut measures a small fraction of that, and skin resistance also falls as the applied voltage rises because the outer layer breaks

down. The practical consequence is that the same contact that gives a harmless tingle in a dry workshop in winter can be fatal in a humid hangar to someone who has been working hard.

Shock current from Ohm's law:

$$I = \frac{V}{R}$$

where V is the voltage across the body and R is the total resistance of the path, including the contact resistance at both ends.

Worked example — the same person, two supplies.
Take an engineer whose hand-to-hand resistance is 1,000 Ω because their hands are damp. Contact across the 230 V workshop supply gives 230 / 1,000 = 0.23 A, which is **230 mA** — far beyond the let-go range and above the 50–150 mA band for which the table below lists respiratory arrest and possible death. The table's figures are for current flowing hand to foot for one second, so comparisons like these are only indicative; HSE regards hand-to-hand contact as possibly the most dangerous path of all.

Now take the same person with dry hands, resistance about 10,000 Ω, touching a 28 V DC busbar: 28 / 10,000 = 0.0028 A, or **2.8 mA**. That is barely above the threshold of perception. This is why the aircraft's 28 V DC system is rarely a shock hazard — and it is exactly why the warning box further down this section is about burns rather than shock. The battery that supplies those 28 volts can deliver hundreds of amps into a near-zero resistance, so a ring, a watch strap or a dropped spanner bridging a busbar to structure carries a current limited only by the metal itself, and reaches red heat in seconds while welding itself to the finger it is on.

The general rule that follows is that *every* supply is treated as dangerous until measured, and that the question "how many volts is it?" is never a complete risk assessment on its own: you also have to ask how much current the source can deliver, and what the resistance of the path will be.

Effects of Electric Current on the Human Body

Current (hand to foot, 1 second)	Effect
1 mA	Tingling sensation — threshold of perception
5 mA	Slight shock, not painful but disturbing — most people can still let go
6–25 mA (women); 9–30 mA (men)	Painful shock, loss of muscular control — the **"let-go" range**, where the victim may be unable to release the grip
50–150 mA	Extreme pain, **respiratory arrest**, severe muscular contractions — death possible
1,000–4,300 mA	The heart's rhythmic pumping stops (**ventricular fibrillation**, in NIOSH's wording) — death likely
10,000 mA	Cardiac arrest, severe burns — death probable

These figures are from OSHA's publication *Controlling Electrical Hazards*, which NIOSH reproduces. Published thresholds, especially for ventricular fibrillation, differ widely with the duration and path of the current — 7.4 compares them.

Electrical Safety Precautions

- **Isolate before work** — use lockout/tagout (LOTO) procedures. Apply personal padlocks and warning tags to circuit breakers and switches.
- **Verify dead** — after isolation, test the circuit at the point of work with a suitable voltage detector, and prove that detector on a *proving unit* (or a known live source of similar voltage) immediately before and after the test.
- **Static (ESD) protection** — when handling electrostatic-sensitive devices, wear an approved anti-static wrist strap (which contains a 1 MΩ current-limiting resistor) connected to the aircraft/equipment structure. This is an ESD precaution, not a shock precaution — never bond yourself to structure on a circuit that may be live.
- **Residual Current Devices (RCDs)** — all portable power tools and workshop outlets should be RCD-protected (typically 30 mA trip).
- **Capacitor discharge** — capacitors in power supplies and ignition units can retain lethal charges long after power is removed. Always discharge and short-circuit before handling.
- **Low-voltage working** — use reduced-voltage tools (110 V centre-tapped to earth, giving 55 V to earth) where practical.
- **High-voltage areas** — radar systems, HF transmitters, and weather radar magnetrons produce hazardous voltages (up to 30 kV). Observe exclusion zones and interlocks.

> ⚠ **Critical:** Never assume a circuit is dead. Always test. Even 28 V DC can cause serious burns if short-circuited through a low-resistance path (e.g. a ring or watch strap). Remove all jewellery before working on electrical systems.

Safe Isolation: Prove, Test, Prove

"Verify dead" is a three-part sequence, and leaving out any part of it destroys the whole thing. First **prove** the test instrument on a known live source or on a dedicated proving unit. Then **test** the isolated circuit, checking every combination of conductors — line to neutral, line to earth and neutral to earth on a single-phase supply, and every pair on a three-phase one, because a circuit can be dead between two conductors and live between two others. Then **prove the instrument again**, on the same known source. The final step exists because an instrument that fails between the first proving and the test would give you a comforting zero on a live circuit, and you would have no way of knowing.

The instrument itself matters. A general-purpose multimeter is the wrong tool for proving dead: it can be on the wrong range, its internal fuse can be blown, its battery can be flat, and its high input impedance can display a misleading "ghost" voltage induced from an adjacent cable — or, worse, display nothing at all when the fault is in the meter. The correct instrument is a two-pole voltage indicator with fixed leads, no range switch and a self-check, designed so that a failure shows up as a failure rather than as a zero. A non-contact voltage "pen" indicates only the presence of a field; it never proves absence, and it must not be used for this purpose.

Where work genuinely cannot be done dead — and that is a decision for an authorised person against a written justification, not one an individual makes at the panel — the controls stack up rather than substitute for one another: tools insulated and rated for the voltage present, insulating gloves and matting, screening or barriers over adjacent live parts, a defined working space, and a second competent person in attendance who knows how to isolate and how to release the casualty. Insulated tools are examined before every such use, because a nick or a crushed spot in the insulation puts the hand at conductor potential without any visible warning. The single most common failure on high-voltage work is treating one of those controls as sufficient on its own — a pair of rubber gloves is not a system of work.

The residual current device mentioned above works on a completely different principle from a fuse or a circuit breaker, and understanding the difference tells you what it will and will not do. It compares the current flowing out in the line conductor with the current returning in the neutral. In a healthy circuit they are equal; if some current is leaking to earth — for example through a person — they are not, and the device trips on that imbalance in a few tens of milliseconds. A 30 mA device therefore does not prevent the shock, it limits how long it lasts, which is precisely the duration factor discussed above. Two limitations follow directly from the principle: an RCD *cannot* detect a shock in which the current passes from line to neutral through the body, because that current still balances, and it offers no protection at all against contact with the DC side of a supply or a charged capacitor. Its test button is there to be used, and it tests only the tripping mechanism, not the earthing.

Reduced-voltage tooling attacks the problem from the other end. A 110 V supply fed from a transformer whose secondary winding is centre-tapped to earth puts each of the two output conductors 55 V from earth, so the worst voltage a person standing on the hangar floor can be exposed to is halved. The tool still receives its full 110 V between the conductors and works normally — the protection is in where the earth reference sits, not in the tool.

Lockout/Tagout Covers Every Energy Source, Not Just Electricity

The word "isolation" is easy to read as "switch the power off", and that reading has killed people. A maintenance task is isolated only when *every* source of hazardous energy that could reach the part being worked on has been positively cut off, locked and, where it is stored rather than supplied, released or restrained:

- **Electrical** — aircraft batteries, external ground power, the APU generator, and any local mains supply to test or servicing equipment.
- **Hydraulic** — pumps off is not enough; accumulators hold system pressure indefinitely after the pumps stop and must be discharged in accordance with the manual.
- **Pneumatic** — bleed air, shop air, and charged bottles for emergency gear extension or brakes.
- **Stored mechanical energy** — springs, gas struts, counterbalances, tensioned control cables and wound-up trim systems.
- **Gravity** — anything raised: a landing gear that could retract, a control surface or door that could fall, a jacked aircraft, a suspended engine.
- **Thermal and chemical** — hot brakes, hot engine and bleed components, an oxygen system still pressurised, residual fuel vapour.

The discipline that makes lockout work is that the lock is *personal*. Each person working on the system fits their own padlock, to a multi-hasp if several people are involved, so the isolation physically cannot be removed until the last person has finished and removed their own lock. Nobody removes another person's lock; where that becomes unavoidable it is a documented management process, not a decision taken with bolt croppers at the end of a shift. The tag that accompanies the lock carries the information the next person needs: what is isolated, why, when it was applied, and who to contact. A tag on its own, without a lock, is a warning rather than an isolation — it relies entirely on the reader choosing to obey it.

Isolating an Aircraft Electrical System

Before work begins on an aircraft electrical system the relevant circuit breakers are pulled and collared or tagged, the battery is disconnected or its master switch selected off and secured, external ground power is disconnected, and warning placards are fitted at the flight-deck controls and at the isolation points so that nobody restores power to a circuit somebody else is inside. Simply selecting the battery master off does not satisfy this: it leaves individual circuits available the moment another engineer, quite reasonably, switches it back on to test an unrelated system.

One trap is worth committing to memory. A circuit breaker is both a protective device and a convenient isolation point, but not every load is behind one that the crew can reach: most transport aircraft have a **hot battery bus** that is permanently connected to the battery and feeds items such as entry lighting, the clock, fire-extinguisher circuits and some fuel shut-off valves regardless of the battery master switch position. Assuming a circuit is dead because a master switch is off, without consulting the wiring diagram to establish where that circuit is actually fed from, is a well-worn route to an arc, a burn or an inadvertent system operation. Consult the diagram, isolate at the source, and then prove dead.

Aircraft Batteries

An aircraft battery combines an electrical hazard, a chemical hazard and, while charging, an explosion hazard, and the two chemistries in service are handled differently.

- **Disconnection order.** Remove the earth (negative) lead first and reconnect it last. With the earth lead off, a spanner that slips and touches airframe structure while on the positive terminal completes no circuit; with the earth lead still on, the same slip is a dead short across the battery.
- **Hydrogen.** A lead-acid battery on charge, and an overcharged nickel-cadmium battery, gas hydrogen. Hydrogen is flammable in air over an exceptionally wide range — roughly 4% to 75% by volume — needs almost no energy to ignite, and being much lighter than air it collects at the highest point of the compartment or battery room. So: ventilate, no sparks or naked flames, and switch the charger off *before* disconnecting the leads so the break is not made in a hydrogen-rich atmosphere while current is flowing.
- **Electrolyte.** Lead-acid batteries contain sulphuric acid; nickel-cadmium batteries contain potassium hydroxide, a strong alkali. Each is neutralised by a different agent, and a trace of one contaminates the other permanently, so the two types are serviced in separate areas with separate tools, separate PPE and separate spill kits. Cadmium itself is toxic and a cell should never be heated, ground or cut.
- **Thermal runaway.** A nickel-cadmium battery on constant-voltage charge can enter a self-reinforcing loop: rising cell temperature lowers internal resistance, which increases the charging current, which raises the temperature further. Left unchecked the battery boils, vents and can catch fire. The defences are temperature sensing, a battery-charger fault warning, and treating any battery that is hot on removal as a hazard rather than a puzzle.
- **Handling.** Batteries are heavy, awkward and top-heavy in their trays; they are also short-circuit sources of enormous current, so the terminals are covered whenever the battery is out of the aircraft and it is never carried by its leads.

Static Electricity and Electrostatic-Sensitive Devices

Static presents two entirely separate hazards, and confusing them leads to the wrong control being applied.

The first is an **ignition** hazard. A charge separated by fuel flowing through a hose, by a synthetic coverall being pulled off, or by a plastic sheet being dragged across a wing can reach many thousands of volts, and the resulting spark carries more than enough energy to ignite a fuel/air mixture or anything at all in an oxygen-enriched atmosphere. It is why static precautions belong on the list of things that matter around fuel and around oxygen systems just as much as around avionics, and the fuelling case is dealt with in detail later in this note.

The second is a **damage** hazard to electrostatic-sensitive devices. Many electronic components can be damaged by discharges of less than 100 volts, and many disk-drive components by less than 10 volts, according to the ESD Association — yet the same association lists a worker at a bench typically generating

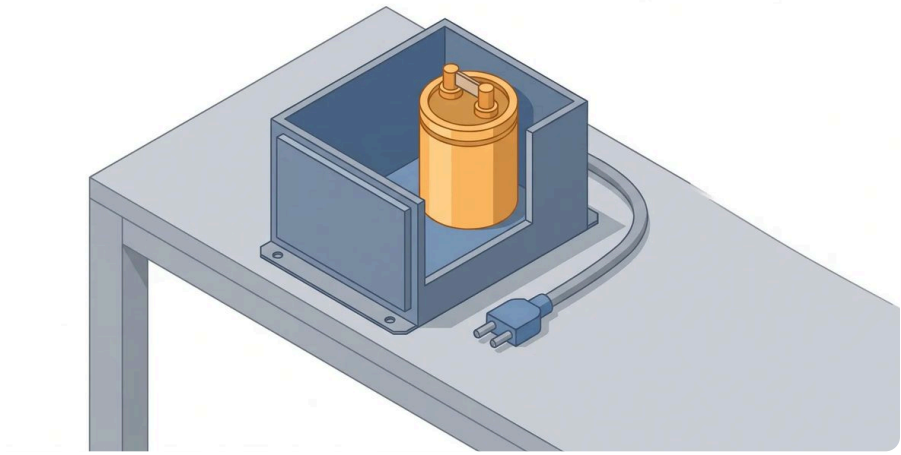
6,000 volts of static at 10 to 25% relative humidity, and still 100 volts at 65 to 90%. Worse, much of the damage is *latent* — the device passes its functional test, is fitted to the aircraft, and fails weeks later in service, at which point the cause is untraceable. The controls are therefore preventive and are followed whether or not anything seems to have gone wrong:

- Work at a designated ESD-protected bench: a dissipative mat, bonded to a common earth point, with the wrist strap, the mat and the operator all referenced to the same point so no potential difference can exist between them.
- Wear the wrist strap correctly, in contact with the skin, and test it — a strap with a broken cord or a dry contact is worse than none, because it produces confidence without protection. The series resistor it contains, of the order of one megohm, is what makes the strap safe for the wearer: it drains static in milliseconds while limiting the current that could flow if the wearer touched a live conductor.
- Transport and store sensitive items in the correct packaging. A pink or blue dissipative bag prevents charge generation but does not screen the contents from an external field; a metallised shielding bag does both. Boards are handled by their edges or their front panel, never by the connector pins or the components.
- Remember that low humidity dramatically increases charge generation and retention — a dry, heated hangar in winter is the worst case, and it is the same condition that makes handling avionics in a nylon coverall so effective at destroying them.

⚠ **Never confuse the two:** an anti-static wrist strap is protection for the *component*, not for the person. It deliberately connects the wearer to structure, which is the opposite of what electrical shock precautions require. It is fitted only for work on de-energised equipment, and it comes off before any work on a circuit that might be live.

Stored and Residual Energy

Removing the supply does not remove the energy. Capacitors in power supplies, filters and ignition exciters can hold a lethal charge for a long time after shutdown, and the higher the voltage the longer it takes to leak away. The correct way to discharge a high-voltage capacitor is through a suitable resistive discharge stick and then to fit a shorting link, leaving it fitted while the work proceeds. Shorting the terminals directly with a screwdriver is unacceptable: the peak current welds and pits the contacts, throws molten metal at the operator's face, and can leave a capacitor that partially recovers charge afterwards.



With the supply removed, a capacitor stays shorted by a link across its terminals while work proceeds.

Turbine ignition systems deserve individual mention because their exciter units store energy specifically in order to release it as a very high-energy spark, and that stored energy remains after the system is switched off. The manual states the waiting period and the discharge procedure before the unit or its lead may be disturbed, and an igniter lead is never left disconnected with the system live. Radar transmitters, HF transmitters and the high-voltage sections of older displays carry the same message: the interlocks and the exclusion zones referred to above are engineering controls, and defeating one to "just have a quick look" removes the only thing standing between the engineer and tens of kilovolts. The separate, non-electrical hazard of the radiation these transmitters emit is dealt with in the section on hazards around a live aircraft.

Gas Safety

Oxygen

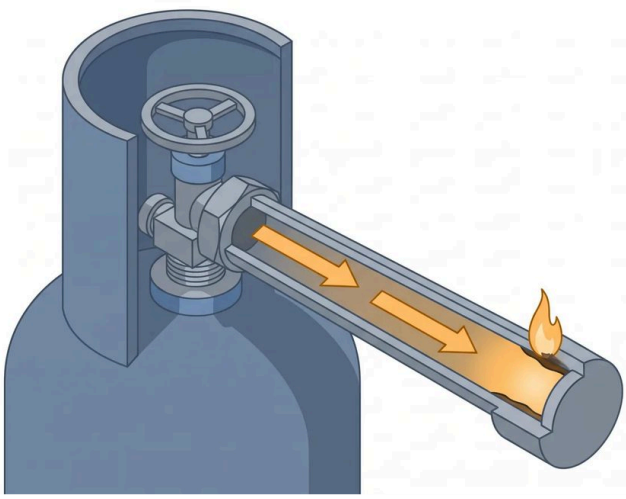
Oxygen is widely used in aircraft for crew and passenger breathing systems (typically stored at 1,800–2,200 psi / 124–152 bar). It is **not** flammable itself but is an extremely powerful **oxidiser** — it dramatically accelerates combustion.

- Keep all oil, grease, and hydrocarbons **away** from oxygen systems — even a fingerprint of grease on a fitting can ignite in pure O_2.
- Use only oxygen-clean tools and components (cleaned with approved solvents, sealed in protective bags).
- Never use PTFE tape on oxygen fittings — use only approved thread sealants.
- Store oxygen cylinders upright, secured, away from heat sources, and separated from fuel/oil storage by at least 6.1 metres (20 feet), or by a non-combustible barrier at least 1.5 m high with a 30-minute fire rating.
- Oxygen-enriched atmospheres (above 23.5%) are extremely fire-hazardous. Ventilate work areas when oxygen systems are open.

Two mechanisms explain every one of those rules, and knowing them is the difference between remembering a list and being able to work out the right answer for a situation the list does not cover.

Enrichment changes what "non-flammable" means. Combustion rate depends on the oxygen concentration at the reaction front. Raise it above the 21% of ordinary air and three things happen together: the energy needed to start ignition falls sharply, the flame temperature rises, and the burning rate increases. Materials that will not sustain a flame in air — steel wool, many fire-resistant fabrics, some fluoropolymers, aluminium in fine form — will burn fiercely in an oxygen-rich atmosphere, and a fire that starts in one spreads at a speed that gives no time to react. This is why the enrichment threshold quoted above is treated as an alarm level and not as an interesting statistic, and why an oxygen leak in a small compartment is a fire emergency even though nothing is burning yet.

The ignition source is usually the oxygen itself. A hydrocarbon film needs something to light it, and high-pressure oxygen provides that without any external spark. When a valve is opened quickly into a dead-ended line, the gas already in the line is compressed almost instantaneously by the incoming high-pressure gas. Because the compression is too fast for the heat to escape, the temperature of that trapped gas rises dramatically — the same adiabatic heating that makes a bicycle pump warm, but compressing to 1,800 psi rather than to a few tens of psi. Add the thinnest smear of oil, grease or even a fingerprint at that point and it ignites, and once it ignites in pure oxygen the fitting itself can burn. A second, related mechanism is particle-impact ignition: a fragment of swarf or Teflon tape carried at high velocity by the gas stream strikes a bend or a seat, and the impact energy does the same job.



Opening a valve fast compresses the gas trapped in a dead-ended line, heating it enough to ignite grease.

Those two mechanisms generate the practical rules directly. Open oxygen cylinder and system valves *slowly*, so that downstream pressure builds gradually and there is no shock compression. Stand clear of the regulator and gauge faces while doing it, in case a component fails. Never use oxygen as a substitute for compressed air — not to blow down a bench, not to cool yourself, and above all not to inflate a tyre or charge a strut, where

it would meet grease. And never allow the system to be opened up in an atmosphere that is not being ventilated.

Oxygen Cleanliness

"Oxygen clean" is a controlled condition, not a description of how the part looks. Components and tools are degreased with an approved solvent that leaves no residue, dried, inspected, and then sealed in protective bags which are only opened at the point of installation; anything that has been out of its bag on a general workbench has to be treated as contaminated again. Lint-free cloths are used because a thread of cotton is fuel. The only lubricant permitted anywhere near the system is the specific oxygen-compatible product named in the maintenance manual, applied in the quantity the manual states and nowhere else — ordinary greases are hydrocarbons and are exactly the fuel described above.

The prohibition on PTFE thread tape has two reasons behind it. Shredded tape can migrate into the system, block a small orifice or a regulator seat, and provide the particle for an impact ignition; and the tape has no sealing function on the tapered metal-to-metal and coned fittings normally used in oxygen systems anyway, so applying it is both hazardous and pointless. Where a thread sealant is required, only the approved oxygen-service compound is used.

Residual Pressure and Charging

An oxygen cylinder is never emptied. A minimum residual charge is always retained, and the reason is that the positive internal pressure keeps the outside world out. The figure itself comes from the maintenance manual for the particular cylinder and system; around 500 psi is commonly quoted. Two things go wrong if the pressure falls below it:

- **Moisture enters the cylinder.** Ambient air carries water vapour, and once inside a steel or aluminium cylinder that moisture condenses and starts internal corrosion, which both weakens the pressure vessel and contaminates the gas with rust particles.
- **Moisture reaches the regulator.** Gas expanding through a regulator or a metering orifice cools sharply. Water that has been drawn into the system therefore freezes exactly where the passages are narrowest, and the regulator blocks — potentially at the moment the crew need the system. Contamination carried in with the moisture can jam the mechanism in the same way.

A cylinder or system found below the stated residual is therefore not simply topped up. It is treated as contaminated, and the maintenance manual's procedure for purging, drying and, where required, returning the cylinder for internal inspection applies.

Charging is done slowly for a reason that catches people out. Compressing gas into a cylinder heats it, and hot gas exerts more pressure than the same quantity of cold gas. Fill quickly and the gauge shows the target

pressure while the contents are still warm; as the cylinder cools to hangar temperature the pressure falls and the cylinder turns out to be undercharged. Slow filling, or filling to a temperature-corrected figure from the chart in the manual, avoids handing the aircraft back with less oxygen on board than the paperwork says.

> ✈ **Exam point:** the hazard of oxygen is not that it burns — it does not. It is that it makes everything else burn faster, hotter and more easily, and that pressurised oxygen meeting a hydrocarbon can supply its own ignition. An answer describing oxygen as "flammable" or "toxic to breathe" is wrong on both counts; the correct description is a powerful oxidiser that vigorously supports combustion.

Nitrogen

Nitrogen is used for tyre inflation, strut charging, and as an inert gas blanket in fuel tanks (Nitrogen Inerting System — OBIGGS). It is an **asphyxiant** — it displaces oxygen in confined spaces without warning (it is odourless and colourless).

- Never enter a nitrogen-purged space without self-contained breathing apparatus (SCBA).
- Monitor oxygen levels in confined spaces — 19.5% is the action threshold, not the point of collapse: below it the atmosphere is classed as oxygen-deficient and must not be entered without breathing apparatus. The immediately-dangerous-to-life condition is nearer 16%, where mental effectiveness, vision and muscular coordination are already measurably impaired.
- Nitrogen cylinders are pressurised to 200–300 bar. Handle with care, secure during transport.

The reason an asphyxiant gives no warning is physiological rather than chemical. The body's urge to breathe is driven almost entirely by the carbon dioxide concentration in the blood, not by the oxygen concentration. In a nitrogen-rich atmosphere the casualty continues to exhale carbon dioxide normally, so no breathlessness, no panic and no distress is felt — the first symptoms are impaired judgement and clumsiness, which the casualty is by definition in no state to recognise. In a severely depleted atmosphere, unconsciousness can follow within a couple of breaths and it happens without any struggle. That is the whole argument for measuring the atmosphere instead of relying on how it feels, and it is why the oxygen concentration limits given above are treated as hard gates rather than guidance.

The margin between the action threshold and the point of obvious harm is deliberate: anyone who has already been affected is unlikely to make a good decision about leaving.



A colleague has collapsed inside: the person outside stays out and raises the alarm.

⚠ **The rescuer is the second casualty:** a large proportion of deaths in oxygen-deficient spaces are would-be rescuers who went in without breathing apparatus to help a colleague who had collapsed. The collapsed person is evidence that the atmosphere is lethal. Raise the alarm, use the retrieval line, and let equipped rescuers enter.

Why Nitrogen Is Used for Tyres, Struts and Fuel Tanks

Nitrogen is chosen for these jobs for two properties: it is dry, and it is inert.

- **Tyres.** A heavily braked wheel gets extremely hot. If the gas inside the tyre contains oxygen, the hot rubber liner can decompose and the resulting volatile products can ignite inside the tyre — a mechanism that has destroyed wheels and killed people on the ramp. Filling with nitrogen removes the oxygen and the reaction cannot occur, and airworthiness requirements for large aeroplanes cap the oxygen content of the inflation gas for exactly this reason. Dryness matters too: moist air condenses and freezes at altitude, and water inside a magnesium or aluminium wheel promotes corrosion. What nitrogen does *not* do is behave differently from air as the temperature changes — the pressure rise with temperature is essentially the same, so that is not the reason for using it.
- **Oleo struts and accumulators.** The gas is in direct contact with hydraulic fluid at high pressure and temperature. Oxygen would accelerate oxidation of the fluid and, in an accumulator, would sit in contact with hot oil at full system pressure. Charging a strut or accumulator with oxygen or shop air instead of nitrogen is one of the most serious errors possible on a ramp, and it is made possible only by using the wrong trolley — which is why the couplings, the trolleys and the cylinder markings are all different.
- **Fuel tank inerting.** The on-board inert gas generating system referred to above separates nitrogen-enriched air from bleed air and feeds it into the tank ullage, holding the oxygen concentration below the level at which the fuel/air mixture can propagate a flame. For maintenance this has a direct consequence: the ullage of an inerted tank is an asphyxiating atmosphere as well as a flammable-vapour one, and a

nitrogen purge used to make a tank safe from fire has made it lethal to enter. A tank purged with nitrogen must afterwards be purged with air and re-tested before anyone goes in.

Nitrogen is also used as a service gas for pressure testing and for blanketing components in storage, and the same rule applies every time: any space that has been filled or purged with it is an oxygen-deficient space until it has been ventilated and measured.

Compressed Air and Hydraulic Gases

- Pneumatic systems operate at up to 3,000 psi (207 bar). Never direct compressed air at the body — it can cause air embolism.
- Hydraulic accumulators are charged with nitrogen — depressurise before removing.
- Bleed all pressure from lines before disconnecting fittings.

Why Compressed Air Is Not Harmless

Shop air is treated so casually that its hazards are worth stating plainly. An air jet directed at skin can force air through a pore or a small wound into the bloodstream, producing an embolism that can be fatal; directed at the ear it can rupture the eardrum; directed at clothing it can inflate and rupture the bowel. It also picks up whatever it lands on and turns it into a projectile, which is why swarf blown off a bench ends up in somebody's eye across the workshop.

- Air is never directed at any person, including yourself, and is never used to blow dust off clothing or hair — a brush or a vacuum does that job.
- Where air is used for cleaning it is reduced in pressure — workshop rules commonly cap cleaning air at about 30 psi, roughly 2 bar — and used with a nozzle designed to bleed off if it is blocked, with eye protection and with screening to protect other people.
- Couplings are fitted with safety devices that vent before the connection releases, and hoses over any distance are fitted with whip-check restraints, because a hose that comes off under pressure flails hard enough to cause serious injury.
- Line pressure is released before any fitting is broken. The energy stored in a charged line does not care that the compressor has been switched off, and a fitting released under pressure becomes a projectile while the hose behind it whips.
- Receivers and line filters are drained of condensate on schedule; water and carried-over compressor oil in the air line contaminate paint, blast media and pneumatic tools, and oil in an air line supplying anything shared with an oxygen system is a serious hazard in its own right.

✈ **Aviation Context:** Gas cylinders are colour-coded. In Europe, EN 1089-3 colours the cylinder shoulder: white = Oxygen, black = Nitrogen, grey = Carbon Dioxide. The US CGA scheme differs — green = Oxygen, black = Nitrogen. Always verify by reading the cylinder label — never rely on colour alone.

Cylinder Identification: Two Schemes, One Reliable Answer

There is a good reason the rule is to read the label. Europe changed its cylinder colour code, and the two schemes disagree with each other about several common gases — so a hangar can easily hold cylinders marked to both. Under the current European standard the *shoulder* carries the identification and the body colour has no meaning at all; under the superseded British scheme the whole *body* was coloured. Both are shown below because both are still met, on older cylinders and in older examination questions.

Gas	EN 1089-3 shoulder colour (current European)	Superseded British body colour
Oxygen	White	—
Nitrogen	Black	Grey body, black neck
Carbon dioxide	Grey	Black
Compressed air (industrial)	Bright green	Grey
Acetylene	Maroon	Maroon
Argon	Dark green	Blue
Helium	Brown	Brown
Hydrogen	Red	Red
Breathing air	White and black quartered	—

Read across the carbon dioxide and air rows and the problem is obvious: grey means carbon dioxide under one scheme and air under the other, and black means nitrogen under one and carbon dioxide under the other. Colour narrows the field; it never settles it. What does settle it is the permanently attached label, which gives the gas name, the hazard pictograms and the composition, and the markings stamped into the cylinder shoulder, which give the specification, the working and test pressures, the tare weight, the date of the last hydrostatic test and the date the next one is due. Those stamped markings are also the reason a cylinder outside test date is quarantined rather than used.

Handling, Storage and Transport

A full cylinder is a pressure vessel with a fragile brass valve on top, and virtually every serious cylinder accident starts with that valve. If a cylinder falls and the valve shears off, the escaping gas turns the cylinder into an unguided projectile capable of going through a hangar wall.

- The valve protection cap stays fitted whenever the cylinder is not connected in service, and the cylinder is never lifted, dragged or manoeuvred by its valve, its cap or its regulator.
- Cylinders are moved on a proper trolley with the retaining chain fastened, not carried, rolled along the floor or "walked" on their base across a hangar. Where a crane must be used, it is with a purpose-made cradle or cage — never a sling round the neck and never a magnet.
- They are secured upright at all times, in use and in store, by chain or bracket at about two-thirds of their height.
- Full and empty cylinders are stored separately and labelled, so that an empty is never connected in the belief that it is a spare, and stock is rotated so that older cylinders are used first.
- Stores are outdoors or well ventilated at both high and low level, because a leak of a light gas collects at the ceiling and a leak of a heavy one collects at the floor.
- Oxidising gases are separated from fuel gases and from combustible materials as described in the oxygen storage rule above. Cylinders are kept away from heat sources and out of direct sunlight, and never in a stairwell, an escape route or an unventilated cupboard.
- Regulators are matched to the gas and the cylinder, checked for damage before fitting, and never forced onto a thread that does not accept them — the thread forms and left/right-hand conventions differ by gas precisely so that a fuel-gas regulator cannot be fitted to an oxygen cylinder.

Worked example — why heat matters.
A cylinder is a fixed volume, so its pressure rises in direct proportion to its *absolute* temperature:

$$P_2 = P_1 \times \frac{T_2}{T_1}$$

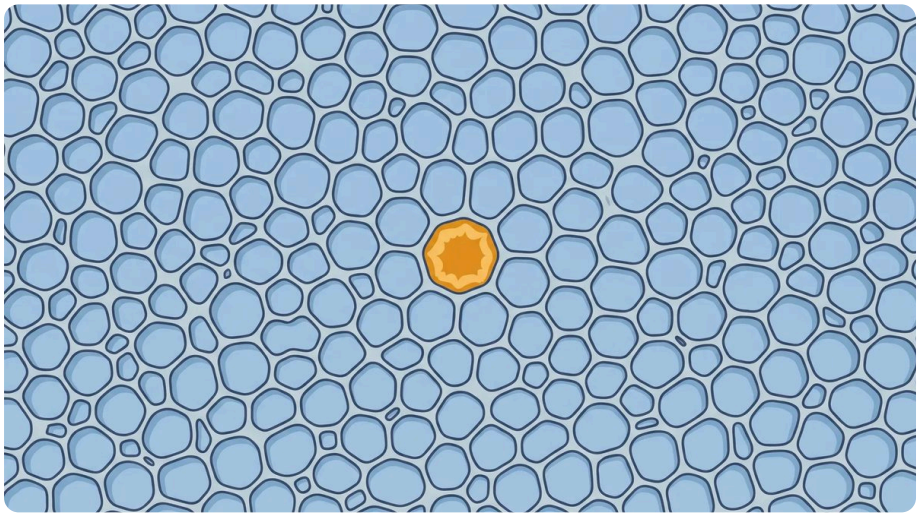
Take a nitrogen cylinder charged to 200 bar in a store at 15 °C, which is 288 K, and stand it in the sun outside a hangar until it reaches 40 °C, which is 313 K:

$$P_2 = 200 \times \frac{313}{288} = 217 \text{ bar}$$

That is a 17 bar rise for a very ordinary summer afternoon, on top of a charging pressure that may already be near the cylinder's rated maximum. Now consider the same cylinder near a fire, where the temperature rise is measured in hundreds of degrees rather than tens: the pressure relief device is the only thing preventing a burst, and if it is painted over, obstructed or facing the wrong way there is nothing. This is also why a cylinder involved in a fire is cooled with water from behind cover and why the area is evacuated, rather than approached.

Acetylene: The Cylinder That Is Not Simply a Bottle of Gas

Acetylene is the one common workshop gas that is dangerous on its own, without any oxidiser present. Above a modest pressure — regulations in the United States cap its use at 15 psi gauge, and other authorities set a similar low limit — free acetylene can decompose explosively into carbon and hydrogen, releasing a large amount of energy with no air involved at all. It therefore cannot be stored as a simple compressed gas.



Inside an acetylene cylinder, a porous mass soaked in acetone splits the gas into countless small cells.

- **Construction.** An acetylene cylinder is filled with a porous mass saturated with acetone, and the acetylene is dissolved in the acetone. The porous mass divides the interior into countless small cells so that a decomposition cannot propagate through the cylinder.
- **Always upright.** Lay a cylinder down and liquid acetone can reach the valve, so that opening it discharges acetone into the regulator and hoses and leaves the cylinder depleted. A cylinder that has been horizontal must stand upright for the period the supplier specifies before it is used.
- **Never copper or silver.** Acetylene reacts with copper and with silver to form metallic acetylides, which are unstable compounds detonated by shock or friction. Fittings, pipework and regulator internals in acetylene service therefore avoid copper and high-copper alloys, and silver-bearing brazing alloys are not used in the gas path. Tin, by contrast, forms no such compound and presents no equivalent hazard.
- **Flashback protection.** Flashback arrestors and non-return valves are fitted in both the fuel-gas and the oxygen lines, so that a flame travelling back up a hose is stopped before it reaches a cylinder.
- **Leaks.** Acetylene is detected by its characteristic garlic-like smell and confirmed with an approved leak-detection fluid — never with a flame. Its flammable range in air is exceptionally wide, so a leak in an enclosed space produces an ignitable mixture almost immediately.
- **Fire.** A cylinder that has been heated in a fire can continue to decompose internally for hours after the flames are out. The area is evacuated, the cylinder is cooled with copious water from behind protection, and it is not approached, moved or declared safe until the fire service says so.

Oils, Fuels, and Chemical Safety

Aviation Fuels

Fuel	Type	Flash Point	Colour Code
Jet A-1	Kerosene	38°C min	Black lettering on white
Avgas 100LL	Gasoline	–40°C	Red lettering on white

- **Avgas vapour** is heavier than air and can accumulate in pits, tanks, and low areas — creating explosive atmospheres.
- **Jet fuel** (kerosene) is less volatile but still flammable above its flash point. Fuel-soaked clothing is extremely dangerous.
- Prevent static discharge during refuelling — always bond the fuel bowser to the aircraft before connecting hoses.
- No smoking, naked flames, or spark-producing equipment within **15 metres** (50 feet) of fuelling operations or fuel storage.
- Prolonged skin contact with aviation fuels causes dermatitis — use barrier cream and chemical-resistant gloves.

Flash Point, Flammable Range and Autoignition

A liquid does not burn. Its *vapour* burns, mixed with air in the right proportion, so three separate properties decide how a fuel behaves in a hangar and they are routinely confused with one another.

- **Flash point** is the lowest liquid temperature at which the fuel gives off enough vapour to form an ignitable mixture at its surface. It is a property of the liquid's volatility and it says nothing about how hot the ignition source must be.
- **Flammable range** is the band of vapour concentrations that will actually propagate a flame, bounded by the lower explosive limit (below it the mixture is too lean) and the upper explosive limit (above it too rich). Both ends matter: a mixture can be too rich to burn and become dangerous simply by being diluted with fresh air.
- **Autoignition temperature** is the surface temperature at which the vapour ignites with no spark or flame at all — a bleed duct, an exhaust, a brake unit or an unlagged hot pipe.

Put the two fuels in the table above through those definitions and their hazards turn out to be almost mirror images. Avgas, with a flash point far below any temperature the hangar will ever see, is always producing

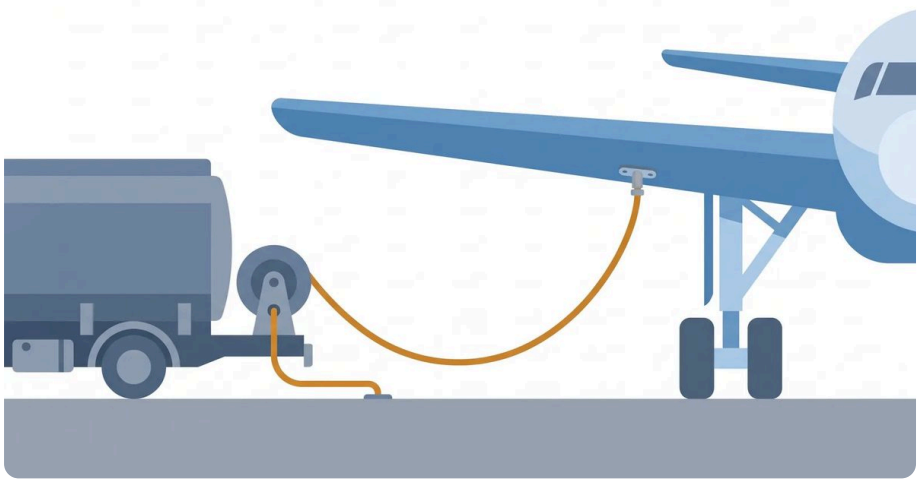
vapour; the space above the liquid in a partly filled Avgas tank is normally *too rich* to ignite, and the danger lies outside the tank — at the vent, in the drip tray, in the pit or bilge where the heavier-than-air vapour has collected and mixed down into its flammable range. Jet A-1, with a minimum flash point of 38 °C, is normally *too lean* above the liquid at ordinary temperatures, which is why kerosene feels so much safer to handle. The two ways that safety disappears are heat and mist: fuel warmed by the sun, by a bleed leak or by equipment running below the tank moves the ullage up into the flammable band, and kerosene atomised into a spray by a leaking high-pressure line, or absorbed into a rag, presents so much surface area that it ignites readily well below its flash point.

Autoignition reverses the ranking again. Kerosene has by far the higher flash point of the two, but its autoignition temperature is substantially the lower — so it is the kerosene, not the Avgas, that will ignite on contact with a hot bleed duct with no spark present. "Higher flash point" therefore does not mean "harder to ignite" in every circumstance, only in the specific circumstance of an open surface at ambient temperature.

Static, Bonding and Earthing During Fuelling

Fuel flowing along a hose, and especially fuel passing through a filter/water separator, strips electrons from the surfaces it touches and carries a charge into the receiving tank. The filter element is by a wide margin the biggest charge generator in the system. That charge accumulates on the fuel surface inside the tank, and if it reaches a high enough potential relative to the surrounding structure it will discharge as a spark — inside the one place in the aircraft where a flammable vapour is guaranteed to be present.

Three defences are used together. **Bonding** connects the bowser, the hose nozzle and the aircraft so that they all sit at the same potential and no spark can jump between them; it is connected *before* the filler cap is opened or the hose coupled, and removed *last*, after the cap is back on, because the moment of connection and the moment of disconnection are when a difference in potential is most likely to exist. **Earthing** to a ground point additionally drains charge to earth where the procedure calls for it. And a **static dissipator additive** blended into jet fuel raises the fuel's electrical conductivity so that charge separated in the hose relaxes back through the fuel in a fraction of a second instead of accumulating.



A bonding cable joins bowser and aircraft at one potential, and an earth lead drains charge to ground.

The remaining precautions follow the same logic: no switching of electrical equipment, no radio transmission and no mobile telephones in the fuelling zone; the exclusion distance for ignition sources given above; a fire extinguisher positioned and manned; the bowser parked so it can be driven away rather than having to reverse past the aircraft; and drip trays and a spill kit to hand. Filling through an overwing nozzle demands particular care because the operator, the nozzle and the fuel are all in the open: the nozzle's bonding clip goes to the designated point on the filler adapter before the cap comes off, and the nozzle stays in contact with the filler neck throughout.

Grade Markings, Fuel Dyes and Lead

Two different colour conventions apply to aviation fuel, and they are easily confused. The colour code in the table above is the **equipment marking** convention — the lettering on bowsers, hydrant couplings, filler-point placards and hose decals, which is how the correct product is identified before a nozzle goes anywhere near an aircraft. Separately, the **fuel itself is dyed** so a sample drawn into a jar can be recognised: Avgas 100LL is blue, Avgas 100 is green, and Jet A-1 carries no dye and appears clear to straw-coloured. A clear sample from a piston aircraft's tank is therefore a reason to stop, not a sign of clean fuel — it may be jet fuel or it may be water.

The "LL" in 100LL stands for low lead, not no lead: the fuel still contains tetraethyl lead as an anti-knock additive. Lead compounds are cumulative toxins absorbed through the skin as well as by inhalation, so Avgas is never used as a cleaning solvent or a degreaser, gloves are worn when handling it and when draining sumps, and hands are washed before eating. The same rule applies to the sludge drained from a sump or a filter bowl, which concentrates whatever the fuel was carrying.

Two further contamination hazards belong with fuel handling. **Water** collects in tank low points by condensation and by settling out of the fuel; besides the icing and engine-stoppage risk it enables **microbiological growth** at the fuel/water interface, and the resulting sludge blocks filters, holds water against

the tank skin and causes serious corrosion. Sump draining, sampling and the handling of biocide treatments are therefore routine maintenance tasks, and the biocide itself is a hazardous substance requiring its own precautions.

Finally, the reason fuel-soaked clothing is singled out in the list above is worth spelling out, because it is a mechanism rather than a warning. Fabric behaves as a wick: it holds the fuel spread thinly over an enormous surface area, in intimate contact with air, at exactly the geometry that makes a candle burn steadily. A splash of kerosene on a concrete floor may not sustain a flame; the same quantity in a coverall will burn, on a person, until it is out. Contaminated clothing comes off at once and is not left to "dry off", and the skin underneath is washed — prolonged contact with fuel held against the skin by wet fabric causes chemical burns without any fire at all.

Hydraulic Fluids

Aircraft hydraulic systems commonly use:

- **MIL-PRF-5606** — mineral-oil based (red). Petroleum-based, flammable.
- **MIL-PRF-83282** — fire-resistant synthetic hydrocarbon base (red). A drop-in replacement for 5606 with a far higher flash point.
- **Skydrol (Type IV/V)** — phosphate-ester-based (purple/dark amber). Fire-resistant but **highly corrosive** — attacks skin, eyes, and many plastics/paints.

⚠ **Skydrol Warning:** Phosphate-ester hydraulic fluid (Skydrol) causes severe eye damage and skin irritation. Wear splash-proof goggles and butyl rubber gloves. If skin contact occurs, wash immediately. If eye contact occurs, irrigate for at least 15 minutes and seek medical attention. Skydrol also attacks many common rubbers, paints, and wire insulation — use only approved seals and materials.

Three Fluid Families, and Why They Are Never Mixed

The fluids listed above fall into three chemical families, and the distinction that matters on the hangar floor is not which is "better" but that each one is matched to the seals, hoses and paint finishes of the system it serves. Mineral oil and the synthetic hydrocarbon that replaced it are compatible with nitrile seals; phosphate esters are not, and require butyl or ethylene-propylene elastomers instead. The incompatibility runs in both directions and neither direction is recoverable in service:

- Put a mineral-based fluid into a phosphate-ester system and it swells and softens every butyl seal in it.
- Put a phosphate ester into a mineral-oil system and it attacks every nitrile seal in it.

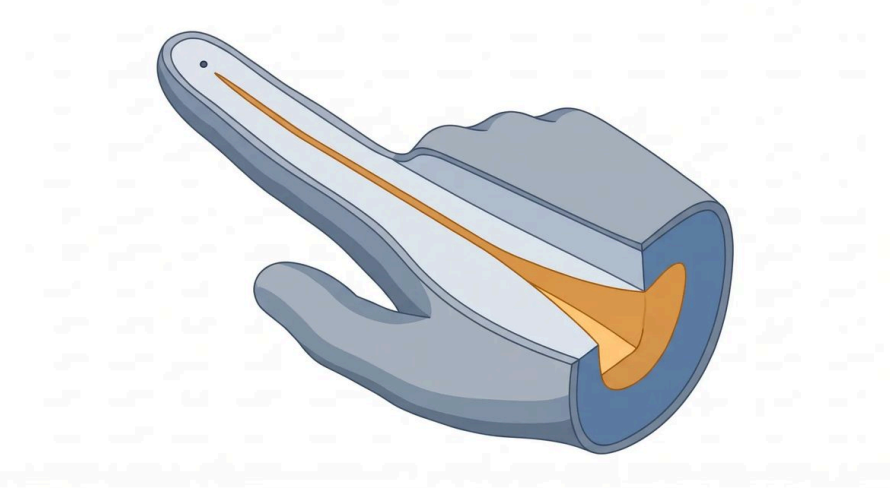
In either case the consequence is not a single leaking joint discovered next week; it is progressive failure of every seal in the system, which means a complete drain, flush and seal replacement, and it is a reportable occurrence rather than an embarrassment to be topped up quietly. The safeguards against it are dedicated servicing trolleys and cans marked for one fluid only, and reading the placard at the reservoir filler rather than relying on memory or on the colour of what is already in the sight glass.

Colour helps but does not decide. Both the mineral fluid and its fire-resistant synthetic hydrocarbon replacement are dyed red, so red tells you which family you are in but not which specification; the phosphate esters are purple. The two red fluids are close relatives — the synthetic hydrocarbon was developed specifically as a drop-in replacement with a much higher flash point, which is why a fine spray of it from a burst line will not ignite as readily. Its one significant limitation is that it is more viscous at very low temperature, which is why a further synthetic hydrocarbon specification exists for aircraft operating in extreme cold. Which fluid a given aircraft takes is stated in the maintenance manual and on the placard, and an aircraft is never converted from one to the other on the basis that they are "both red".

The material attack described in the warning above has a maintenance consequence beyond personal protection. A phosphate-ester leak that runs down inside a structure will lift paint, soften PVC wire insulation and degrade some transparent plastics, so a persistent seep above a wiring loom is a potential electrical fault in the making and is cleaned up and rectified rather than monitored. Spills are wiped up immediately with the approved absorbent and the area is cleaned according to the manual, not left to evaporate — it does not evaporate.

High-Pressure Fluid Injection Injury

Transport aircraft hydraulic systems run at 3,000 psi, and several modern types at 5,000 psi. A pinhole in a pipe, a hose or a seal at that pressure produces a jet fine enough to be almost invisible and fast enough to pass straight through skin without the person necessarily feeling more than a sting. This is one of the most under-rated injuries in aircraft maintenance, and it is dangerous precisely because it looks trivial:



Outside, a pinprick mark; inside, injected fluid spreading along the finger and into the palm.

- The entry wound may be a mark the size of a pinprick, with little bleeding and little immediate pain, so the casualty carries on working.
- The injected fluid tracks along the tendon sheaths and tissue planes, spreading far beyond the entry point — a finger injection can reach the palm and the forearm.
- Swelling inside the closed compartments of a finger or hand cuts off the blood supply, and the fluid itself is chemically toxic to tissue. Phosphate ester and the additive packages in modern fluids are particularly aggressive.
- Within hours the tissue begins to die. Treated late, the outcome is frequently amputation; treated early, the hand is usually saved.

⚠ **Never search for a hydraulic leak with your hand.** Depressurise the system first; if a leak must be located under pressure, pass a piece of stiff card or use a mirror, keeping every part of the body out of the jet. If an injection injury does occur it is a **surgical emergency**, not a first-aid matter: go straight to hospital, say explicitly that it is a high-pressure injection injury and not a simple puncture, and take the fluid's safety data sheet with you so the surgeon knows what has been injected. Do not wait to see whether it gets worse.

The same stored energy that makes the jet dangerous makes a disconnection dangerous. A hydraulic system holds pressure long after the pumps stop, because accumulators are designed to store it — that is their purpose. So before a joint is broken the system is depressurised through the correct valve or ground connection, the accumulator is discharged to the figure the manual gives, and the reservoir is checked to establish where the fluid will go. Control surfaces, gear doors and flaps can move as a system depressurises, so the area is cleared and the appropriate ground locks and safety pins are in place before anyone starts.

Solvents and Cleaning Agents

- **MEK (Methyl Ethyl Ketone)** — highly flammable, CNS depressant. Use only in ventilated areas.
- **Acetone** — extremely flammable, rapid evaporation, drying to skin.
- **Isopropyl Alcohol (IPA)** — commonly used for cleaning electrical connectors. Flammable.
- **Chromate compounds** — used in some primers and conversion coatings. *Hexavalent chromium (Cr^{6+})* is carcinogenic. Wear respiratory protection when sanding/spraying chromate products.
- **Two-part sealants and adhesives** — catalysts (hardeners) are often sensitisers causing allergic reactions.

How Chemicals Get In, and the Difference Between Irritation and Sensitisation

A substance can only harm you if it reaches you, and there are only four ways in: **inhalation** of vapour, mist, dust or fume; **absorption** through intact skin or through the eyes; **ingestion**, almost always indirectly by eating, drinking or smoking with contaminated hands; and **injection** through a puncture, a cut or a high-pressure jet. Identifying the route is what selects the control: extraction and respiratory protection defeat inhalation, gloves and coveralls defeat absorption, hand-washing discipline defeats ingestion, and nothing defeats injection except keeping the body out of the line of a pressurised jet.

Effects divide into **local** — damage where the substance lands, such as a solvent defatting the hands or an acid burning the eye — and **systemic**, where the substance is carried around the body and does its damage elsewhere, as lead does to the nervous system and cadmium does to the kidneys. Systemic effects are the ones that accumulate silently over a career, and they are the reason exposure is controlled long before anybody feels unwell.

The early signs of solvent over-exposure deserve to be recognised, because they are routinely mistaken for ordinary tiredness: headache, dizziness, a feeling of mild intoxication, poor coordination and slowed reactions. An engineer in that state is not merely uncomfortable — they are a hazard on an access stand, on a live circuit and behind the wheel afterwards. Ventilation restored and fresh air taken at the first symptom, rather than at the last, is the correct response.

A distinction worth getting right is between an irritant and a sensitiser, and it is a genuine difference in kind:

- An **irritant** causes harm in proportion to the dose. Keep the exposure low enough and the effect is small; remove the exposure and the skin or airway generally recovers. Most solvents behave this way.
- A **sensitiser** provokes the immune system. The first exposures may produce no reaction at all. Once a person is sensitised, however, an exposure far below any exposure limit — a concentration that affects nobody else in the workshop — triggers dermatitis or asthma, and the sensitivity is usually permanent. The

isocyanate hardeners in two-pack polyurethane paint, the amine hardeners in epoxy resin and adhesive systems, and chromate compounds are the sensitisers an aircraft engineer meets most often.

That is why the catalyst side of a two-part product is treated with more respect than the resin side, and why respiratory sensitisation is career-ending in a way that a solvent headache is not: a sensitised sprayer cannot go back into the paint shop, whatever protection is offered. Exposure limits are published as an eight-hour time-weighted average and, for fast-acting substances, as a fifteen-minute short-term limit; both assume the extraction is running and are not a licence to work up to the number.

Skin Care, Barrier Creams and Dermatitis

Occupational dermatitis is one of the commonest reportable ill-health conditions in engineering trades and its mechanism is simple. The outer layer of skin is waterproofed by a thin film of natural lipids. Solvents dissolve that film. Repeatedly degreasing the hands — and washing them in thinners, MEK or Avgas does it far more effectively than any hand cleanser — strips the lipid layer faster than the body can rebuild it, and the skin becomes dry, red, itchy and cracked. Cracked skin is then an open door: chemicals that would never pass through intact skin get in, and so do bacteria.

Three products are involved and they do three different jobs:

- **Barrier or pre-work cream** is applied to clean, dry skin *before* work begins. It makes contamination easier to remove afterwards and gives a little protection against mild irritants. It is not a glove, it does not restore anything, and it is not a substitute for the correct protective glove.
- **The correct cleanser** removes contamination after work. A proprietary hand cleanser, not a solvent, and warm water — grit-loaded cleaners used habitually cause their own damage.
- **A restorative preparation applied afterwards** replaces the natural oils that the work has removed. This is where the traditional workshop answer for resin and solvent work belongs: after cleaning off the residue, an acetone and lanolin preparation returns oil to the skin, lanolin being the fatty component that does the restoring.

Alongside the creams sit the habits: gloves selected for the chemical and changed when contaminated, no solvent-soaked rag stuffed in a pocket against the thigh, hands washed before eating and before using the lavatory, and a regular look at the hands for early redness or cracking. Dermatitis caught in its first weeks resolves; left for a year it can become chronic.

Dust, Fume and Fibre

Several of the worst exposures in a hangar are solids rather than liquids, and they are generated by the work rather than poured from a tin — which means the safety data sheet for the product as supplied may say

nothing at all about them.

- **Composite dust.** Cutting, drilling, sanding or grinding cured carbon or glass fibre produces a fine dust of sharp fibre fragments that irritates skin, eyes and airways. Carbon dust has a second property that catches people out: it is electrically conductive, so it settles into connectors, relays and open equipment and causes tracking and intermittent faults far from where the work was done. Control it at source with local exhaust ventilation and on-tool extraction, protect adjacent equipment, wear respiratory protection of the appropriate class along with gloves, eye protection and covered skin, and clean up with a suitable vacuum — never by dry sweeping and never with a compressed-air line, both of which simply put the dust back into the air everyone is breathing.
- **Uncured resins and hardeners.** These are the sensitisers described above, and they are absorbed through the skin. Once fully cured the matrix is far less hazardous; partially cured swarf and dust from an incomplete cure is not.
- **Asbestos.** Older aircraft may contain asbestos in brake and clutch friction materials, in gaskets and packings, in thermal and acoustic insulation and around firewalls. Undisturbed it is not a risk; drilling, grinding, scraping or blowing it out with an airline releases fibres that lodge permanently in the lung and cause asbestosis, lung cancer and mesothelioma decades later. If a material is suspected of containing asbestos the work stops until it has been identified. An ordinary nuisance-dust mask offers no useful protection against asbestos fibres, and removal is a job for licensed specialists working under controlled conditions, not for the shift.
- **Cadmium.** Many aircraft fasteners and steel parts are cadmium plated. Cadmium fume is severely toxic to the lungs and kidneys and it is produced by any process that heats the plating — welding, brazing, flame cutting or aggressive grinding of a cadmium-plated part. The rule is simply not to apply heat to cadmium-plated components, and to treat the dust from grinding them as a hazardous substance.
- **Chromates.** The carcinogenic hexavalent chromium referred to above is released as dust when chromate primer is sanded and as mist when it is sprayed; both need extraction at source in addition to respiratory protection.

The control order here is the hierarchy in the first section applied to airborne contaminants: substitute the product if a safer one is approved, extract at the point of generation so the contaminant never enters the room, ventilate the room generally as a second line, and use respiratory protection last — because the mask protects only the person wearing it, while the extraction protects everybody in the bay.

Safety Data Sheets (SDS / MSDS)

Every chemical substance used in the workshop must have an SDS available. The SDS contains 16 sections covering identification, hazards, composition, first aid, fire-fighting measures, handling, exposure controls, physical properties, stability, toxicology, disposal, transport, and regulatory information. Always consult the SDS before using an unfamiliar substance.

Those sections appear in the same fixed order on every data sheet in the world, which is what makes the document usable under pressure: you do not have to read it, you have to know where to look. Identification, hazards identification and composition are sections 1 to 3; first aid is section 4 and fire-fighting measures section 5, deliberately near the front so that the two things needed in an emergency are the two you reach first. Further down, accidental release measures tell you how to deal with a spill, handling and storage gives the compatibility and segregation rules, exposure controls and personal protection names the actual glove material and filter class to use rather than "suitable PPE", physical and chemical properties gives the flash point and vapour density, stability and reactivity lists what it must not be mixed with, and disposal gives the waste route.

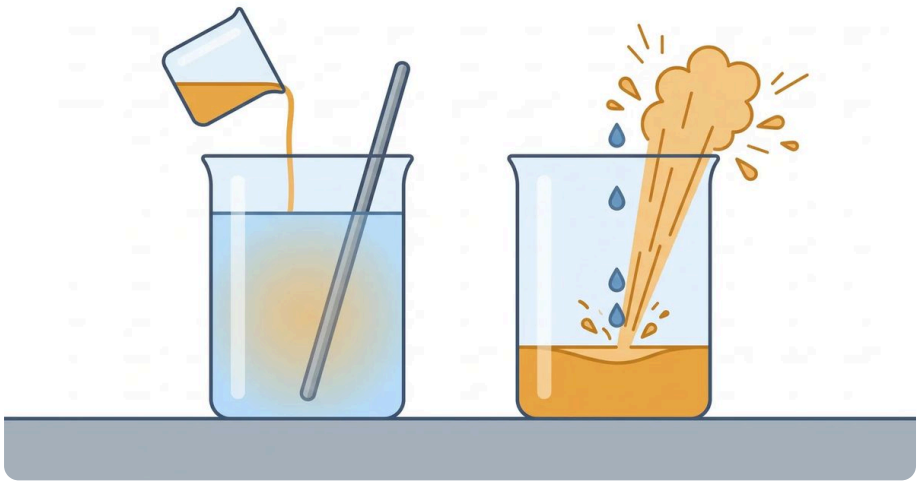
Two limitations are worth understanding. First, the data sheet describes the substance **as supplied**. It will not describe the dust produced by sanding the cured product, the fume produced by heating it, or what happens when it is mixed with the other substance already on the component — all of which are your assessment to make. Second, the data sheet is information, not a control: the legal duty is to carry out an assessment of the substance *in the way your organisation uses it*, and to put controls in place. A file of data sheets in the office satisfies nothing on its own.

The container label carries the same information in compressed form. The internationally harmonised hazard pictograms are red-bordered diamonds with a black symbol, and nine of them cover everything: a flame for flammable, a flame over a circle for oxidising, an exploding bomb for explosive or self-reactive, a gas cylinder for gases under pressure, a corrosion symbol for substances that attack skin, eyes or metal, a skull and crossbones for acute toxicity, an exclamation mark for harmful or irritant, a "health hazard" torso symbol for the serious long-term effects such as carcinogenicity and respiratory sensitisation, and a dead fish and tree for environmental hazard. Signal words — *Danger* for the more severe category, *Warning* for the lesser — sit alongside them.

⚠ **Never decant into an unlabelled container.** Solvent in an unmarked jar, a soft-drink bottle or a coffee cup has killed people, and it defeats every control in this section at once: the data sheet cannot be found, the pictograms are gone, the spill responder does not know what they are dealing with and the casualty department does not know what has been swallowed. If a substance is transferred, the receiving container is labelled with the product name and the hazard information before the transfer, not afterwards.

Acids, Alkalis and Battery Electrolyte

Diluting a concentrated acid releases a surprising amount of heat. Because water has an exceptionally high specific heat capacity, adding the acid slowly to a large volume of water spreads that heat through the whole body of water and the temperature rise stays modest. Do it the other way round and the first drops of water land on concentrated acid, absorb the heat of dilution in a tiny volume, flash into steam and eject a spray of concentrated acid out of the vessel and into the face of whoever is pouring. Hence the rule and the mnemonic that goes with it: **always add acid to water**, slowly, with stirring, never water to acid.



Acid poured slowly into plenty of water stays calm; water dropped onto acid flashes to steam and sprays.

It is also worth knowing that an alkali splash in the eye is generally *more* damaging than an acid splash of similar strength, which surprises people. An acid coagulates the surface protein it meets, and that coagulated layer partly limits how deep the acid can penetrate. An alkali saponifies the fatty components of the tissue and keeps advancing, so the injury continues to develop after the initial contact. That is the reason irrigation must be prolonged and must not be stopped as soon as the pain eases.

Battery type	Electrolyte	Neutralising agent for a spill
Lead-acid	Sulphuric acid (acidic)	Dilute solution of sodium bicarbonate (bicarbonate of soda), then flush with water
Nickel-cadmium	Potassium hydroxide (alkaline)	Weak acid — a boric acid solution or dilute acetic acid — then flush with water

Read the table the right way round and it is obvious: an acid is neutralised by a mild alkali and an alkali by a mild acid. Two errors follow from getting it backwards, and both appear as examination distractors. Washing spilled acid off with water alone dilutes and spreads it without neutralising it, so the acid is still acid and is now over a wider area. Reaching for caustic soda to deal with spilled battery acid is worse still: it is a strong alkali, it attacks aluminium alloy vigorously in its own right, and the neutralisation reaction itself is violently exothermic. Nor does a coating of grease or petroleum jelly neutralise anything — it may protect a clean terminal afterwards, but it does nothing to acid already on the structure.

Because a trace of one electrolyte permanently poisons a cell of the other type, lead-acid and nickel-cadmium batteries are serviced in separate rooms, with separate tools, separate hydrometers, separate protective clothing and separate spill kits. Full-face protection, an apron and gauntlets are worn for any electrolyte handling, and an eyewash must be immediately available in the battery shop.

The same acid-and-alkali logic applies to workshop reagents. Where caustic soda has been applied to bare aluminium alloy — for example as an identification reagent — it is neutralised afterwards with a chromic anhydride (chromic acid) solution, which cancels the alkali and leaves a passivating film on the exposed metal. Simply rinsing with water leaves the alkali working on the aluminium and leaves the surface unprotected.

Storage, Spillage and Waste

Hazardous substances live in designated, clearly labelled, ventilated cabinets or stores with spill containment underneath — a bund is normally sized to hold at least 110% of the largest single container in it, so that the biggest credible leak is caught. Quantities kept in the workshop are limited to what the day's work needs, with the bulk stock in the flammable store, and the store itself is of fire-resisting construction, kept away from escape routes and marked with the appropriate warning signs.

Segregation matters as much as containment, because the hazard is often in the combination rather than in either substance alone. Oxidising agents are kept away from flammable liquids and from organic materials such as rags, paper and wood; acids are kept apart from alkalis, and both from cyanide and sulphide compounds, which liberate highly toxic gas on contact with acid; and compressed gases are stored outside the chemical store altogether. The stability and reactivity section of each data sheet names the incompatibilities, and a segregation plan for the store is drawn up from those rather than from what fits on the shelf.

When something is spilled, the order of actions matters:

- **Raise the alarm and protect people first.** Warn everyone nearby, evacuate the immediate area if the material is volatile or toxic, and remove ignition sources. Do not attempt a spill that is beyond the training and equipment available — call the emergency services.
- **Stop the source** if that can be done safely: upright the drum, close the valve, plug the leak.
- **Contain it,** and above all keep it out of drains, gullies and watercourses. A drain mat or a bund of absorbent placed downhill of the spill is the difference between a workshop clean-up and a pollution offence with a criminal penalty attached.
- **Absorb and collect** using the correct medium from the spill kit — a general absorbent for oils and fuels, and the appropriate neutralising medium for acids and alkalis.
- **Dispose of the waste as hazardous waste,** through a licensed route and with the transfer documentation completed. The used absorbent is now itself contaminated waste and cannot go in the general bin.
- **Report and review.** A spill is an occurrence: it is recorded, investigated and used to fix whatever allowed it.

Never wash a spill down a drain to "dilute" it. Solvents and fuels entering a drainage system create an explosive vapour hazard in the drains themselves as well as an environmental one, and diluting a hazardous substance does not stop it being one.

Routine waste needs the same discipline. Waste streams are segregated at source — used oils, solvents, batteries, aerosols, contaminated absorbent, composite offcuts, metal swarf — because a mixed drum is expensive to dispose of and may be illegal to move. One category deserves special mention: **rags soaked in solvent, oil, or particularly in drying oils and some resins, can heat themselves.** Oxidation of the oil generates heat, a crumpled heap of rags insulates that heat instead of losing it, the temperature climbs and the pile can ignite with no external source at all. Contaminated rags therefore go into a lidded metal bin, which is emptied at the end of every shift.



Inside a crumpled heap of soaked rags, oxidation heat is trapped at the core until the pile can ignite.

Fire Safety and Extinguishing Agents

A fire requires three elements (the **fire triangle**): fuel, oxygen, and heat (ignition source). Removing any one element extinguishes the fire. For some fires, a fourth element — the chemical chain reaction — must also be broken (the **fire tetrahedron**).

Fire Classes and Extinguishing Agents

Class	Fuel Type	Examples	Suitable Extinguisher
A	Ordinary combustibles	Wood, paper, cloth, rubber	Water, foam, dry powder
B	Flammable liquids	Fuel, oil, grease, solvents	Foam, CO_2, dry powder
C	Flammable gases	Propane, acetylene, hydrogen	Dry powder (shut off gas supply first)

Class	Fuel Type	Examples	Suitable Extinguisher
D	Combustible metals	Magnesium, titanium, lithium	Specialist dry powder (e.g. L2, Ternary Eutectic Chloride)
Electrical (not a fuel class under BS EN 2 — classify by what is burning)	Energised electrical equipment	Panels, motors, avionics	CO_2, dry powder (never water or foam)
F	Cooking oils and fats	Deep-fat fryers, galley equipment	Wet chemical (never water)

Key Extinguishing Agents:

- **Water** — cools the fuel below ignition temperature. Class A only. *Never* use on electrical, oil/fat, or metal fires.
- **Foam (AFFF)** — smothers by forming a film that excludes oxygen. Classes A and B.
- **CO_2 (Carbon Dioxide)** — smothers by displacing oxygen. Leaves no residue — ideal for avionics and electrical fires. Very little cooling effect — fire may re-ignite.
- **Dry Powder (ABC)** — chemical interruption of the chain reaction. Versatile but leaves residue that can damage avionics.
- **Halon alternatives (HFC-227ea / Novec 1230)** — clean agents that extinguish the fire without leaving residue. Used in aircraft engine/APU fire suppression and avionics bays.

⚠ **Magnesium fires:** Burning magnesium (common in aircraft wheels and casings) produces intense white light, extreme heat (3,100°C), and reacts violently with water — producing hydrogen gas and risk of explosion. Use only specialist Class D extinguishing agents. Never apply water, foam, or CO_2.

How Each Agent Attacks the Fire

Every extinguishing agent works by removing one or more sides of the triangle or by breaking the chain reaction, and knowing which side an agent attacks tells you immediately where it will fail.

- **Cooling** removes the heat. Water is unmatched at this because of the energy it absorbs in turning to steam, and because the steam produced also displaces air above the fuel. Its weakness is everything else: it conducts electricity, it floats burning liquids and spreads them, and on hot metal it reacts.
- **Smothering** removes the oxygen. Foam does it by floating a blanket on a liquid surface and sealing the vapour underneath; carbon dioxide does it by displacing the air in the immediate vicinity. Both fail the moment the blanket is broken or the gas disperses, which is why a foam blanket must not be walked through and why a carbon dioxide discharge in a draught achieves very little.

- **Starvation** removes the fuel — shutting a valve, isolating a tank, or letting a gas fire burn out under control. For a pressurised gas fire this is not one option among several, it is the only real answer.
- **Chain-breaking** interferes chemically with the free radicals that propagate combustion, extinguishing the flame without necessarily cooling anything. Halons and dry powders work this way, which is why they knock a flame down so fast, and also why the fuel underneath is still hot enough to re-ignite the instant the agent clears.

Worked example — why water is the best coolant, and why carbon dioxide is not.
Warming one kilogram of water from 20 °C to boiling absorbs about 4.18 kJ for every degree, so $4.18 \times 80 = 334$ kJ. Turning that same kilogram from water at 100 °C into steam at 100 °C absorbs a further 2,260 kJ — without the temperature changing at all. The total is roughly 2,594 kJ, of which about $2260/2594 = 87\%$ comes from the change of state alone.

That is why water extinguishes by cooling in a way no gas can match, and why the steam it forms — expanding to well over a thousand times the volume of the liquid — smothers as a bonus. Carbon dioxide, by contrast, leaves the horn at about –78 °C as a mixture of gas and dry-ice snow, and will certainly frost-burn skin. The snow sublimes rather than melting, absorbing about 570 kJ for every kilogram — but that is only about a quarter of water's latent heat of vaporisation, the mass of snow formed is small, and much of that change of state takes place in the plume rather than on the burning surface. It removes very little heat from a burning mass, which is exactly why a fire it has apparently put out can re-ignite from residual hot surfaces once the gas has dispersed.

Two consequences of the agent list above are worth stating plainly. Dry powder is the fastest flame knock-down available and is effective across several classes, but the residue is fine, pervasive and mildly corrosive, so a discharge inside an equipment bay or a flight deck writes off connectors and boards that the fire itself might not have touched — which is why the clean agents are preferred anywhere near avionics. And foam has to be applied gently, off a vertical surface or the back wall of a tray, so that the blanket forms and flows; directed straight into a burning liquid it simply punches through and splashes it about.



Foam aimed at the back wall of the tray runs down and floats across the burning liquid as a sealing blanket.

Reading the Class Table

The European classification groups fires by *what is burning*, and the table above follows it: ordinary combustible solids, flammable liquids, flammable gases, combustible metals and, at the end, cooking oils and fats. Cooking oils get their own class because they burn at a temperature far above the boiling point of water: applying water to a burning pan does not cool it, it flashes to steam instantly and throws burning fat across the room. A wet chemical agent works instead by reacting with the fat to form a soapy crust that seals the surface. On an aircraft the class is not academic — galley equipment and deep-fat fryers are exactly where it applies.

Electricity is treated differently, and this catches candidates out. In the European scheme it is **not a class of its own**, because electricity is an ignition source rather than a fuel: what is actually burning in an "electrical fire" is the cable insulation, the plastic housing or the oil in a transformer, so the fire is classified by that. What the electrical involvement changes is the *agent*, which must be non-conductive so the operator does not receive a shock back up the jet, and which must not leave a conductive or corrosive residue on equipment. Isolating the supply is the first action and often the decisive one: once the power is off, what remains is an ordinary fire of whatever material was ignited. Some other national schemes do label energised equipment as a separate class, which is why the letter appears in older material and in questions written outside Europe.

One nuance about carbon dioxide follows from its very limited cooling. It is rated primarily for flammable liquids and is the standard choice around live electrical equipment because it is non-conductive and leaves no residue. Directed at a surface fire in ordinary combustibles it will knock the flames down, but it cannot extinguish a deep-seated fire in wood, paper or upholstery, which will re-ignite from within — and that is why the class table above lists water, foam and dry powder against ordinary combustibles rather than carbon dioxide.

Extinguisher Identification and Colour Coding

Since the European standard was adopted, every portable extinguisher body is **signal red**, and the contents are identified by a coloured band or panel on the body together with the printed label. Before that, the whole body was finished in the agent's colour, and older extinguishers in that form are still encountered — and still appear in examination questions. The agent colour itself did not change; only how much of the extinguisher it covers. Alongside the colour, the label states in class letters exactly what that particular unit is approved for, and it is the label rather than the colour that is authoritative.

Agent	Identification colour	Class letters printed on the label
Water	Red	A
Foam	Cream	A and B
Dry powder	Blue	A, B and C, plus a marking for use near live electrical equipment
Carbon dioxide	Black	B, plus a marking for use near live electrical equipment

Agent	Identification colour	Class letters printed on the label
Wet chemical	Yellow	F (many units are also rated A)
Halon, e.g. BCF (legacy)	Green	Withdrawn from general use; still met on aircraft

Halons are ozone-depleting substances, banned for general use under the Montreal Protocol and the European regulation implementing it. Civil aircraft fire protection is one of a shrinking list of "critical uses" for which they remain permitted, with separate phase-out dates set for hand-held extinguishers, lavatory systems, cargo bays and engine and APU installations — which is why an engineer still meets halon on the aircraft while it has disappeared from the hangar wall. The replacements named in the agent list above take halon's place without the ozone depletion.

The label also carries a *fire rating* — a number and a letter stating the size of standard test fire the unit will extinguish, which is what distinguishes two identically blue-panelled powder extinguishers of different capacities — and the operating instructions. Separately, an extinguisher carries a service label: it is inspected regularly, weighed or pressure-checked, and it is taken out of service after *any* discharge, however brief, because a partially discharged unit will not perform when it is next picked up.

Metal Fires: Why the Normal Rules Reverse

The reason burning magnesium behaves so violently with the ordinary agents is that at the temperature quoted in the warning above it is hot enough to strip oxygen out of compounds that are normally inert in a fire.

- **With water:** the magnesium takes the oxygen from the water molecule, leaving magnesium oxide and free hydrogen — and hydrogen released into a fire that hot promptly explodes. Applying water also causes violent steam explosions that scatter burning metal.
- **With carbon dioxide:** the same thing happens. Magnesium reduces carbon dioxide to carbon, taking the oxygen for itself, so the "inert" gas actually supplies the oxidiser. Magnesium will continue to burn in an atmosphere of pure carbon dioxide, which is precisely why carbon dioxide is listed as prohibited alongside water.

The only agents that work are dry, chemically inert powders that form a crust over the burning metal and exclude the air without reacting with it — the specialist Class D powders identified in the class table, which include graphite-based, sodium-chloride-based and ternary-chloride formulations, together with the dry limestone or chalk powders used for burning swarf in a machine shop. Older manuals also list dry asbestos wool with chalk for the same purpose, and it works on exactly the same principle of an inert dry blanket, though asbestos is no longer used in new work because of its own health hazard. Two practical points: dry *sand* is not a reliable substitute because it almost always holds some moisture, which supplies the water reaction above; and a Class D agent is applied gently, building a blanket over the metal rather than blasting it,

because a high-velocity discharge scatters burning particles across the shop. Once covered, the metal is left to burn out under the crust rather than disturbed.

Titanium behaves the same way. Bulk titanium is difficult to ignite, but titanium *swarf*, with its enormous surface area, will burn fiercely and reacts with water and carbon dioxide exactly as magnesium does — which is why swarf is cleared from machines and kept in covered metal bins rather than allowed to accumulate.

Wheel, Brake and Engine Fires

These two aircraft fires call for different agents, and the reason is the fuel:

- **A wheel or brake fire takes dry powder.** Wheels and some brake components are magnesium and aluminium alloys, so an established fire there is a metal fire with all the chemistry above. There is a second reason to keep water and carbon dioxide away even before the metal ignites: an aircraft wheel after a heavy braking event is extremely hot, and quenching it locally causes thermal shock that can crack the wheel and turn a hot brake into an explosive wheel failure. The correct response to a merely overheated brake is to leave it alone in the designated cooling position, allow it to cool in air or with a cooling fan, and let the fusible plugs do their job — they are designed to melt and deflate the tyre in a controlled way rather than allow it to burst.
- **A ground engine fire takes carbon dioxide.** What is burning is fuel, oil and electrical insulation, so a clean, non-conductive smothering agent that leaves no residue in the gas path is what is wanted; water would spread burning fuel and create a shock hazard, and dry powder would contaminate the engine. Discharge is directed into the intake, the exhaust or the drains as the maintenance manual specifies, and the aircraft's own fire bottle is used if the installation allows it. Do **not** open the cowl doors to get at the fire: opening them admits air and can turn a contained fire into a flare-up in the operator's face.



Ground engine fire: carbon dioxide is discharged into the intake while the cowl doors stay shut.

✈ **Exam discriminator:** two aircraft fires, two opposite answers, and the difference is the fuel. Brake and wheel fire — burning metal — dry powder, never water or carbon dioxide. Engine fire — burning fuel and oil around electrical components — carbon dioxide, never water. If a question offers carbon dioxide for a wheel fire or water for an engine fire, both are wrong for the same underlying reason: the agent has to match what is actually alight.

Extinguishers Inside the Aircraft

The cabin imposes an extra requirement that no workshop extinguisher has to meet: whatever the agent does to the fire, it must not poison the people who cannot leave. Cabin extinguishers are therefore water or water/glycol units for fires in ordinary combustible materials such as seat covers, paper and furnishings, plus a halon unit — historically BCF — or its modern clean-agent replacement for everything else, halon's comparatively low toxicity at the concentrations used being the reason it survived in the cabin long after it left the hangar.

Two agents once carried in aircraft are now prohibited, and the reason in both cases is toxicity rather than fire performance. Carbon tetrachloride decomposes on hot surfaces into phosgene and other highly toxic products, so using it in an enclosed occupied space can kill the occupants faster than the fire; methyl bromide is acutely toxic in its own right. Neither may be used in a cabin, and the general principle carries across to the hangar — agent selection in any confined occupied space has to account for what the agent and its decomposition products will do to the people in there.

The lithium cells in passenger electronic devices deserve a separate note, because the class table above lists lithium among the combustible metals and that classification applies to lithium *metal* primary cells. In a rechargeable lithium-ion device the fire is dominated by burning electrolyte and by cell-to-cell thermal runaway propagating through the pack, so the published procedure is to knock the flames down with the available extinguisher and then to cool the device with a non-flammable liquid to stop the propagation, rather than to treat it as a metal fire. Which procedure applies is set out in the operator's manuals, and it is one of the few cases where "put water on it" is the correct answer for something that looks like a battery fire.

Using an Extinguisher

A portable extinguisher is for a fire in its earliest stage and for protecting an escape route. It is not a fire-fighting capability, and the decision to use one is made in a couple of seconds:

- **Raise the alarm first,** before doing anything else, so that the evacuation and the fire service response are already running whether or not your attempt works.
- **Only tackle a small fire,** and only with the right agent to hand. If it is larger than about the size the extinguisher is rated for, if it is producing significant smoke, or if you would have to enter the smoke to

reach it, leave it.

- **Keep your escape route behind you.** Stand between the fire and the way out, never the other way round, and get out if the fire is not going out.
- **Aim at the base of the flames,** from the edge nearest you, sweeping across; on a liquid fire in a tray, work from the near edge to the far one so the blanket builds towards the fire rather than being punched into it.
- **Do not hold the horn** of a carbon dioxide extinguisher unless it is a frost-free type — the horn reaches the discharge temperature within a second or two.
- **Ventilate afterwards** if carbon dioxide has been discharged in a small or enclosed space; the agent works by displacing oxygen, and it does not distinguish between the fire's air supply and yours.
- **Watch for re-ignition,** particularly after a dry powder or carbon dioxide discharge, and report every use so the unit is replaced or recharged.

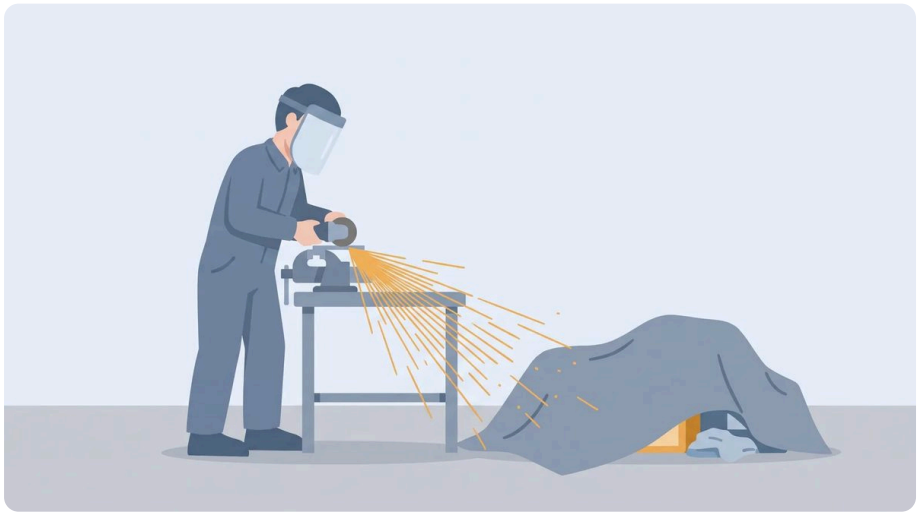
Fire Prevention in the Hangar

An aircraft in a hangar concentrates all three sides of the triangle in one building: tonnes of fuel, fuel-gas cylinders, solvents, paints, composite dust and packaging on the fuel side; the oxygen system and the hangar's own air on the second; and hot work, grinding sparks, electrical faults, static and smoking materials on the third. Hangar fire rules are therefore stricter than ordinary workshop rules, and all of them apply, not merely the smoking ban:

- No naked flames or unauthorised spark-producing work; smoking only in designated areas away from the building.
- Flammable liquids kept in the flammable store, with only the working quantity in the hangar, in approved containers, and returned at the end of the shift.
- Fire doors kept closed — a wedged-open fire door protects nothing, and its only function is to hold back fire and smoke for the time the escape needs.
- Escape routes, exits, fire alarm call points, extinguisher points, hydrants and hose reels kept permanently clear. "Temporarily" parking a stand or a tug across an exit is exactly the situation the rule exists to prevent, and posting somebody to watch it does not substitute for keeping it clear.
- Aircraft positioned so they can be towed out, with a towbar and tug available and the crew trained to use them, and with the aircraft's fuel state and configuration considered as part of the fire plan.
- Housekeeping treated as fire prevention: rags, packaging, composite offcuts and dust are all fuel.
- Fixed detection and suppression systems — hangar deluge or high-expansion foam — maintained and understood. An inadvertent foam discharge is itself a hazard: it fills the building in minutes, hides obstacles and open pits, and is extremely slippery.
- Everybody knowing the alarm signal, the nearest two exits and the assembly point, and a roll call being taken so that nobody is searched for who is already outside — or missed who is not.

Hot Work

Hot work — welding, brazing, soldering, grinding, cutting, heat-shrinking, anything that produces a flame, an arc, a hot surface or sparks — is one of the most common causes of industrial fires, and near an aircraft it is controlled by permit. The permit is issued by an authorised person after the precautions have been checked, and it is specific to the location and the time period:



Combustibles that cannot be removed from a hot-work area are covered with fire blankets before work starts.

- Combustible material is removed from the area, or where it cannot be removed it is covered with fire blankets or screened; this includes what is *below* the work, because sparks fall through floor gaps, cable runs and open panels and start a fire on the deck beneath while everybody is watching the torch.
- Fuel and oxygen systems are isolated and, where the work is on or near a tank, the space is gas-tested and certified free of flammable vapour before the permit is issued — and re-tested if work is suspended.
- Suitable extinguishing equipment is positioned within reach before work starts, and a **fire watch** is posted: a person whose only job is to watch for ignition, with an extinguisher and the means to raise the alarm.
- The fire watch continues *after* the work stops, for the period the permit specifies — commonly at least half an hour — because material heated by grinding or welding can smoulder unnoticed for a long time before it flames. The area is checked again before the permit is handed back.
- Aircraft transparencies, composite structure, wiring and soft furnishings near the work are protected from heat and spatter, and adjacent work is stopped or screened.

A hot-work permit cannot be self-issued by the person doing the work, for the obvious reason that the point of it is independent verification that the precautions are in place. Nor is it a marine or heavy-industry formality: it applies in an aircraft hangar with more force than almost anywhere else, because of what is parked in the middle of it.

✈ Hazards Around a Live Aircraft

The hazards dealt with so far belong to substances and equipment. The aircraft itself is a hazard of a different kind, because it stores energy in a dozen forms and because it is surrounded by other people's work. Almost every serious ramp and hangar injury comes from the same root cause: a system did something the person nearby had assumed it could not do.

Propellers and Rotors

A propeller is treated as live at all times, by everybody, whether or not the engine has been shut down and whether or not the switches are off. Two mechanisms justify that rule rather than mere caution. A piston engine's magneto generates its own ignition and is stopped only by the switch *earthing* its primary lead; if that lead is broken, disconnected or its switch has failed, the magneto is still hot with the switch in the OFF position and the engine can fire when the propeller is turned by hand. And a turboprop or a helicopter rotor with residual power available, a wind gust, or a hydraulic or electrical system re-energising, can move without anyone touching it.

The practical rules are simple and absolute: never stand, walk or place equipment in the plane of rotation of a propeller or rotor; never approach a running or recently-shut-down aircraft without eye contact and a signal from the crew; treat "engine shut down" and "ground power connected" as reasons for more care rather than less. Where a propeller must be moved by hand for a maintenance task, it is done to the procedure in the manual, with the ignition confirmed dead and everyone positioned clear of the arc.

Helicopters add their own geometry. The main rotor droops when it slows and flaps in wind, so the disc can dip to head height at the edge; the tail rotor is close to invisible when turning and is the reason nobody approaches a helicopter from the rear. The standard approach is from the front or the front quarter, in the pilot's field of view, crouched, having been signalled in. On sloping ground the approach is always from the **downhill** side, because the disc is nearer the ground on the uphill side. Anything carried — a ladder, a panel, a length of cable — is carried horizontally, never on the shoulder.



A slowing main rotor droops at its edge, so the approach is crouched with anything carried held horizontally.

Engine Intake and Exhaust Danger Zones

A running turbine produces two lethal zones that must be observed at all power settings, not only during high-power running:

- **The intake** generates enough suction at ground idle to ingest a person, and people have been killed by walking too close to an engine that was, by any casual judgement, "only ticking over". Loose items — hats, ear defenders, identity cards, rags, panels, unsecured ground equipment — go the same way and destroy the engine on the way through.
- **The exhaust** produces a high-velocity, high-temperature blast that remains dangerous a long way behind the aircraft and can blow over people, equipment and light aircraft, and can pick up stones and turn them into projectiles.

Both zones grow substantially with thrust, which is why the manuals give separate danger-area diagrams for ground idle and for take-off power, and why the higher-power zones are the ones respected during a ground run. Before any ground run the intake, the surrounding area and the surface upwind are checked for debris, the area is roped off or marked, communication with the flight deck is established and tested, and hearing protection is worn by everyone within the noise zone. Nobody enters a danger area on the assumption that the engineer at the controls has seen them.

Radar, HF and Other Radio-Frequency Radiation

Radio-frequency energy at the powers used by weather radar and HF transmitters is absorbed by tissue and turned into heat. The eye and the testes are the most vulnerable organs because both are poorly supplied with blood and therefore cannot carry the heat away; the recognised injuries range from a superficial warming

sensation through deep tissue burns to cataract, and because the deeper heating produces no immediate pain there is no reliable warning at the time.

Three rules cover almost all of it. First, the minimum safe distance and the safe sector are given in the aircraft maintenance manual for the specific installation, and they are not a matter of judgement. Second, weather radar is not operated inside a hangar or with anyone in front of the antenna — reflections from hangar walls and from nearby aircraft can produce field strengths well above those in the open, and a scanner parked in one position concentrates the whole output into a narrow sector. Third, the transmitter is placarded off and the circuit breaker collared before any work in front of the antenna, because a built-in test or a system self-check can energise it without anybody selecting it.

Two further points belong here. Radiated energy can ignite fuel vapour and can fire electro-explosive devices, so radar and HF transmission are prohibited during refuelling and during work on any pyrotechnic or cartridge-actuated system such as a fire-bottle squib or an escape-slide inflator. And touching an HF antenna while the set is transmitting produces a deep RF burn, so the aircraft is earthed and the HF is placarded off during any work near it.

Wheels, Tyres and Brakes

An inflated aircraft tyre stores an enormous amount of energy, and the wheel that contains it is a two-piece assembly held together by tie bolts. A wheel failure releases that energy in a fraction of a second, and the debris — the wheel halves, the tie bolts and the bead flange — is thrown outwards along the axis of the axle, to either side. The recognised danger zone is therefore the region in line with the wheel axis, and the safe position for approaching or working near a wheel is **fore or aft of it, in the plane of the tyre**. This is the same geometry as the hot-brake rule, and for the same reason.

- Inspect the tyre and the wheel for cuts, bulges, flat spots, damaged or missing tie bolts and evidence of overheating *before* inflating anything. Inflating a damaged wheel is how the assembly fails.
- Inflate through an extension hose with a remote gauge, from outside the danger zone, and never exceed the rated pressure — which is a specific figure for the specific tyre and aircraft, taken from the manual, and is not "whatever it had last time".
- Deflate the tyre **completely** before removing a wheel from the axle and before slackening any tie bolt. Splitting a wheel that still holds pressure is one of the classic fatal maintenance accidents, and the fact that the tyre "looks flat" is not evidence that it is.
- Inflate with dry nitrogen, for the reasons given in the gas section, and never with oxygen.
- After heavy braking, leave the assembly alone. Approach only fore or aft, do not spray water or any other coolant onto a hot wheel, allow the cooling period the manual specifies or use an approved cooling fan, and remember that the fusible plugs are designed to release the tyre pressure in a controlled way rather than let the tyre explode — a deflated tyre after a hot-brake event is the system working, not a separate fault.

Stored Energy and Unexpected Movement

A parked aircraft is full of things that can move on their own or under someone else's command:

- **Landing gear.** Ground locks and gear pins are fitted before anyone works in or passes through a wheel well, and they are removed only by the person authorised to do so, at the point in the procedure where the manual says. The streamers on them are long and red for exactly one reason.
- **Flying controls, flaps, slats, spoilers and thrust reversers.** These move under hydraulic or electrical power with no warning to someone standing in the way, and they also move under their own weight when a system is depressurised. Isolation, warning placards at the controls and a check that nobody is in the path come before any system operation.
- **Doors and escape slides.** A door with its slide armed will fire the slide when it is opened, at a speed and force that has broken bones and thrown people off stands. Doors are disarmed before being opened for maintenance, and the arming state is checked rather than assumed.
- **Accumulators, gas struts, springs and tensioned cables.** These retain their energy indefinitely after the aircraft is shut down and are released only through the procedure in the manual.
- **Hot and pressurised systems.** Bleed ducts, engine components and brakes stay hot for a long time; oxygen and hydraulic systems stay pressurised until deliberately vented.

Every safety pin, ground lock, blanking cap and cover fitted during maintenance is an item that must be accounted for and removed before the aircraft flies, and the record of them is part of the release paperwork. A gear pin left in place after maintenance and a gear pin missing during maintenance are both serious — one grounds the aircraft, the other endangers the person under it.

Warning Notices and Tags

A red warning tag or placard fitted to an aircraft is a mandatory instruction, not advice. It signifies a hazardous condition, an inoperative system or maintenance in progress, and the aircraft or the system must not be operated until the responsible person has cleared the work and removed it. Only the authorised maintenance person who owns the work removes it — not a ramp agent who wants to reposition the aircraft, not the crew, and not the next shift because it looks finished. That, along with the physical lock in the isolation described earlier, is what stops one person's convenience becoming another person's injury. Tags are numbered or logged so that their removal can be verified before release to service.

Ground Movement, Towing and Jacking

Towing is a whole-team operation, and the injuries it produces are crushing injuries. The tow crew works to briefed roles — a person in charge, a brake rider on the flight deck where required, wing walkers with a clear view of the tips and the tail, and intercom or agreed hand signals throughout — and the aircraft stops the

instant communication is lost. The towbar and head are matched to the aircraft type, the nose-gear steering is bypassed or the torque link disconnected as the manual requires, and the **nose-gear turn limits are never exceeded**: overturning the nose gear damages the steering mechanism and the damage is not always obvious afterwards, so an over-turn is a reportable event rather than something to drive out of.

Jacking concentrates a very large mass on three or four small points and is unforgiving of shortcuts:

- Jack only at the designated jacking points, with the correct adapters, on a level and load-bearing surface, ideally inside; outdoors there are wind limits and they are observed.
- Jacks are positioned squarely, plumb under the point, and raised together and evenly so the aircraft is never twisted or allowed to run off a jack head.
- Tail-tipping types are supported with a tail stand or ballasted before anything is removed.
- Locking collars are wound down to follow the jack as it rises, so that a hydraulic failure cannot drop the aircraft; nobody works under an aircraft supported only by hydraulic pressure.
- The centre of gravity is not changed while the aircraft is on jacks — removing an engine, defuelling, or moving a heavy component — except in the sequence and with the ballast the manual specifies.
- The area under and around the aircraft is controlled: no towing, no other movement, warning notices at the flight deck and at the jacking points, and the aircraft is left jacked only if the procedure allows and the notices are in place.

Accident Remedial Action and First Aid

- **Electric shock** — isolate the supply first (switch off or pull plug). Do not touch the victim while they are in contact with the live source. If breathing has stopped, begin CPR immediately.
- **Chemical burns** — flush affected area with copious clean water for at least 20 minutes. Remove contaminated clothing. Seek medical attention.
- **Chemical splash in eyes** — irrigate with clean water or eyewash for at least 15 minutes, holding eyelids open. Seek immediate medical attention.
- **Fuel/solvent inhalation** — move to fresh air. If unconscious, place in recovery position and call emergency services.
- **Fire on a person** — stop, drop, and roll. Smother with fire blanket if available. Do not use CO_2 extinguisher directly on a person (risk of cold burns and suffocation).

The Order of Actions

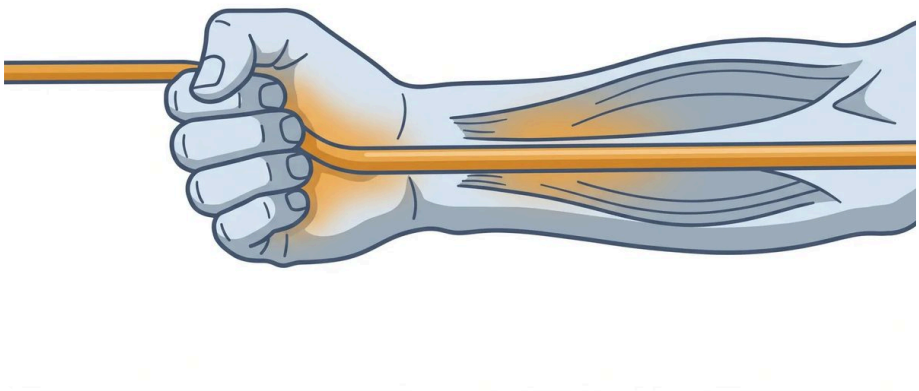
Every one of the responses listed above fits inside the same framework, and knowing the framework is what lets an engineer act correctly in a situation the list does not cover. It is usually remembered as **D–R–A–B–C**:

- **Danger.** Make the scene safe before you touch anybody. This is first for a hard practical reason rather than a legalistic one: the second casualty is the person who ran in to help, and two casualties are far more than twice as bad as one, because there is now nobody left to raise the alarm or to guide the ambulance in. Isolate the supply, stop the machine, ventilate the space, remove the ignition source.
- **Response.** Speak to the casualty and gently shake the shoulders. A casualty who answers has an airway and is breathing, which changes everything that follows.
- **Airway.** An unconscious casualty's airway is most often blocked by their own tongue; opening it by tilting the head and lifting the chin is the single most valuable thing an untrained bystander can do.
- **Breathing.** Look, listen and feel. If the casualty is not breathing normally, start resuscitation without delay and send someone else for help — the instruction to send *someone else* is deliberate, because leaving a non-breathing casualty to go and find a telephone yourself costs the minutes that decide the outcome.
- **Circulation.** Control severe bleeding with direct pressure and elevation.

An unconscious casualty who *is* breathing normally and has no suspected spinal injury is placed in the recovery position, which keeps the airway open and lets vomit drain rather than be inhaled. A casualty who has been in an atmosphere that could still affect them is moved into clean air first, provided that can be done without the rescuer entering it.

Electric Shock: What Makes It Different

Three features of electrical injury change the response, and all three follow from the mechanisms described earlier in this note.



Above the let-go threshold, current clamps the hand shut on the conductor, so the casualty stays energised.

- **The casualty may still be part of the circuit.** Because a current above the let-go threshold clamps the hand shut on the conductor, the casualty frequently cannot release it and is still energised when you arrive.

Touching them makes you the next casualty at the same current. Isolate at the switch, the breaker or the plug — that is always the first choice. Only if isolation is genuinely impossible is the casualty pushed or dragged clear using a dry insulating object, by someone standing on a dry insulating surface and touching nothing else.

- **High voltage is different again.** At the voltages present in radar and HF installations, and in airport high-voltage distribution, current can arc to a person who never touches the conductor, and can flow through the ground around a fallen cable. Nobody approaches such a casualty until the responsible authority confirms the supply is isolated and earthed, however long that takes.
- **The visible injury understates the damage.** Current passing through the body burns along its whole path, and the entry and exit marks on the skin can be small while the muscle and nerve damage between them is extensive. Cardiac rhythm disturbances can also develop hours after an apparently minor shock. **Anyone who has received an electric shock goes to hospital**, even if they feel perfectly well, and even if the shock was "only" from the workshop supply.

Burns, Cold Burns and Special Chemical Cases

For a thermal burn, cooling is the treatment and it works far better than most people expect — cool running water for a prolonged period both limits the depth of the injury and relieves pain. Clothing that is not stuck to the skin is removed along with rings and watches before swelling starts; anything adhering is left alone. Blisters are not burst, no cream, fat or ointment is applied, and the burn is covered with a sterile non-fluffy dressing or clean cling film, which does not stick.

Cold burns need mentioning because they come from equipment nobody thinks of as dangerous. Carbon dioxide discharges at about –78 °C and will freeze skin on contact with the horn or the jet; escaping high-pressure gas cools sharply as it expands, and cryogenic liquids are colder still. A cold injury is rewarmed gently with body heat or tepid water — never rubbed, never heated directly — and treated as a burn thereafter.

Three chemical situations in this trade have a specific response that generic first aid does not cover:

- **Hydrofluoric acid,** present in some etchants, deoxidisers and paint strippers, is uniquely dangerous. It penetrates the skin and destroys tissue from within, the pain can be delayed for hours, and absorbed fluoride disrupts the body's calcium balance and can cause cardiac arrest from a burn covering only a small area. Where it is used, the specific calcium gluconate antidote gel must be immediately available, applied after flushing, and the casualty goes to hospital regardless of how the burn looks.
- **High-pressure fluid injection** is a surgical emergency dealt with under hydraulic fluids above — the point to carry here is that it presents as a trivial puncture and must never be triaged as one.
- **Oxygen-saturated clothing.** After a leak or a purge, clothing and hair hold a high concentration of oxygen for some time and will flare violently at any spark or flame. The casualty or the worker moves to well-

ventilated fresh air, stays well away from any ignition source, and the clothing is aired thoroughly before they go anywhere near hot work, a vehicle or a smoking area.

Rescue from a Confined Space or an Oxygen-Deficient Atmosphere

This is the situation in which the instinct to help is most likely to kill the helper, so it is worth having decided in advance. A colleague who has collapsed inside a tank, a duct or a purged compartment is direct evidence that the atmosphere in there is lethal, and it will do exactly the same to anyone who follows them in. The standby person's job is to raise the alarm, to attempt recovery from *outside* using the retrieval line and harness that were rigged before entry, and to stop anybody else going in. Entry rescue is carried out only by trained rescuers with breathing apparatus. It is also the reason the rescue arrangements are made a condition of the entry permit, rather than something improvised once a person is already down.

First-Aid Provision

Provision is set by assessment: how many people, spread over what area and shift pattern, exposed to which hazards, and how far away the emergency services are. The practical minimum in a maintenance organisation is trained first-aiders on every shift with their names and locations posted, stocked and in-date first-aid kits, and an accident record.

Two items are worth singling out because they are effective only if their siting is right. **Emergency eyewash** must be immediately accessible wherever chemicals are used — the widely used standard requires it to be reachable within about ten seconds' travel from the hazard and to deliver a continuous flush for at least fifteen minutes — and every person working in the area must know where it is without looking, because a casualty with a chemical in their eyes cannot see to find it. A unit "somewhere in the building" is not compliant and, more importantly, is not useful. Sealed single-use eyewash pods serve the same purpose out on the line where plumbed-in stations do not exist. **Emergency showers** serve the same function for a body splash and are sited on the same reasoning. Both are tested and flushed on schedule, because a stagnant eyewash line grows contamination that will be flushed straight into an injured eye.



Eyewash sited beside the chemical work, so a casualty who cannot see can reach it within seconds.

Finally, know the practicalities before you need them: the emergency number, the site address and gate the ambulance should use, who will meet it, and where the assembly point is. A hangar on an airfield is not a location an emergency vehicle can find unaided.

Reporting

All accidents, near-misses, and hazardous occurrences must be reported immediately to the supervisor and documented in the company's safety reporting system. In EASA-regulated environments, serious incidents must also be reported to the competent authority.

"Immediately" is doing real work in that sentence. A report made at the time captures the scene as it was, the position of the equipment, the state of the guard, the witnesses while they are still on shift and their memories before they have been discussed and reshaped. A report written up at the end of the week captures a reconstruction. Where it is safe to do so and nobody needs treatment, the area is left undisturbed for the investigation.

Near Misses Are the Cheap Lessons

A near miss is an event that could have caused injury or damage and, by chance rather than by design, did not: the spanner that fell from the stand and landed beside somebody, the panel that was found unsecured, the wrong fluid that was noticed before it went into the reservoir. It carries the same information about the failed control as an accident does, at none of the cost.

That is the whole argument for reporting them. Serious injuries are rare, so an organisation that only investigates injuries is learning from a tiny and randomly-selected sample of its problems, always after the damage. Near misses are numerous, and each one identifies a defective control before it collects on the debt.

An organisation whose near-miss reporting rate *rises* is usually getting safer, not more dangerous — the underlying events were always happening and are now being seen.

Reporting only survives if it is easy and if it is safe. A form that takes twenty minutes, or a culture in which the reporter is asked what they did wrong, produces exactly the silence it deserves; the events continue and nobody hears about them until one of them injures somebody. Waiting for visible damage before treating something as worth reporting discards precisely the preventive value the system exists for.

Occurrence Reporting and Just Culture

Two reporting regimes run in parallel in a maintenance organisation and an engineer needs to recognise both.

- **Aviation occurrence reporting.** European law requires occurrences that endanger, or which if not corrected would endanger, an aircraft or its occupants to be reported through the organisation's mandatory reporting scheme and passed to the competent authority, normally within 72 hours of the reporter becoming aware of the occurrence. A Part-145 organisation is required to run an internal safety reporting scheme as part of its management system, both to feed that external reporting and to capture the hazards that never reach the threshold of an occurrence. Typical maintenance examples are an incorrect assembly found in service, a tool left in an aircraft, an incorrect fluid uplift, and damage discovered but not recorded.
- **Workplace health and safety reporting.** National legislation separately requires certain injuries, occupational diseases and dangerous occurrences at work to be reported to the health and safety authority. The two regimes have different thresholds and different recipients, and a single event — an engineer injured by an escape slide firing during maintenance, say — can be reportable under both. Satisfying one does not satisfy the other.

Both regimes depend on **just culture**, which is frequently misdescribed as a no-blame culture. It is not. A just culture undertakes that people will not be punished for actions, omissions and decisions that are commensurate with their experience and training — that is, for honest error — while making equally clear that wilful violations, gross negligence and destructive acts are not tolerated. The distinction is what makes it work: a promise of total immunity would be neither credible nor safe, and a policy of punishing error simply hides the error. The reporter's identity is protected, and reports are used to fix systems rather than to build a case against individuals.

Investigation and Corrective Action

A report that is filed and not acted on is worse than no report, because it creates a record that the hazard was known. The investigation looks for the conditions that allowed the event rather than for the person at the end of the chain: what the procedure actually said, whether the tooling was available, how the shift handover worked, whether the lighting or the access made the correct action difficult, what pressure the team was under. Those are the factors that can be changed; "be more careful" cannot.

The output is a corrective action with an owner and a date, and the loop is closed only when someone has verified that the action was taken and that it worked. The findings then have to travel — through the safety notice, the toolbox brief, the amended procedure or the revised training — because the next person to meet the same trap will be somebody who was not there. Trend analysis over many small reports is what identifies the hazard that no single event made obvious, and that is the reason the small reports are worth collecting in the first place.

Additional Workplace Hazards

Noise

Aircraft engines, APUs, pneumatic tools, and riveting guns produce noise levels well above 85 dB(A) — the threshold for mandatory hearing protection. Prolonged exposure causes permanent noise-induced hearing loss (NIHL). Wear ear defenders or ear plugs in all designated hearing protection zones.

Two features of noise damage explain why the rules look the way they do. The first is that it is **painless and permanent**. The sound energy destroys the outer hair cells of the cochlea, which do not regenerate; there is no soreness the next day and no recovery. The loss begins in the region around 4 kHz, which sits above most of the energy in speech but right on the consonants, so the earliest symptom is not "everything is quieter" but "I can hear people talking in the crew room, I just cannot make out what they are saying". By the time the loss is obvious in conversation, a large part of it has already happened. Ringing in the ears or muffled hearing at the end of a shift — a temporary threshold shift — is the warning that the exposure that day was too high.



Early noise loss: speech is still heard in the crew room, but the words can no longer be made out.

The second is that the decibel scale is **logarithmic**, so the numbers do not behave the way intuition suggests. An increase of 3 dB doubles the sound energy reaching the ear, which means it halves the time you can safely

be exposed. The damage depends on the total dose — level combined with duration — not on the peak level alone.

Worked example — equal daily noise dose.
Working the 3 dB exchange rate upwards from an eight-hour shift at 85 dB(A) gives four exposures that are all the same daily dose:

- 85 dB(A) for 8 hours
- 88 dB(A) for 4 hours
- 91 dB(A) for 2 hours
- 94 dB(A) for 1 hour

A rivet gun, an air chisel or a running APU comfortably exceeds the top of that list, so an hour of riveting can do as much damage as a whole shift on the hangar floor — and it is the short, "not worth putting the defenders on" jobs that quietly build the dose.

Action Values and What Each One Triggers

European noise legislation sets three figures, expressed as a daily or weekly personal exposure, and each one triggers a different obligation:

Level	Daily exposure	Peak sound pressure	What it requires
Lower action value	80 dB(A)	135 dB(C)	Assess exposure; provide information and training; make hearing protection available on request
Upper action value	85 dB(A)	137 dB(C)	Reduce exposure by technical and organisational means; hearing protection becomes mandatory; mark hearing protection zones; provide health surveillance
Exposure limit value	87 dB(A)	140 dB(C)	Must not be exceeded — and is measured *inside* the protector, so the attenuation counts towards it

The last row is the one that is most often misread. The limit value is not "85 plus a bit of tolerance": it is an absolute ceiling on what actually reaches the ear, so the protection worn is part of the calculation. The peak column exists separately because impulse noise — a rivet gun, a pneumatic chisel, a cartridge tool — can injure the ear in a single event regardless of how quiet the rest of the day was, and an average measured over eight hours hides that completely.

Choosing and Wearing Hearing Protection

Hearing protection is PPE and therefore the last line, which means the first question is always whether the noise can be reduced at source or the person moved away from it — a quieter tool, damping, a screen, a longer air hose so the operator is not standing beside the compressor. What is left is then attenuated:

- **More attenuation is not automatically better.** Over-protection cuts the wearer off from speech, from alarms, from the sound of a reversing tug and from the change in an engine's note that tells them something is wrong — and the usual result is that the protector gets lifted off at exactly the wrong moment. The aim is to bring the level at the ear down to a comfortable working range, not to the lowest number available.
- **Fit decides the real performance.** The attenuation printed on the box is achieved in a laboratory. In practice, disposable foam plugs deliver a fraction of it unless they are rolled down and inserted correctly with clean hands, and ear defenders lose a great deal to spectacle side-arms, long hair, hoods and a headband stretched over a hard hat.
- **Match the type to the job.** Defenders are quick to put on and take off for intermittent noise and are easy to supervise, but they are hot and awkward in a confined space; plugs suit long continuous exposure and tight access; communication headsets combine protection with intercom for ground running and towing.
- **Hygiene and replacement.** Foam plugs are single use, reusable plugs are washed, and defender cushions harden with age and stop sealing — a set with cracked cushions is decoration.

Where the upper action value is exceeded the area is designated and signed as a hearing protection zone, and the requirement inside it is absolute rather than a matter of personal choice: hearing loss accumulates across a career, and the engineer who takes their defenders off "for a minute" to hear a colleague is taking the dose that mattered. Health surveillance — audiometry — exists to catch the early loss the individual cannot yet notice.

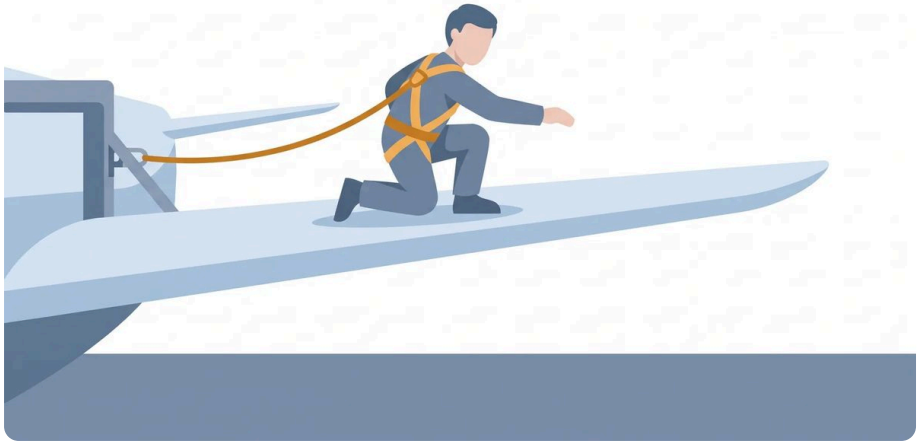
Working at Height

- Use approved platforms, docking systems, and scaffolding. Never stand on boxes, drums, or improvised platforms.
- Use fall-arrest harnesses when working on top of the fuselage, wings, or tail.
- Ensure platforms have guard rails, toe boards, and non-slip surfaces.

Falls are among the largest causes of serious and fatal injury in maintenance, and aircraft work presents an unusually awkward version of the problem: curved, slippery surfaces with no natural edge protection, few designated anchor points, and access equipment that has to be positioned close to expensive structure without touching it.

When does "working at height" start? The answer depends on where you are. European work-at-height law sets *no* minimum height: it applies to any place from which a person could fall a distance liable to cause personal injury, which includes falling into an open fuel tank or through an unguarded floor opening, and even a fall of less than a metre onto a hard hangar floor or into machinery qualifies. Other jurisdictions set a numeric trigger instead, and the figure most commonly quoted is 1.8 m, or 6 ft. Either way the requirement comes from the applicable regulation and not from the height of the aircraft, from a supervisor's opinion or from the individual's confidence — and it applies to a narrow-body wing exactly as it does to a wide-body crown.

The hierarchy of controls from the first section maps directly onto height work, and the order is worth following deliberately because the natural instinct is to start at the bottom of it and reach for a harness:



Restraint: the lanyard is set so short that the wearer physically cannot reach the edge of the wing.

- **Avoid the work at height.** Can the component be removed and worked on at floor level? Can the inspection be done from a docking floor rather than from the top of the fuselage?
- **Prevent the fall** using a guarded working platform — the collective measure that protects everybody on it without any action on their part.
- **Prevent the fall from a personal anchor** using a *restraint* system: a lanyard adjusted so the wearer physically cannot reach the edge. Nobody falls, so no clearance and no rescue are required. Where it is workable, restraint is always better than arrest.
- **Arrest the fall** with a harness and energy-absorbing lanyard or an inertia reel, accepting that a fall will happen and planning for it.
- **Mitigate the consequences** — nets, airbags, exclusion of people from underneath.

Fall arrest carries two obligations that get overlooked. The first is **clearance**: the total fall distance is the free fall before the lanyard becomes taut, plus the extension of the energy absorber as it deploys, plus the height of the person below the attachment point, plus a safety margin. Added up, a two-metre lanyard needs several metres of clear space below the anchor, and an anchor at foot level roughly doubles the free fall compared

with one overhead — so on a low working level a fall-arrest lanyard can let the wearer reach the floor before the system has finished doing anything at all. The second is **rescue**. A person hanging motionless in a harness can lose consciousness within minutes as blood pools in the legs and returns poorly to the heart, so a plan and the means to recover them quickly must exist before the work starts. "Call the fire brigade" is not a rescue plan.

Anchorage is the other half of the system and is aircraft-specific. Harnesses are attached only to the anchor points, running lines or docking structures designated for the purpose in the maintenance manual or by the equipment supplier — never to a hydraulic pipe, an antenna, a duct, a control cable or a convenient bracket, none of which is rated for the several kilonewtons an arrested fall applies, and some of which will simply come off in your hand.

Access equipment brings its own rules. Platforms and docking are positioned with the brakes applied or the wheels chocked, are not moved with people, tools or parts on them, and carry a stated safe working load that includes the equipment as well as the people. Gaps between the platform edge and the aircraft skin are closed or guarded, both to stop a person or a tool falling through and to protect the skin. Outdoor work has a wind limit, and the equipment supplier states it. Ladders and stepladders are a last resort for short-duration, light work only, used at about a 1-in-4 lean — roughly 75 degrees — secured at the top or footed, with three points of contact maintained and never used as a working platform or leaned against the aircraft skin.

Finally, working at height creates a hazard for the people underneath. Tools are tethered or kept in a closed bag, loose items come out of pockets before climbing, the area below is excluded or barriered, and anyone who has to work there wears head protection. A dropped spanner from wing height is capable of killing someone standing below, and it is also the beginning of a foreign object damage investigation.

Confined Spaces

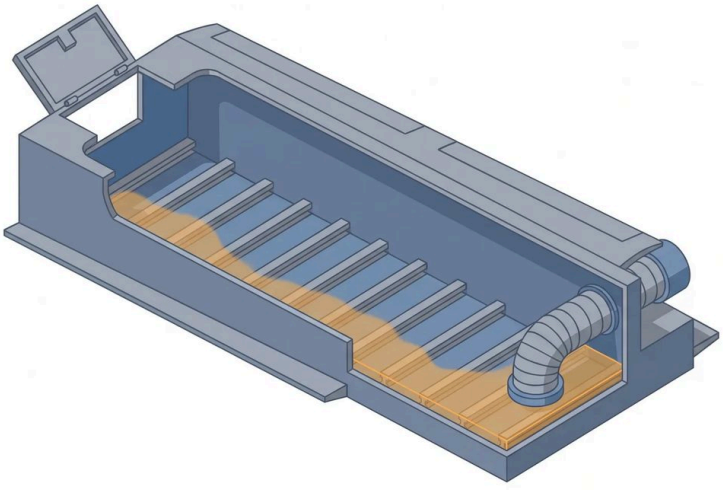
Fuel tanks, cargo compartments, and wheel wells are confined spaces. Entry requires a **permit to work**, atmospheric monitoring (oxygen level, flammable gas concentration), a safety attendant outside, and rescue equipment on standby.

A confined space is not defined by being small. It is a substantially enclosed space in which a *specified risk* exists — oxygen deficiency or enrichment, a flammable atmosphere, a toxic atmosphere, engulfment by a liquid or a free-flowing solid, or excessive heat. A large cargo hold with an inerting system discharged into it is a confined space; a cramped avionics bay with normal ventilation and no specified risk is not. That distinction is what determines whether the permit regime applies.

Fuel tank entry is the case that combines almost every hazard in this note, and the preparation sequence has a logic that is worth learning as a sequence rather than a list:

- **Defuel and drain.** Remove the bulk fuel, then drain the sumps and low points. Residual fuel clinging to structure and trapped in stringer bays continues to evaporate long after the tank reads empty, so "drained" is a starting condition and not a safe one.
- **Open up for through ventilation.** Access panels are opened at more than one point so that air can flow across the space rather than swirl in one corner. Kerosene vapour is heavier than air and pools at the lowest point of the tank, which is where the extraction duct goes.
- **Purge.** Force air through the tank, or use the manual's inert-gas purge. If a nitrogen purge is used to remove the flammable atmosphere it must itself be purged out with air afterwards — a nitrogen-purged tank is safe from fire and lethal to enter, and confusing the two states is a fatal error rather than a paperwork one.
- **Test the atmosphere,** with a calibrated instrument, in the order oxygen, then flammable gas, then toxic contaminants — oxygen first because the readings of the other sensors depend on the oxygen concentration being normal. Test at high, middle and low level, because a heavier-than-air vapour will sit undetected at the bottom while the reading taken at the hatch is clear.
- **Ventilate continuously while the space is occupied,** and re-test after any break in the work, after any change in conditions, and at the intervals the permit specifies. Vapour continues to come out of sealant, out of the tank coating and out of pockets of trapped fuel for as long as anyone is in there.

The atmospheric criteria are conventional and worth memorising: oxygen between 19.5% and 23.5% by volume — the upper limit matters as much as the lower, because an enriched atmosphere makes clothing and hair flammable — and flammable vapour typically below 10% of the lower explosive limit. The 10% figure is a wide safety margin over the concentration that would actually burn, and it exists because a gas detector samples one point at one moment while the space contains gradients and pockets.



Kerosene vapour is heavier than air: it pools at the tank's lowest point, where the extraction duct goes.

Everything taken into the space is chosen with the flammable atmosphere in mind. Lighting and tools are explosion-protected and low voltage, or air-driven; ordinary torches, mobile telephones, cameras and personal

electronic devices stay outside. Clothing is cotton or approved antistatic rather than nylon, footwear is non-sparking, and jewellery and anything that can snag or be dropped comes off. Skin is protected against fuel and sealant, and where the manual requires it, breathing air is supplied from outside rather than filtered — a filter cartridge cannot make oxygen.

The human arrangements matter as much as the equipment:

- The **standby attendant** stays outside for the whole duration, maintains continuous communication with the person inside, controls who enters, keeps the entry log, and never leaves the position or enters the space themselves — the reason for that last rule is set out in the first-aid section above.
- An **entry and exit log** records who is inside at any moment, so that an evacuation or a rescue starts from a known number rather than a guess.
- The person inside wears a **harness with a retrieval line** where the geometry allows recovery from outside, and the rescue equipment and plan are in place *before* the first entry.
- **Work periods are limited and rotated.** Tank work is hot, cramped and disorienting; heat stress and fatigue degrade judgement in exactly the environment where judgement is needed, and dehydration builds quickly inside a wing in summer.
- The **permit** ties all of it together: the isolations, the gas test results and their time, the ventilation, the equipment permitted, the people, the validity period and the hand-back. Work stops when the permit expires, and it is re-tested and re-issued rather than extended by verbal agreement.

One last point that catches people out: the hazard does not end when the work does. A tank that has been open, ventilated and worked in is re-sealed with fresh sealant, which continues to give off solvent for a considerable time, and the tank is not treated as a normal enclosed volume again until the manual's cure and ventilation requirements have been met.

Foreign Object Debris (FOD)

FOD is any object or substance in the aircraft environment that could damage the aircraft or injure personnel. Tools, fasteners, safety wire, rags, and personal items must be accounted for at all times. Many organisations use **tool shadow boards** and **FOD walks** to maintain clean working areas. Lost tools require an aircraft search and documentation before the aircraft can return to service.

The three letters are used for two different things and it is worth keeping them apart: foreign object *debris* is the loose material, and foreign object *damage* is what it does. Debris is generated by the work itself far more often than it blows in from outside — rivet tails and drill swarf, lockwire offcuts, cable-tie tails, blanking caps, split pins, washers, sealant cartridge nozzles, packaging, gloves, ear plugs, pens, identity cards, torch batteries and coins out of a pocket. Outside contributions include stones, ramp fittings, catering and baggage-handling hardware, and ice shed from an aircraft.

The damage is not confined to engines, although an intake ingestion is the most expensive single outcome. Loose articles jam control runs and cable pulleys, foul flap and slat tracks, short across electrical connections and busbars, block drain paths so that water collects where it should not, cut tyres, and become high-velocity projectiles in jet blast or propeller wash. A washer dropped into an open panel and not recovered may do nothing for a year and then find its way somewhere it matters.

Tool Control

Tool control is a formal system, not tidiness, and its whole design goal is to make a missing item *visible immediately* rather than discovered later:

- **Shadow boards and foam-cut trays** show an empty outline the moment a tool is out, so a gap is obvious at a glance to anybody walking past, not just to the person who owns the box.
- **Individual identification.** Every tool is etched, engraved or laser-marked with the box or organisation number, so a tool found in an aircraft can be traced to a task and a person, and so that a tool from another box is not signed back into yours to fill the gap.
- **Issue and return control.** Tools are booked in and out, whether by a store, a tally system or a digital record, and the box is inventoried at the end of the task and at shift handover.
- **Count before you close.** The critical check is a tool and equipment count immediately before any panel, cowling, access door or tank is closed. Once the panel is on, finding a missing spanner means taking it off again — and that is the cheap version of the outcome.
- **Special-purpose and hired-in equipment** is controlled the same way. Borrowed tooling, calibrated equipment and consumable-fitted items such as torque adapters are frequently outside the normal box system, which is exactly why they get left behind.

When a tool cannot be accounted for, the response is escalation rather than an assumption that it will turn up. Work stops, the area and the aircraft are searched, the supervisor is informed, and if the item is still not found the aircraft cannot simply be signed off: the loss is recorded, the possible locations are assessed against the work carried out, and the decision on whether the aircraft can be released — and under what restrictions — is a documented one taken by the responsible person. Concealing a lost tool is a disciplinary matter precisely because the person who knows where it might be is the only person who can find it cheaply.

Consumables, Housekeeping and Checks

Tools are only part of the problem, and consumables are the part that generates the most debris:

- Drilling into a structure or over an open bay is done with a chip catcher, a vacuum attachment or a temporary catchment, and the swarf is cleared and counted out of the area rather than left to migrate.

- Lockwire offcuts and rivet tails go straight into a pocketed pouch or a bag as they are produced; a rivet tail dropped into a wing rib is effectively unrecoverable.
- Blanks, caps and protective covers are treated as controlled items, tallied like tools, and removed to a count.
- Pockets are emptied before working over an open panel, an intake or a fuel tank, and loose personal items — badges on lanyards, phones, pens, jewellery — are secured or left behind.
- The work area is cleaned as the job proceeds rather than at the end, and the bay is swept and inspected before the aircraft is moved.
- Organised **FOD walks** across the ramp and hangar floor, FOD bins at the exits from working areas, and magnetic sweepers for the ramp catch what individual discipline missed.

A final inspection of the intake, the exhaust, the surrounding surface and the area upwind is carried out before every engine ground run, and any panel or cowling opened during the work is confirmed closed and latched. It is worth noticing that the FOD check and the tool count are two different checks with two different purposes — one asks "is the area clear?" and the other asks "do I still have everything I brought?" — and passing one does not substitute for the other.

> ✈ **Aviation Context:** FOD damage costs the aviation industry an estimated $4 billion per year. A single bolt left in an engine intake can destroy turbine blades worth hundreds of thousands of dollars. Every tool should be accounted for — shadow boards, etched markings, and calibrated tool inventories are standard practice.

Manual Handling and Lifting Equipment

Musculoskeletal injury is the largest single cause of lost working days in maintenance, and unlike most of the hazards in this note it accumulates: it is the twentieth awkward lift rather than the first that does the damage. The assessment considers four things — the **task** (how far, how often, twisting, reaching above shoulder height or below knee height, holding away from the body), the **individual** (strength, height, training, pregnancy, existing injury), the **load** (weight, size, rigidity, grip points, whether the contents shift) and the **environment** (floor condition, lighting, space, temperature, working on a stand or in a tank).

The hierarchy applies here exactly as it does elsewhere, and it is routinely skipped. The first question is whether the load has to be moved by hand at all: a trolley, a hoist, a component dolly or the type-specific handling fixture called up in the maintenance manual is an engineering control and outranks any amount of lifting technique, which is only an administrative one. Where a manual lift is unavoidable, the technique is the familiar one — assess and plan the route first, feet apart, bend the knees and keep the back's natural curve, load close to the body, no twisting, lower by reversing the process — and a team lift has one person calling the moves so that everybody lifts and lowers together.

Mechanical lifting gear brings its own regime:

- Every item — sling, chain, shackle, eyebolt, hoist, crane, jack — is marked with a unique identification and its **safe working load**, and it is never loaded beyond it.
- It is inspected by the user before every use for cuts, broken wires, distortion, corrosion, damaged stitching and illegible markings, and it is subject to periodic thorough examination by a competent person, with a current certificate. The interval is set by the applicable national regulation or by a written scheme of examination; under the United Kingdom regulations, for example, it is at least every six months for lifting accessories and for any equipment used to lift people, and at least every twelve months for other lifting equipment.
- A load is never suspended over people and never left suspended unattended; tag lines are used to control swing rather than hands.
- The lift is planned: the weight is known rather than estimated, the centre of gravity is identified, and the attachment points are the ones the manual designates — not a convenient lug.

Worked example — why sling angle matters.
A 1,000 kg component is lifted on a two-leg sling. Each leg carries half the weight only if the legs are vertical. As the legs are spread, the tension in each rises:

$$T = \frac{W/2}{\cos\theta}$$

where θ is the angle of each leg from the vertical, so the included angle between the legs is 2θ.

- Included angle 60 degrees, so $\theta = 30^\circ$: $T = 500/0.866 = 577$ kg-force in each leg.
- Included angle 120 degrees, so $\theta = 60^\circ$: $T = 500/0.5 = 1000$ kg-force in each leg.

At a 120 degree spread each leg is carrying the *whole* weight of the load, and a pair of slings rated at 750 kg each — which looks like a comfortable margin on a 1,000 kg lift — is now overloaded by a third. Spreading the legs to clear an obstruction is the commonest way lifting gear is unknowingly overloaded, and it is why the marked safe working load of a multi-leg sling is quoted against a stated angle.

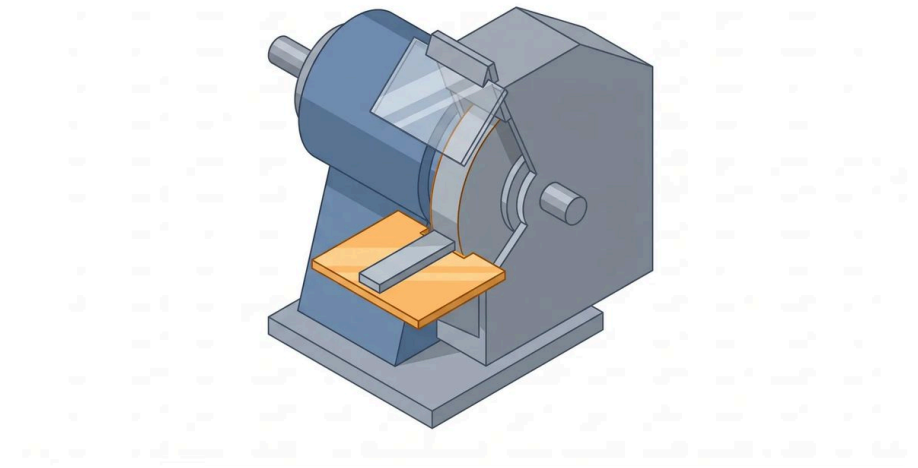
Machinery and Workshop Equipment

Guards and interlocks are engineering controls, which is why defeating one is treated so seriously: it does not reduce the level of protection by a step, it removes the tier of the hierarchy the machine was designed around and leaves nothing but the operator's attention. A machine with a guard removed, an interlock bridged or a trip bypassed is taken out of service and tagged, not used carefully.

- **Rotating machinery.** No gloves, no loose clothing, no cuffs, no ties, no lanyards, no rings or wristwatches, and long hair tied back and covered. The workpiece is clamped rather than held; swarf is cleared with a brush or hook after the machine has stopped, never with fingers and never while it is running.
- **Abrasive wheels.** A grinding wheel is a pressure vessel of sorts — it stores rotational energy and fails explosively. Wheels are mounted only by a person trained and authorised to do so, after checking that the wheel's maximum speed is not less than the machine's spindle speed and after a ring test for cracks; the

machine is then run up to speed from behind cover before use. In service, the work rest is kept adjusted close to the wheel so that the workpiece cannot be dragged into the gap, the guard and eye shield stay in place, and eye protection is worn regardless of the shield. Grinding on the *side* of a straight wheel not designed for side loading is a classic cause of wheel burst.

- **Portable power tools.** Inspected before use, protected by a residual current device, cables routed clear of walkways and sharp edges, and disconnected from the supply — or the air line — before changing a bit, blade or disc.
- **Hot processes.** Grinding, cutting and drilling produce hot chips and sparks that travel a surprising distance, so the hot-work precautions described earlier apply whenever this work is done near an aircraft, and screens are used to protect people and structure nearby.



Bench grinder: guard and eye shield in place, work rest set close so the workpiece cannot be dragged in.

Slips, Trips and Housekeeping

Slips and trips produce more workplace injuries than any other single mechanism, and in a hangar they have the additional property of putting somebody on the floor next to whatever they were carrying, or off the edge of a stand. The causes are almost always mundane and almost always fixable at the moment they arise:

- Fluid on the floor. Hydraulic fluid, oil and de-icing fluid on a painted or sealed hangar floor produce a surface with virtually no friction, and they spread further than they look. Spills are contained and cleaned immediately, not marked with a cone and left for the next shift.
- Trailing cables and air hoses. Route services overhead where possible, along walls where not, and across walkways only in a cable ramp.
- Clutter and changes of level. Packaging, offcuts, jack pads, chocks and toolboxes left in the walking route; unmarked steps, kerbs and pit edges; open floor panels and inspection hatches, which are guarded or covered as soon as they are opened.

- Poor lighting. Task lighting is brought to the job rather than working in the shadow of the aircraft, and general lighting failures are reported instead of tolerated.
- Footwear. Non-slip soles, in good condition, are part of the control and not a substitute for the other four.

"Clean as you go" is the operative principle, and it does double duty: the same discipline that keeps the floor safe to walk on keeps loose material out of the aircraft.

Hand-Arm Vibration

Riveting hammers, air chisels, needle scalers, die grinders and impact wrenches transmit vibration into the hands, and prolonged exposure causes hand-arm vibration syndrome: permanent damage to the small blood vessels and nerves of the fingers. The early sign is blanching of the fingertips triggered by cold, accompanied by tingling and numbness; as it progresses, grip strength and fine dexterity are lost — which for an aircraft engineer is an occupational disability as well as a personal one. Like noise damage, it is cumulative, painless in its early stages and irreversible.

European vibration legislation follows the same two-threshold pattern as noise, expressed as a daily exposure averaged over eight hours: an **exposure action value of 2.5 m/s²**, at which the employer must assess and control the risk and provide health surveillance, and an **exposure limit value of 5 m/s²** which must not be exceeded. Because the figure is a daily dose, a high-vibration tool used for a short time can be equivalent to a low-vibration tool used all day — which is why the control is usually a combination of choosing lower-vibration tooling, keeping tools maintained and their consumables sharp, limiting trigger time, rotating the task between people, and keeping the hands warm and dry, since cold makes the vascular symptoms markedly worse.

Working Conditions

The physical conditions of the job are a safety control in their own right, because they act on the person doing it. Heat stress builds quickly inside a wing tank, on a ramp in summer or under a hangar roof, and it degrades concentration and decision-making before the person notices they are affected; the counter-measures are rotation, rest breaks out of the heat and access to drinking water. Cold has the mirror effect, reducing dexterity and grip on a winter line station and making a vibration exposure worse at the same time. Lighting has to be adequate for the inspection standard being applied, not merely for finding the fastener — a crack that could have been found under proper light is not "not there".

Environmental limits also apply directly to the work itself. Painting, sealing, bonding and composite repair all have temperature and humidity windows specified in the manual, outside which the material will not cure or bond correctly. A task carried out outside those limits produces a defect that may not be visible on completion, so the correct response to a hangar that is too cold, too damp or too dusty is to stop and reschedule the task rather than to press on and hope — which is the same conclusion, reached from the technical side, as the one the working-environment requirement reaches from the regulatory side.

Module 8 — Basic Aerodynamics

8.1 — Physics of the Atmosphere

Introduction

Every aircraft flies within the Earth's atmosphere, and the behaviour of that atmosphere — its temperature, pressure, and density — has a **direct and profound effect** on aircraft performance, engine power output, instrument readings, and aerodynamic forces. An aircraft that produces 250 kN of thrust at sea level will produce significantly less at 35,000 ft because the air is thinner. An airspeed indicator calibrated for sea-level conditions will read differently at altitude unless corrections are applied. For these reasons, a thorough understanding of atmospheric physics is essential for every aircraft maintenance engineer.

It is worth being precise about what "the atmosphere" means to an engineer, because the word does a lot of work. To a meteorologist it is a weather machine. To a maintenance engineer it is a **working fluid** — a substance with measurable properties that the aircraft continuously uses, and whose properties change from hour to hour and from one end of a climb to the other. Four separate jobs depend on it, and every one of them degrades in the same direction when the air thins:

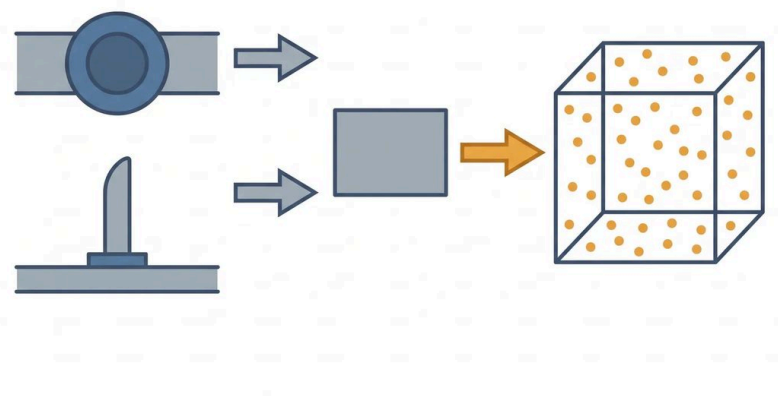
- **It is the source of lift and drag.** The wing does not "hold the aircraft up" by itself; it deflects a stream of air, and the reaction to that deflection is lift. Both the lift and the drag equations contain air density as a straight multiplier, so the aerodynamic forces are proportional to how much mass the wing has to work on.
- **It is the working fluid and the oxidiser of the engine.** A gas turbine is an air pump before it is anything else, and a piston engine fills its cylinders from the same source. Thrust and power both follow the *mass* of air handled per second, not the volume, so an engine breathing thin air is handicapped even though the volume flow through it is unchanged.
- **It is the reference the flight instruments measure against.** The altimeter, the airspeed indicator, the vertical speed indicator and the Mach meter are all pressure instruments. They do not measure height or speed at all; they measure pressures and convert them using an assumed model of the atmosphere. When the real atmosphere differs from the model, the instrument is not faulty — it is answering a different question from the one the pilot asked.
- **It is the supply for the cabin and for many aircraft services.** Pressurisation air, air-conditioning, wing and engine anti-icing, and on many types the hydraulic reservoir head and the potable-water pressurisation all come from bleed air taken from the compressor. The mass available to bleed falls with altitude for exactly the same reason thrust does.

The four properties, and which one actually appears in the equations

An engineer needs to be fluent in four atmospheric quantities. It is worth separating them at the outset, because students routinely treat them as interchangeable and the exam does not:

Property	Symbol	SI unit	What it physically is	How it is obtained on the aircraft
Static pressure	p	Pa (N/m²)	The force per unit area exerted by the weight of the air column above, acting equally in all directions	Measured directly, at the static ports
Temperature	T	K	A measure of the average kinetic energy of the air molecules	Measured directly, at the total air temperature probe
Density	ρ (rho)	kg/m³	The mass of air contained in unit volume	**Never measured directly** — always computed from pressure and temperature
Humidity	—	several	The quantity of water vapour mixed into the air	Not sensed by the air data system at all; allowed for by rule of thumb

The fourth row of that table is the one that catches people out, and the third row is the one that matters most. **Density is the property the aerodynamics and the engine actually care about, and it is the one property no aircraft instrument measures.** There is no "densitometer" in the pitot-static system. Density is always inferred, through the equation of state, from a pressure that is measured and a temperature that is measured. That single fact explains a large part of this paragraph of the syllabus: the reason so much attention is paid to pressure and temperature is that they are the two measurable quantities from which the important one is derived.



Density is never measured: it is computed from the measured static pressure and temperature.

The dependency chain to memorise: pressure and temperature are the *inputs*; density is the *output*; lift, drag, thrust, power and true airspeed are the *consequences*. Most atmospheric questions in the examination are questions about one link in that chain. If you can say which link a question is asking about, the answer usually follows.

Units: the same quantity in five different clothes

Aviation is unusually careless about units, largely for historical reasons: pressure is quoted in millibars by European meteorologists, in inches of mercury by North American ones, in pounds per square inch on hydraulic and pneumatic gauges, and in pascals in the certification documents. A maintenance engineer meets all of them, sometimes on the same aircraft, and the examination tests whether you can move between them. The conversions below are worth committing to memory, because a question that looks like a physics question is frequently a units question in disguise.

Quantity	Units you will meet	Standard sea-level value expressed in each	Conversion
Pressure	hectopascal (hPa), millibar (mb), pascal (Pa)	1013.25 hPa = 1013.25 mb = 101,325 Pa	1 hPa = 1 mb = 100 Pa exactly
	inch of mercury (in Hg)	29.92 in Hg	1 in Hg = 33.86 hPa
	millimetre of mercury (mmHg, torr)	760 mmHg	1 mmHg = 1.333 hPa
	pound per square inch (psi)	14.696 psi	1 psi = 68.95 hPa
	bar, atmosphere (atm)	1.01325 bar = 1 atm	1 bar = 1000 hPa
Temperature	degree Celsius (°C)	15 °C	—
	kelvin (K)	288.15 K	K = °C + 273.15
	degree Fahrenheit (°F)	59 °F	°F = (°C × 9/5) + 32
Speed	knot (kt), nautical mile per hour	—	1 kt = 1.852 km/h = 0.5144 m/s
	statute mile per hour (mph)	—	1 kt = 1.1508 mph
Distance	nautical mile (NM), statute mile, kilometre	—	1 NM = 1.852 km = 6,076 ft = 1.1508 statute miles

Worked units conversion: an aircraft is flying at a true airspeed of 720 knots. Express that in statute miles per hour.

A nautical mile is longer than a statute mile — 1,852 m against 1,609 m — so the number of statute miles covered per hour must be *larger* than the number of nautical miles. The factor is

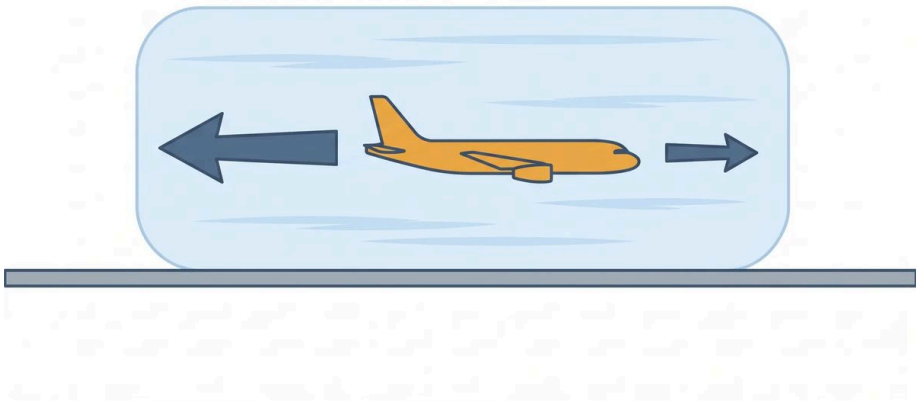
$$1852 \div 1609.344 = 1.1508$$

so $720 \times 1.1508 = 828.6$ mph, or about 829 mph.

The direction check that stops the classic error: because a statute mile is the *shorter* unit, it takes more of them to cover the same ground, so the mph figure must always exceed the knots figure. Multiplying by a factor smaller than one converts the other way, statute to nautical; the 0.83 offered as a distractor in this conversion comes in two forms, and the direction check settles only one of them. Multiplying by 0.83 would have produced 598 — a number smaller than the knots value, which is immediately recognisable as wrong without doing any arithmetic at all. Dividing by 0.83 gives 867, which is larger than 720 and so survives the direction check; that form has to be rejected on size instead. The correct reverse factor is $1 \div 1.1508 = 0.8690$, not 0.83, and dividing by that returns the same 829 mph.

Airspeed, ground speed and the air mass

One distinction belongs here rather than later, because it underlies everything said below about airspeed. **An aircraft flies through an air mass, and that air mass is itself moving over the ground.** True airspeed is the aircraft's speed relative to the air around it; ground speed is its speed relative to the earth. The two differ by the wind, and the aerodynamics only ever respond to the first of them.



The aircraft flies through an air mass that itself moves over the ground; the difference is the wind.

Worked example: an aircraft is flying at 350 mph through the air, directly into a 75 mph headwind. What is its ground speed?

The air mass is moving backwards over the ground at 75 mph relative to the aircraft's direction of travel, so the speed made good over the ground is

$350 - 75 = 275$ mph.

With a 75 mph tailwind instead, the same aircraft would make 425 mph over the ground — but in *both* cases the wing, the pitot tube and the engine see exactly the same 350 mph of airflow, so the lift, the drag, the indicated airspeed and the thrust are all identical. Wind changes the navigation problem and the fuel burn per mile; it does not change the aerodynamics.

A persistent misconception: that flying into a headwind somehow "loads" the wing or increases the stalling speed. It does not. In steady flight the aircraft is simply part of the moving air mass, and a constant wind of any strength is aerodynamically invisible. Only a *change* of wind — wind shear, or the wind gradient on the approach — produces a real aerodynamic effect, and it does so precisely because the aircraft's inertia prevents it from adopting the new air-mass velocity instantly.

Composition of the Atmosphere

The Earth's atmosphere is a mixture of gases held in place by gravity. By volume, dry air at sea level consists of:

Gas	Chemical Formula	Volume %
Nitrogen	N_2	78.08 %
Oxygen	O_2	20.95 %
Argon	Ar	0.93 %
Carbon dioxide	CO_2	0.04 %
Trace gases	Ne, He, CH_4, etc.	< 0.01 %

This composition remains remarkably constant up to about 80 km altitude. The two dominant gases — nitrogen and oxygen — together make up 99 % of the atmosphere. Although oxygen is the gas essential for both human life and combustion in aircraft engines, nitrogen plays an important role as an inert dilutant that moderates the chemical reactivity of pure oxygen.

In addition to these fixed gases, the atmosphere contains variable amounts of **water vapour** (0 – 4 % by volume). Water vapour is critically important for aviation because:

- **Water vapour is less dense than dry air** — the molecular weight of water (H_2O = 18) is lower than that of nitrogen (N_2 = 28) or oxygen (O_2 = 32). When water vapour displaces these heavier molecules, the resulting humid air is *less dense* than dry air at the same temperature and pressure.
- **Lower density reduces performance** — less dense air means less aerodynamic lift for a given airspeed and angle of attack, and less mass of air entering the engine per second, reducing thrust or power output.

- **Water vapour causes weather** — condensation of water vapour produces clouds, rain, snow, ice, and fog — all of which affect aviation operations.
- **Water vapour can cause icing** — when moisture in the atmosphere freezes on aircraft surfaces, it degrades aerodynamic performance and can block pitot tubes and engine intakes.

Percentage by volume is not percentage by mass

The figures in the table above are **by volume**, and that is the convention the examination uses. For a mixture of gases at the same temperature and pressure, percentage by volume is identical to percentage by *molecule count*, because Avogadro's principle says equal volumes of any gas at the same temperature and pressure contain equal numbers of molecules. So "78.08 % nitrogen by volume" also means "78.08 of every 100 air molecules is a nitrogen molecule".

Percentage by **mass** is different, because the molecules do not all weigh the same. Weighting each gas in the table above by its molar mass, and comparing the result against the volume figures listed there, gives:

Gas	Molar mass (g/mol)	% by mass	How that compares with its share by volume
Nitrogen	28.01	75.5 %	**Lower** — a nitrogen molecule is lighter than the average air molecule, so it carries less than its numerical share of the mass
Oxygen	32.00	23.1 %	**Higher** — an oxygen molecule is heavier than average
Argon	39.95	1.3 %	**Higher** — an argon atom is about 38 % heavier than the average air molecule, so it carries appreciably more than its numerical share of the mass
Carbon dioxide	44.01	0.06 %	**Higher**, and by the largest proportional margin of the four, carbon dioxide being the heaviest of them
Dry air as a whole	**28.96**	100 %	—

Oxygen's share rises from 20.95 % by volume to 23.1 % by mass because an oxygen molecule is heavier than the nitrogen molecule it is being compared with. The mean molar mass of dry air, 28.96 g/mol, is the single number from which the specific gas constant for air is derived, and it reappears later in this note in the equation of state.

For examination purposes the composition is often quoted as the simple fractions **one fifth oxygen and four fifths nitrogen**. That rounding is close enough for reasoning: one fifth is 20 % against a true 20.95 %, and four fifths is 80 % against a true 78.08 %. Note which way round the fractions go — nitrogen is the large fraction. Reversing them, so that oxygen becomes four fifths, is a common careless error and gives an atmosphere that would make almost anything combustible burn violently.

Argon, carbon dioxide and the variable gases

Argon is worth noticing because it is the **third most abundant gas** in the atmosphere — more than twenty times as abundant as carbon dioxide — and yet it is almost never mentioned. It is a noble gas, chemically

inert, and takes no part in any atmospheric process, which is precisely why it is ignored.

Carbon dioxide, at about 0.04 % by volume, is a trace gas by quantity but not by effect: it is a strong absorber of infrared radiation and therefore matters to the atmosphere's heat balance out of all proportion to its concentration. Unlike nitrogen, oxygen and argon it is not truly fixed — it varies seasonally and has been rising over the long term — so it sits at the boundary between the fixed and variable gases.

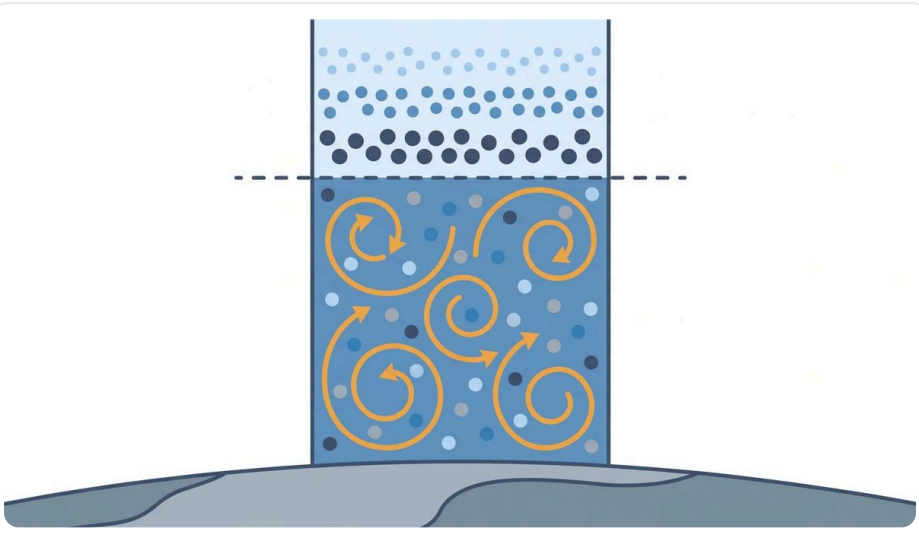
Two other constituents are genuinely variable and are excluded from the "dry air" table for that reason. The first is **water vapour**, dealt with in its own section below. The second is **ozone (O_3)**, a three-atom form of oxygen that is negligible near the surface but is concentrated in a broad band in the stratosphere. Ozone is chemically aggressive and is produced from ordinary oxygen by ultraviolet radiation; the same absorption that creates it is what warms the upper stratosphere, and that is dealt with in the section on the layers.

Why the proportions stay the same all the way up

A natural expectation is that the heavier gases should settle out, leaving the upper atmosphere progressively richer in the light ones. That is exactly what happens in a still container of gases left for long enough, and it is what happens in the atmosphere *above* about 80 km. Below 80 km it does not, and the reason is mechanical rather than chemical.

- **Below about 80 km, turbulent mixing dominates.** Convection, wind, wave motion and general stirring move whole parcels of air far faster than individual molecules can diffuse relative to one another. Because a parcel carries all its gases together, stirring homogenises the mixture. This region is called the **homosphere** — "same sphere" — and within it the proportions of nitrogen, oxygen, argon and carbon dioxide are essentially identical at every altitude.
- **Above about 80 km, molecular diffusion dominates.** The air is so thin that a molecule travels a long way between collisions, and the stirring that homogenises the lower atmosphere has nothing left to stir. Each gas then settles according to its own molar mass, so the mixture separates with the heavier gases below and the lighter ones above. This region is the **heterosphere** — "different sphere" — and it is dominated successively by molecular nitrogen, then atomic oxygen, then helium, then hydrogen with increasing altitude.

The boundary between the two, at roughly 80 km, is called the **turbopause**. It is far above the operating altitude of any aircraft, which is why for all aviation purposes the composition of the air can be treated as fixed.



Below about 80 km turbulence keeps the gases mixed; above it, each gas settles by its own molar mass.

What actually changes with altitude is the NUMBER of molecules, not the mixture. The ratio of nitrogen to oxygen at 40,000 ft is exactly what it is on the ground: about 78 to 21. What has fallen is the total number of molecules in each cubic metre. This distinction is one of the most tested points about atmospheric composition, and it has a direct physiological consequence — hypoxia at altitude is *not* caused by the air containing a smaller percentage of oxygen, because it does not. It is caused by the low **partial pressure** of that unchanged 21 %.

Partial pressure — the reason "21 % oxygen" is not the whole story

In a mixture of gases, each constituent contributes to the total pressure in proportion to its share of the molecules. That contribution is its **partial pressure**, and Dalton's law states that the total pressure is simply the sum of the partial pressures of all the constituents. For oxygen:

$$p_{O_2} = 0.2095 \times p_{total}$$

Where p_{total} is the ambient static pressure and 0.2095 is oxygen's fraction by volume.

Because the fraction never changes but the total pressure falls steeply with altitude, the oxygen partial pressure falls in exactly the same proportion as the total:

Altitude	ISA pressure (hPa)	Oxygen partial pressure (hPa)	Fraction of the sea-level value
Sea level	1013	212	100 %
8,000 ft (typical cabin)	753	158	74 %
10,000 ft	697	146	69 %
18,000 ft	506	106	50 %
35,000 ft	238	50	24 %
40,000 ft	188	39	19 %

The 18,000 ft row is the one to remember: it is where the total pressure has halved, so an occupant breathing ambient air there is receiving half the oxygen molecules per breath that the same lungs would receive at sea level. That is why 18,000 ft appears repeatedly as a threshold in both physiology and in the practical rules for altimetry, and it is the reason supplemental oxygen is not optional in an unpressurised aircraft operating anywhere near it.

Nitrogen as an engineering material

Nitrogen's inertness has a direct maintenance consequence that goes beyond its role in the atmosphere. Because it is unreactive, contains no oxygen to support combustion, and can be supplied bone dry, **dry nitrogen is the standard servicing gas** for a range of aircraft systems: main and nose landing gear tyres, oleo struts, hydraulic accumulators and reservoirs, fire-extinguisher bottle superchargers, and the purging of fuel and oxygen system components during maintenance. Compressed air would work mechanically but brings two liabilities with it — the oxygen that supports combustion in a hot brake or a ruptured tyre, and the water vapour that condenses and freezes at altitude. Servicing with nitrogen removes both. The gas in the tyre is therefore chemically the same as the majority constituent of the air around it, just without the 21 % that makes air chemically dangerous and without the water that makes it wet.

What else is in the air: aerosols and contaminants

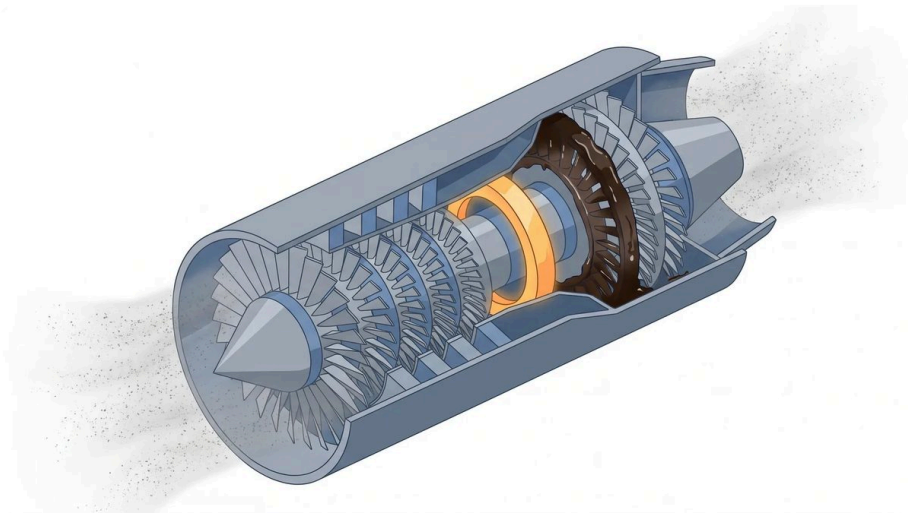
Beyond the gases and water vapour, the atmosphere carries a suspended load of solid and liquid particles collectively called **aerosols**. They are present in quantities far too small to affect density, pressure or any of the calculations in this note, and they are absent from every table of composition — yet several of them are of direct operational significance, and one of them is capable of stopping every engine on an aircraft.

Contaminant	Origin	Where it is encountered	Effect on the aircraft
Volcanic ash	Explosive eruption, injected directly into the troposphere and often into the stratosphere	Ash clouds drift for thousands of miles at cruising levels and are effectively invisible to airborne weather radar	Melts in the combustor and solidifies on turbine nozzle guide vanes, restricting flow and causing surge and flameout; erodes compressor blades and aerofoil leading edges; sandblasts windscreens opaque; blocks pitot and static systems
Sand and dust	Wind lifted from arid surfaces; raised by the aircraft's own downwash or exhaust on unpaved surfaces	Low level in desert regions; dust can be lifted to considerable altitude and carried far	Compressor and turbine erosion, blocked filters, accelerated wear of bearings and seals, and reduced visibility
Salt	Sea spray evaporating and leaving crystals suspended in the air	The marine boundary layer — coastal and over-water operations	A powerful corrosion promoter, particularly with condensation present; the reason for washing and corrosion-prevention programmes on maritime-operated aircraft

Contaminant	Origin	Where it is encountered	Effect on the aircraft
Industrial and combustion pollution	Ground-level sources, concentrated beneath inversions	Low level near urban and industrial areas, especially in stable air	Corrosive deposits and reduced visibility; the accumulation is the direct consequence of the trapping mechanism described later in this note
Condensation nuclei	Salt, dust, combustion products — the same particles listed above	Everywhere in the troposphere	Not a hazard but a necessity: water vapour needs a surface to condense on, and in genuinely clean air it will supersaturate rather than form cloud

The last row is worth pausing on, because it links this list back to the water vapour section that follows. **Cloud, fog and precipitation all depend on the presence of aerosols.** Perfectly clean air can be cooled well past its saturation point without any condensation occurring, and it is the ubiquity of dust and salt particles that keeps atmospheric supersaturation to a fraction of a per cent in practice. The contaminants that cause the engineer so much trouble are also what makes the weather work.

Volcanic ash is the one atmospheric contaminant that is an immediate flight-safety emergency. It cannot be detected by weather radar, because radar responds to precipitation-sized water droplets and ash particles are far too small and too dry. It gives no warning other than possibly a sulphurous smell, static discharge on the windscreen and an engine-related haze in the cabin. And it attacks the engines by a mechanism that gets worse the harder they are worked, because the ash melts in the hottest part of the engine — which is why the standard response includes reducing thrust and leaving the ash cloud by the shortest route rather than climbing through it.



Volcanic ash melts in the combustor and solidifies on the turbine nozzle guide vanes, restricting flow.

Water Vapour and Humidity

Water is the only substance that exists in the atmosphere in all three physical states at ordinary temperatures, and it is the only major constituent whose concentration varies from almost nothing to several per cent within a

few hours and a few hundred miles. Everything the atmosphere does that an aircraft has to be protected against — cloud, precipitation, icing, thunderstorms, fog, reduced visibility — is water changing state. This section deals with how the quantity of water vapour is described, what controls it, and why its effect on air density runs in the opposite direction to almost everyone's intuition.

Water vapour is a gas, and it is invisible

The first thing to be clear about is that **water vapour cannot be seen**. It is a true gas, fully mixed with the nitrogen and oxygen, obeying the same gas laws they do. A cloud, a contrail, steam from a kettle or fog on a runway is *not* water vapour: it is water that has already condensed into suspended liquid droplets or ice crystals and has therefore stopped being vapour. The visible boundary of a cloud is the surface at which condensation begins, not the edge of a body of vapour.

This matters practically. Air can hold a large quantity of water invisibly, and it is the transition — vapour to liquid, or vapour directly to ice — that produces every hazard. The engineer's interest in humidity is therefore mostly an interest in *how close the air is to that transition*.

Saturation and saturation vapour pressure

At any given temperature there is a maximum quantity of water vapour a given volume of air can hold. When that limit is reached the air is **saturated**: molecules are leaving the liquid surface and returning to it at the same rate, and any further vapour added condenses out. The partial pressure exerted by the vapour at that limit is the **saturation vapour pressure**.

The critical property of saturation vapour pressure is that it depends on **temperature alone**, and it depends on it very steeply — roughly doubling for every 10 °C rise. Warm air can hold far more water than cold air:

Air temperature	Saturation vapour pressure	Water held at saturation
0 °C	about 6 hPa	about 4.8 g per cubic metre
15 °C	about 17 hPa	about 12.8 g per cubic metre
30 °C	about 42 hPa	about 30.3 g per cubic metre

Two consequences follow immediately, and both are examined. First, **cooling air is by far the most common way of producing condensation** — not adding water. Air that is nowhere near saturation at 30 °C is well past it at 5 °C without a single extra molecule of water being introduced. Second, the very cold air of the upper troposphere can hold almost no water at all, which is why the stratosphere and the upper troposphere are effectively dry and why virtually all the atmosphere's water vapour is concentrated in the lowest few kilometres.

The five ways of stating how much water is present

Five separate quantities are in routine use, and the examination consistently tests whether you can tell them apart. Three of them are amounts, one is a ratio and one is a temperature. Read the definitions against each other rather than one at a time:

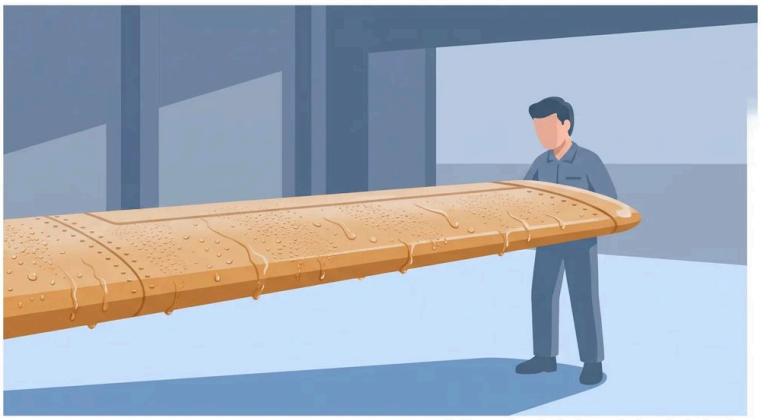
Term	What it states	Typical units	Is it a ratio?
Absolute humidity	The *actual quantity* of water vapour present, expressed per unit **volume** of air	g/m³	No — an amount
Specific humidity	The *actual quantity* of water vapour present, expressed per unit **mass** of moist air	g/kg	No — an amount
Mixing ratio	The mass of water vapour per unit mass of the **dry air** it is mixed with	g/kg	No — an amount
Relative humidity	The vapour *actually present* as a percentage of the maximum the air could hold if **saturated at the same temperature and pressure**	%	**Yes**
Dew point	The **temperature** to which the air must be cooled, at constant pressure, for it to become saturated	°C	No — a temperature

How to answer these questions without hesitating. Ask what kind of thing the answer must be. If the question asks for *the amount of water present*, the answer is **absolute humidity** (per unit volume) or specific humidity (per unit mass) — never relative humidity, which tells you nothing about the amount on its own. If the question uses the word *ratio*, or asks for a comparison with the saturated condition, the answer is **relative humidity**. If the question describes *a temperature the air is cooled to*, the answer is **dew point**. The three are frequently offered together as the only three options, and each is the correct answer to precisely one of these three question shapes.

Why relative humidity on its own tells you nothing about the amount of water

Relative humidity is a percentage of a moving target, because the denominator — the saturation amount — changes with temperature. Air at 0 °C and 100 % relative humidity holds about 4.8 g of water per cubic metre. Air at 30 °C and only 50 % relative humidity holds about 15 g per cubic metre, **three times as much water**, despite the far lower percentage.

This also explains an everyday observation with a maintenance consequence: heating a hangar without adding any moisture drives its relative humidity down sharply, because the amount of water present is unchanged while the amount the air *could* hold has risen. Cooling the same air overnight drives the relative humidity back up, and once it reaches 100 % the surplus condenses onto the coldest available surfaces — airframe skins, control cables, connectors and the inside of unsealed structures. Condensation inside an aircraft is almost always a temperature story, not a leak.



Overnight cooling brings hangar air to 100 % relative humidity, so water condenses on the cold airframe skin.

Dew point and the temperature/dew-point spread

The dew point is the most directly useful of the humidity measures, because it is an absolute statement about the water content expressed in a unit — a temperature — that can be compared directly with the actual air temperature. The gap between the two is the **temperature/dew-point spread**:

- **A large spread** means the air is far from saturation. Cloud, if any, will be high, and fog is unlikely.
- **A small and closing spread** means saturation is near. If the temperature falls to the dew point at the surface, the result is fog or dew; if it does so aloft in rising air, the result is cloud.
- **A spread of zero** means the air is saturated: relative humidity is 100 %, and temperature and dew point are equal.

Because unsaturated air cools at close to the dry adiabatic rate as it rises, while the dew point of that same air falls only slowly, the spread closes with height at a predictable rate. That gives the standard rule of thumb for the base of convective cloud: divide the surface temperature/dew-point spread in degrees Celsius by 2.5 and multiply by 1,000 to obtain the approximate cloud base in feet above the ground. A spread of 10 °C therefore suggests a cloud base of roughly 4,000 ft. It is an approximation and should be treated as one, but it demonstrates the mechanism: cloud base is set by how far the air has to rise before it cools to its own dew point.

Why humid air is LESS dense than dry air

This is the single most counter-intuitive result in the whole of atmospheric physics, and it is examined repeatedly because almost everyone's instinct is the wrong way round. Damp air *feels* heavy, and water is obviously heavier than air, so the expectation is that adding moisture must make the air denser. It does the opposite.

The reason is that water vapour does not *add* to the air — it **displaces** part of it. At a fixed temperature and a fixed total pressure, a given volume contains a fixed number of molecules, whatever those molecules are. That is Avogadro's principle again, and it is the whole of the argument. So every water molecule that enters the volume must push out one molecule of something else, and the something else is nitrogen or oxygen:

Molecule	Formula	Molar mass (g/mol)	Compared with a water molecule
Water vapour	H_2O	18.0	—
Nitrogen	N_2	28.0	56 % heavier than water
Oxygen	O_2	32.0	78 % heavier than water

Each swap therefore replaces a heavier molecule with a lighter one, and the total mass in the volume falls. **The direction is unambiguous: at the same temperature and the same pressure, the more water vapour the air contains, the lower its density.**

Worked example — ranking three mixtures by weight. Consider three samples, each of 100 parts by volume, at the same temperature and pressure. Because equal volumes contain equal numbers of molecules, "parts by volume" is the same as "parts by molecule count", so the mass of each sample is found by weighting the parts by molar mass. Dry air averages 28.96 g/mol and water vapour is 18.02 g/mol:

- 98 parts dry air, 2 parts water vapour: $(98 \times 28.96) + (2 \times 18.02) = 2838 + 36 = 2874$ g
- 50 parts dry air, 50 parts water vapour: $(50 \times 28.96) + (50 \times 18.02) = 1448 + 901 = 2349$ g
- 35 parts dry air, 65 parts water vapour: $(35 \times 28.96) + (65 \times 18.02) = 1014 + 1171 = 2185$ g

The **wettest mixture is the lightest**, at 2,185 g against 2,874 g for the driest. The ranking runs monotonically with water content, so no arithmetic is strictly needed to answer a question of this shape — only the direction of the effect, which is that more vapour always means less mass.

How large is the effect in practice?

Being clear about the direction is not the same as claiming the effect is large. It is real but modest, and it is worth knowing the magnitude so as not to over-state it:

Worked example — saturated air on a hot day. Take sea-level pressure of 1013.25 hPa and an air temperature of 35 °C. Compute the density twice, once for perfectly dry air and once for air saturated at that temperature, using each gas's own partial pressure with its own gas constant (287.05 J/kg·K for dry air, 461.5 J/kg·K for water vapour):

- Dry: the whole 101,325 Pa is dry air, giving $\rho = 1.145$ kg/m³
- Saturated: the saturation vapour pressure at 35 °C is about 5,610 Pa, so 95,715 Pa is dry air and 5,610 Pa is vapour, giving $\rho = 1.122$ kg/m³

The reduction is 0.023 kg/m³, or about **2 %**. Expressed as density altitude, saturating the air adds roughly **700 ft** to an already-high figure — not nothing, but small beside the several thousand feet that the 35 °C temperature itself contributes.

The honest summary is therefore: **temperature is the dominant term, pressure is second, and humidity is a third-order correction** that nevertheless always pushes the same way as the other two on a hot day. That is why performance charts are drawn against pressure altitude and temperature, and why humidity is normally handled by a margin rather than by a chart axis. It is also why the worst-case combination for take-off performance is **high, hot and humid** — the three density-reducing effects reinforcing one another — and why

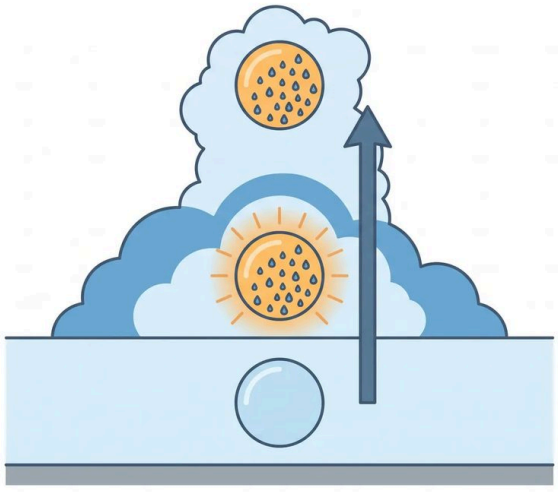
the highest true airspeed at touchdown for a given indicated speed occurs on a hot, humid day rather than a cold, dry one.

A trap in the wording. The distractor to watch for reverses the molecular argument: it explains a claim that humid air is *denser* than dry air by saying the water molecules are heavier than the nitrogen and oxygen they displace. Comparing molecular masses is exactly the right move, but the comparison runs the other way — water vapour is 18.02 g/mol against 28.96 g/mol for dry air, so every water molecule that joins the mixture makes it lighter. Keep the statement to water in the *vapour* state as well: rain falling through the air, or water standing on the wing, genuinely does add mass, but that mass is not part of the air and does not change the density of the air.

Latent heat: the energy that water carries in and out of the air

Changing state does not change temperature, but it does move a great deal of energy. Evaporating one kilogram of water absorbs roughly 2,260 kJ from its surroundings; condensing the same kilogram releases the same amount back into them. Melting or freezing one kilogram of ice moves about 334 kJ. This energy is **latent heat** — "hidden" heat, because it produces no temperature change in the substance changing state while it is being transferred.

The aviation consequence runs in a direction that is easy to reverse and is worth stating explicitly. **Condensation RELEASES heat into the surrounding air and therefore WARMS it.** Consider a parcel of air rising in a developing cumulus cloud: it cools as it expands, reaches saturation, and its water begins to condense. From that moment the latent heat released partially offsets the cooling of expansion, so the parcel cools more slowly than it did before. Cooling more slowly means it stays warmer than the air around it for longer, so it remains buoyant for longer and rises further and faster. Latent heat release is therefore an *amplifier* of instability, and it is the energy source that turns a shallow cumulus into a thunderstorm.



Condensation releases latent heat, keeping a rising parcel warmer than its surroundings so it climbs further.

The same principle explains why the saturated adiabatic lapse rate is smaller than the dry one, a point taken up in the section on lapse rates below. It also explains why evaporation cools: the reverse process removes the same energy, which is why wet surfaces are cold and why evaporating fuel chills a tank.

The Gas Laws and the Behaviour of Air

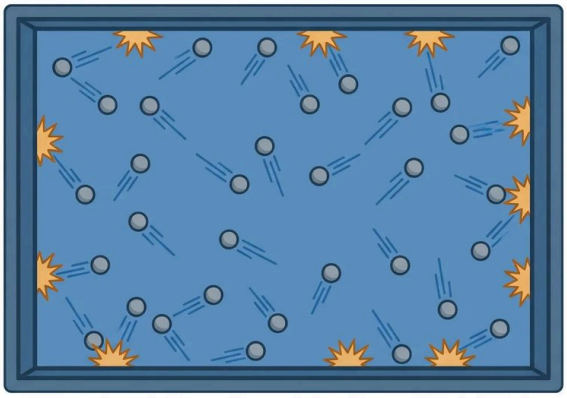
Everything the standard atmosphere says, and every performance correction derived from it, rests on a small group of physical laws describing how a gas behaves. They are worth setting out properly rather than assumed, because a large fraction of the questions in this paragraph are gas-law questions wearing atmospheric clothing.

What temperature and pressure actually are

Air is a swarm of molecules in continuous, random, high-speed motion. Two of the four properties introduced earlier are direct statements about that motion:

- **Temperature is a measure of the average kinetic energy of the molecules.** Hotter means faster. It is not a measure of potential energy, and it is not a measure of the quantity of heat present. At sea-level standard temperature the average air molecule is moving at roughly 500 m/s — comfortably faster than the aircraft — which is worth holding on to, because it explains why the speed of sound is of the same order (about 340 m/s at that temperature): sound is a pressure disturbance passed from molecule to molecule, so it cannot travel much faster than the molecules themselves.
- **Pressure is the aggregate effect of molecular bombardment.** Each molecule striking a surface delivers a tiny impulse; the pressure is the sum of those impulses per unit area per unit time. This is why pressure acts equally in all directions in still air: the molecular motion is random, so there is no preferred direction for the bombardment.

From this picture the two things that raise pressure follow at once: **put more molecules in the same space** (raise the density), or **make the existing molecules move faster** (raise the temperature). Both increase the rate of momentum delivery to the walls.



Pressure is the sum of molecular impacts on the walls, equal in all directions because the motion is random.

Heat and temperature are not the same quantity, and their relationship is direct, not inverse. Temperature is an *intensive* property — a measure of how energetic the average molecule is, independent of how much substance there is. Heat is *energy in transit*, and it flows from the higher temperature to the lower one. Adding heat to a substance raises its temperature; removing heat lowers it. A teaspoon of boiling water is at a higher temperature than a bathful of warm water but contains far less heat energy. Any statement that heat and temperature are inversely related is simply wrong.

Absolute temperature and why Celsius will not do

The gas laws are all proportionalities, and a proportionality only works from a true zero. The Celsius zero is an arbitrary point — the freezing point of water — not a point at which the quantity being measured vanishes. Doubling a temperature from 10 °C to 20 °C does not double the molecular kinetic energy, so it does not double the pressure of a confined gas.

The **absolute** or **Kelvin** scale fixes this by placing its zero at absolute zero, the temperature at which molecular motion would cease. That point is at –273.15 °C, and one kelvin is the same size as one degree Celsius, so:

$$T\,[\mathrm{K}] = t\,[^\circ\mathrm{C}] + 273.15$$

Standard sea-level temperature: $15 + 273.15 = 288.15$ K. Standard tropopause temperature: $-56.5 + 273.15 = 216.65$ K.

Every gas-law calculation uses kelvin, without exception. Substituting a Celsius temperature into any of the relationships below produces an answer that is not merely inaccurate but nonsensical — and for temperatures below freezing, a negative one. If a calculation involving temperature returns an absurd result, the first thing to check is whether the conversion to kelvin was made.

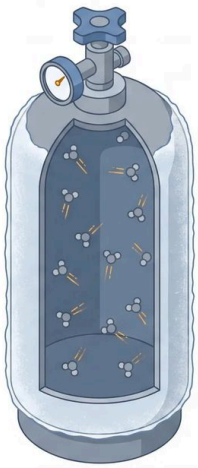
The individual gas laws

Each of the classical gas laws describes what happens when one of the three variables is held constant. All apply to a **fixed mass** of gas:

Law	Held constant	Relationship	In words	A physical example
Boyle's law	Temperature	$p_1 V_1 = p_2 V_2$	Pressure is *inversely* proportional to volume: squeeze it and the pressure rises	A syringe with the outlet blocked; a bubble rising through a liquid and expanding
Charles's law	Pressure	$\frac{V_1}{T_1} = \frac{V_2}{T_2}$	Volume is *directly* proportional to absolute temperature: heat it and it expands	A hot-air balloon; warm air rising because it has expanded and become less dense

Law	Held constant	Relationship	In words	A physical example
Pressure law (Gay-Lussac)	Volume	$\frac{p_1}{T_1} = \frac{p_2}{T_2}$	Pressure is *directly* proportional to absolute temperature: heat a sealed vessel and the pressure rises	A tyre or an oxygen bottle cold-soaking at altitude and losing indicated pressure
Combined gas law	Nothing — general	$\frac{p_1 V_1}{T_1} = \frac{p_2 V_2}{T_2}$	The three individual laws written as one	Any change of state of a sealed mass of gas

The pressure law is worth stating carefully in its own right, because its direction is easy to reverse. **For a fixed mass of gas in a fixed volume, pressure and absolute temperature rise and fall together.** Cool the gas and its pressure falls in the same proportion; the relationship is direct, not inverse. Questions in this module seldom name the individual laws; where the gas laws are examined here it is in the general form set out below.



A sealed bottle cold-soaked at altitude: the same gas, slower molecules, lower pressure, and no leak.

Worked example — a cold-soaked pressure vessel. A sealed bottle is charged on the ground at 15 °C to an indicated (gauge) pressure of 200 psi. The aircraft climbs, and the bottle cold-soaks to –40 °C. What does the gauge read at altitude, assuming no leakage and negligible change of volume?

Work in absolute quantities throughout. Absolute pressure on the ground is $200 + 14.7 = 214.7$ psia. The temperatures are $15 + 273.15 = 288.15$ K and $-40 + 273.15 = 233.15$ K.

$$p_2 = p_1 \times \frac{T_2}{T_1} = 214.7 \times \frac{233.15}{288.15} = 214.7 \times 0.8091 = 173.7 \text{ psia}$$

The gauge, which reads the amount by which the bottle pressure exceeds the surrounding pressure, would show $173.7 - 14.7 = 159$ psi if the surroundings were still at sea-level pressure — a fall of about 41 psi purely from the temperature change, with not a single molecule lost.

Direction check: the gas got colder, its molecules slowed, they strike the walls less energetically, so the pressure fell. A cold bottle reading low is normal physics, not necessarily a leak — which is exactly why bottle charge is checked against a pressure/temperature chart rather than against a single number.

The ideal gas law and the equation of state

The four relationships above are all special cases of one general law. A gas that obeys it exactly is called an **ideal** or **perfect** gas:

Molar form: $pV = nRT$

Where p = absolute pressure (Pa), V = volume (m³), n = amount of substance (mol), T = absolute temperature (K), and R is the **universal gas constant**, $8.314\ \mathrm{J/(mol \cdot K)}$.

The universal gas constant is a genuine constant of nature. It has the same value for every gas, at every altitude, at every pressure and every temperature; it does not vary with the conditions and it is not a property of air in particular. That universality is the whole point of the word "universal", and it is why the law can be applied to the atmosphere from sea level to the mesopause without adjustment.

For working with air it is more convenient to express the same law per unit *mass* rather than per mole. Dividing the universal constant by the mean molar mass of dry air gives the **specific gas constant for dry air**:

$$R_{air} = \frac{8314.32}{28.9644} = 287.05\ \mathrm{J/(kg \cdot K)}$$

and the law becomes the **equation of state** used throughout aerodynamics:

$$p = \rho R T \qquad \text{or, rearranged,} \qquad \rho = \frac{p}{RT}$$

This is the same equation that appears later in this note under density variation, and it is the mathematical statement of the dependency chain set out in the introduction: density is not an independent property of the air but a consequence of its pressure and its absolute temperature.

Do not confuse the two constants. The *universal* gas constant is 8.314 J/(mol·K) and belongs in the molar form $pV = nRT$. The *specific* gas constant for dry air is 287.05 J/(kg·K) and belongs in $p = \rho RT$. They differ by the molar mass of the gas, so the specific constant is different for every gas — it is 461.5 J/(kg·K) for water vapour, which is why the humidity calculation above needed both.

Why air can be treated as an ideal gas — and what that claim does not mean

No real gas is perfectly ideal, because the ideal model assumes the molecules occupy no volume themselves and exert no forces on one another except during collisions. Real molecules do both. However, those assumptions fail only when the molecules are crowded close together or moving slowly — that is, at very high pressure or close to the temperature at which the gas liquefies. Atmospheric air over the whole range of pressures and temperatures an aircraft meets is nowhere near either condition, so it follows $pV = nRT$ closely enough that the residual error is negligible against the other uncertainties in a performance calculation.

"Ideal gas" is not a claim that air is inviscid or incompressible, and both appear as distractors. These are three separate idealisations and they are used in different places:

- **Ideal gas** — the pressure, density and temperature of the air are related by $p = \rho RT$. This is *true* of atmospheric air to a very good approximation, and it is what justifies the standard atmosphere.
- **Inviscid** — the air has no viscosity. This is **false**: air is viscous, and its viscosity is the entire reason a boundary layer exists and the sole origin of skin-friction drag. It is assumed only in restricted parts of aerodynamic theory, outside the boundary layer.
- **Incompressible** — the density does not change with pressure. This is **false** in general, though it is a serviceable approximation at low speed. Compressibility becomes significant above roughly Mach 0.3 and dominates the high-speed behaviour of the aircraft.

A question offering "air has no viscosity" or "air is incompressible" as the reason air is treated as an ideal gas is offering a true-sounding statement about the wrong idealisation, and in both cases the statement itself is also false.

Absolute, gauge and differential pressure

A pressure reading is meaningless until you know what it is measured *from*. Three datums are in use and all three occur on aircraft:

- **Absolute pressure** is measured from a perfect vacuum. It is the pressure that appears in every gas law and in the equation of state, and it can never be negative. Units are written psia, or simply stated as an absolute value in hPa or Pa.
- **Gauge pressure** is measured from the surrounding atmospheric pressure. An ordinary pressure gauge open to the atmosphere on one side reads zero when it is disconnected, because it is comparing the system with its surroundings. Units are written psig, or simply psi on most gauges.
- **Differential pressure** is measured between two arbitrary points, neither of which need be a vacuum or the ambient air. Cabin differential pressure — cabin against ambient — and the pressure across a filter or a heat exchanger are differential readings.

The relationship between the first two is the one that is tested:

$$\text{Absolute pressure} = \text{Gauge pressure} + \text{Atmospheric pressure}$$

and therefore $\text{Gauge pressure} = \text{Absolute pressure} - \text{Atmospheric pressure}$

Worked example: a gauge on a standard day at sea level indicates 25 psi. What is the absolute pressure?

Standard sea-level atmospheric pressure is 14.7 psi, so

$25 + 14.7 = 39.7$ psia.

Direction check. The gauge is blind to the atmosphere pressing on its outside, so it always *under*-states the true pressure inside the system by exactly one atmosphere. The absolute figure must therefore always be the *larger* of the two at sea level. Subtracting instead of adding gives 10.3 psi — smaller than the gauge reading, which is impossible for a system above ambient, and that inconsistency identifies the wrong answer before any arithmetic is checked.

The same reasoning in reverse. If a question gives an absolute pressure and asks for the gauge reading, subtract 14.7 psi — but only if the surroundings are at sea-level pressure. At altitude the atmospheric term is smaller, so the same absolute pressure reads *higher* on a gauge. This is why a tyre or an accumulator gauge reads higher at a high-elevation airfield than it does at sea level with exactly the same quantity of gas inside, and why servicing pressures for some systems are quoted against field elevation.

Putting the gas law to work on the atmosphere

The equation of state settles a question that otherwise looks paradoxical. Density depends on pressure and temperature, and in the troposphere *both* of them fall with altitude. Falling pressure reduces density; falling temperature increases it. The two effects oppose each other, so which one wins?

Worked example — sea level to the tropopause. Take the standard values at each end and form the ratios:

- Pressure falls from 1013.25 hPa to 226.3 hPa, a ratio of $226.3/1013.25 = 0.223$ — the pressure retains 22.3 % of its sea-level value.
- Absolute temperature falls from 288.15 K to 216.65 K, a ratio of $216.65/288.15 = 0.752$ — the temperature retains 75.2 % of its sea-level value.

Since $\rho = p/RT$, the density ratio is the pressure ratio divided by the temperature ratio:

$$\frac{0.223}{0.752} = 0.297$$

so the air at the tropopause has just under **30 % of its sea-level density**. Compare that with the pressure, which has fallen to 22 %: the density has not fallen quite as far as the pressure, and the difference is exactly the amount by which the cooling has partially compensated.

The pressure effect wins decisively, and it wins because pressure falls far faster than absolute temperature does. Over the whole troposphere the pressure loses more than three quarters of its value while the absolute temperature loses only a quarter of its own. This is the quantitative answer to a question that is

asked in several forms: air density decreases with altitude because *both* pressure and temperature decrease, and the reduction in pressure — fewer molecules pressed into each cubic metre — is the dominant term. It also explains a related result: pressure and temperature do not fall at the same rate, and the pressure decline is by far the steeper of the two.

Buoyancy: why anything in the atmosphere rises or sinks

Several arguments in this note have turned on a parcel of air being "buoyant" or "heavier than its surroundings", and it is worth setting the principle out explicitly because it is the mechanism behind convection, behind stability, behind the sea breeze and behind the whole idea of an air mass.

Archimedes' principle states that a body immersed in a fluid experiences an upward force equal to the weight of the fluid it displaces. Applied to a parcel of air within the atmosphere, the displaced fluid is the surrounding air, so:

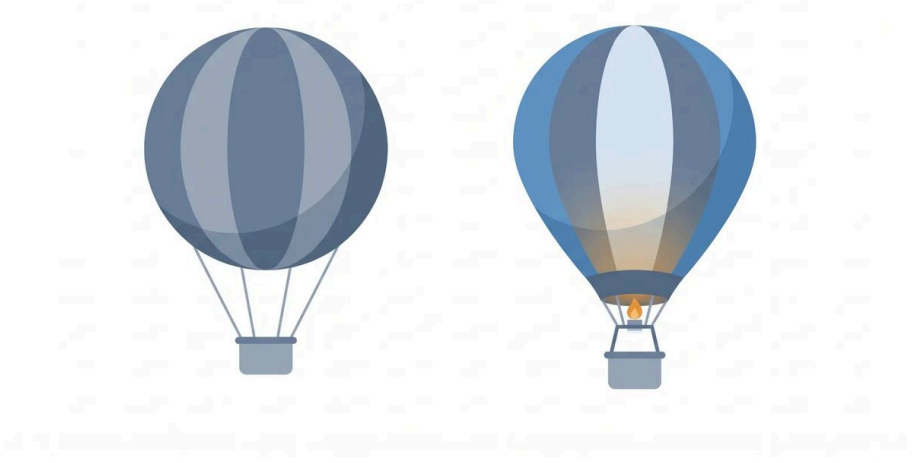
Net upward force $= (\rho_{surrounding} - \rho_{parcel}) \times V \times g$

Where V is the volume of the parcel and g is gravitational acceleration.

The sign of that expression, and therefore the direction of the motion, depends entirely on the **difference in density** between the parcel and its surroundings. A parcel less dense than its surroundings rises; one more dense sinks; one of equal density stays where it is. Since a parcel at the same pressure as its surroundings has a density set by its temperature, and warmer means less dense, the familiar statement "warm air rises" is a shorthand for "air that is warmer than its surroundings is less dense than its surroundings and is therefore buoyant".

The comparison is with the surroundings, always. Warm air does not rise because it is warm; it rises because it is warmer *than the air around it*. Air at +40 °C sitting in air at +50 °C sinks. This is exactly the point of the stability discussion below: a rising parcel cools as it goes, and whether it continues to rise depends on whether it is still warmer than its new surroundings, not on whether it is warm in any absolute sense.

The same principle applied to a body that is not made of air explains lighter-than-air flight. A balloon rises when the total weight of the envelope, its contents and its payload is less than the weight of the atmosphere it displaces. There are two ways to achieve that, and both are density arguments: fill it with a gas of lower molar mass than air — helium at 4 g/mol or hydrogen at 2 g/mol against air's 28.96 — or fill it with ordinary air heated so that some of it has been driven out and what remains is less dense. It also explains why a balloon has a ceiling: as it rises, the surrounding air becomes less dense, so the weight of air displaced falls, and equilibrium is reached where that displaced weight equals the total weight of the balloon.

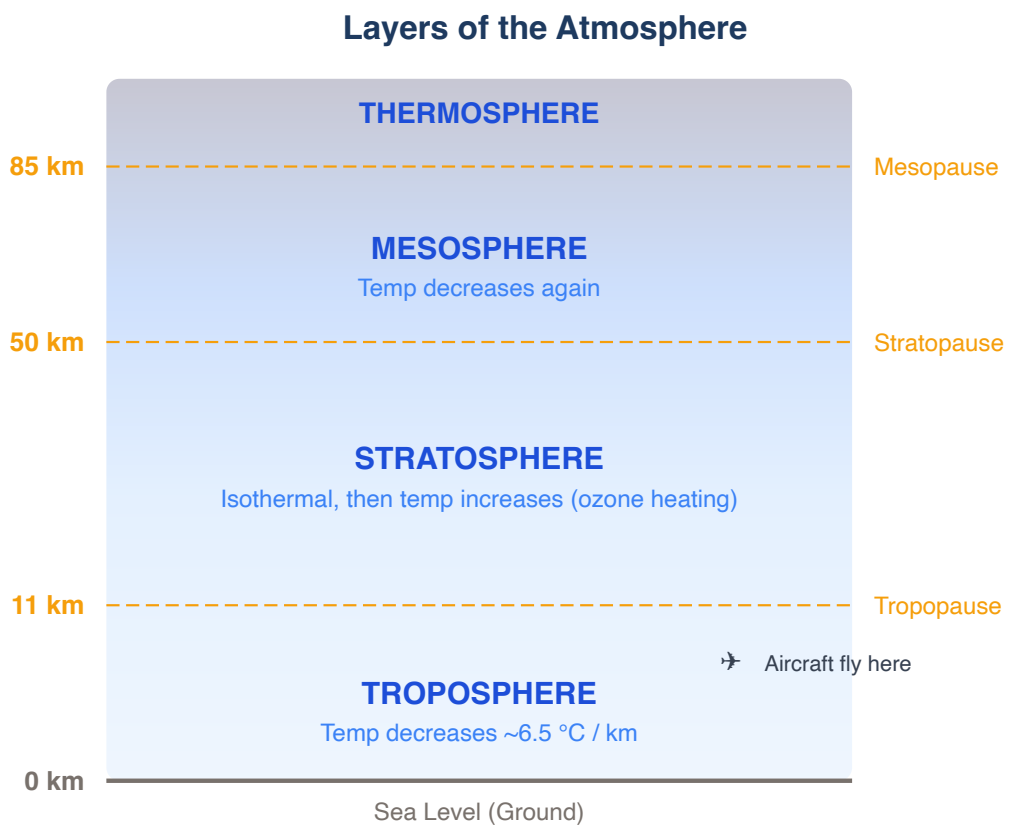


A gas balloon and a hot-air balloon both rise because they weigh less than the air they displace.

Finally, the same effect acts on the aircraft itself, though the numbers make it negligible. An aircraft displaces its own volume of air, so it is very slightly buoyant — a large transport displaces perhaps a tonne or two of air at sea level, against a weight of hundreds of tonnes. It is real, it is why an aircraft weighed indoors and one weighed at altitude differ by a trace, and it is far too small to matter beside the aerodynamic forces this module is otherwise concerned with.

Regions (Layers) of the Atmosphere

The atmosphere is divided into distinct layers based on how temperature changes with altitude. The boundaries between layers are called **pauses** (tropopause, stratopause, mesopause). Understanding these layers is important because each has different characteristics that affect flight.



As in almost every diagram of this kind, the vertical scale above is progressively compressed with height, so that the thin lower layers where aircraft actually operate remain legible alongside layers hundreds of kilometres deep. Read it for the order and the temperature trends rather than for proportions.

How the layers are defined, and why there are exactly four of them

The layers are not arbitrary bands chosen for convenience. Each one is a region in which the temperature changes with height in a consistent *direction*, and each pause is the level at which that direction **reverses**. The result is an alternating sequence, and once you see the alternation the whole structure becomes memorable rather than something to be learned by rote:

Layer	Approximate extent	Temperature with increasing height	Ends at	What drives the temperature behaviour
Troposphere	Surface to about 11 km (36,000 ft) in the ISA	**Falls**	Tropopause	Heated from below by the sunlit surface
Stratosphere	About 11 km to about 50 km	**Constant, then rises**	Stratopause	Heated from within and above by ozone absorbing ultraviolet
Mesosphere	About 50 km to about 80–85 km	**Falls**	Mesopause	Too little ozone to absorb; radiates its heat away to space

Layer	Approximate extent	Temperature with increasing height	Ends at	What drives the temperature behaviour
Thermosphere	About 85 km upward	**Rises**	Thermopause	Extreme ultraviolet and X-radiation absorbed by very sparse gas

Read down the third column: falls, constant then rises, falls, rises. Each pause is a turning point, and each turning point is caused by a change in what is heating the air at that level. The troposphere is warmed from beneath, so it is coldest at the top; the stratosphere is warmed from above and within, so it is warmest at the top; and so on alternately. **A layer's name tells you nothing about its temperature; the heat source does.**

Where the mass is, and how to read it straight off the pressure

There is an exact relationship worth knowing, because it converts the pressure table later in this note into a mass table at no extra effort. Static pressure at any level is the weight of the air column above that level, divided by area. It follows directly that **the pressure ratio at any altitude is the fraction of the atmosphere's total mass lying above it**, and one minus that ratio is the fraction lying below:

Level	ISA pressure	Fraction of the atmosphere's mass BELOW this level
18,000 ft (5.5 km)	about half of sea level	about 50 %
Tropopause, 11 km	226 hPa	about 78 %
20 km	54.7 hPa	about 95 %
32 km	8.7 hPa	about 99 %
Stratopause, 50 km	0.76 hPa	about 99.9 %

So the whole of the stratosphere, mesosphere and thermosphere together — hundreds of kilometres of atmosphere — account for barely a fifth of its mass, and everything above 50 km amounts to one part in a thousand. This is why the layers above the stratosphere are of no aerodynamic consequence whatever: there is not enough air there to generate a usable force. It is also the physical reason the halving rule discussed later works at all, and why the first 18,000 ft of climb sheds as much pressure as the whole of the rest of the atmosphere put together.

Three different ways of slicing the same atmosphere

The four-layer scheme above is only one classification, and questions occasionally mix schemes to see whether the distinction is understood. The atmosphere can equally be divided by composition or by electrical properties, and those divisions do **not** line up with the temperature-based ones:

Basis of division	Regions	Boundary	Why an engineer cares
Temperature gradient	Troposphere, stratosphere, mesosphere, thermosphere	The pauses, where the gradient reverses	This is the standard aviation scheme and the one the standard atmosphere is built on
Composition	Homosphere (well mixed) below; heterosphere (separated by molar mass) above	Turbopause, about 80 km	Explains why the gas proportions are constant throughout the whole flight envelope
Ionisation	Neutral atmosphere below; ionosphere above	Begins about 60 km, extends to roughly 1,000 km	Governs long-range HF radio propagation and affects satellite navigation signals

The important structural point is that the **ionosphere is an overlay, not one of the four layers**. It begins in the upper mesosphere and continues right through the thermosphere, so it overlaps two of the temperature-defined layers and belongs to neither. A question asking which layer has "fairly constant temperature", or which is the outermost, is asking about the temperature scheme, and the ionosphere is not a valid answer to it.

The atmosphere as an operating envelope

Read from the point of view of an aircraft rather than a meteorologist, the same vertical structure divides into bands defined by what becomes possible or impossible in each. The boundaries are set by the pressure profile, and each band is stated below as a fraction of the sea-level pressure:

Band	Pressure and density	What governs operations there
Surface to about 10,000 ft	Pressure 100 % down to about 69 % of sea level	Unpressurised operation is practicable without supplemental oxygen. Weather, terrain and traffic dominate; the friction layer and its wind gradient occupy the lowest part
10,000 ft to about 18,000 ft	Down to about half of sea level	Supplemental oxygen becomes progressively necessary; unpressurised aircraft can operate but with an oxygen system and time limits
18,000 ft to the tropopause	Half down to about 22 %	Pressurisation is effectively mandatory. Turboprops and business aircraft cruise here; the air is thin enough for good true airspeeds but weather is still present
Tropopause to about 45,000 ft	22 % down to about 15 %	The transport cruise band. Above most weather, temperature constant, speed of sound constant, and the aircraft is Mach-limited rather than speed-limited
45,000 ft to about 60,000 ft	15 % down to about 7 %	Specialised high-altitude aircraft only. Wings must be very large or speeds very high; the margin between low-speed and high-speed limits is narrow

Band	Pressure and density	What governs operations there
Above about 63,000 ft	About 6 % of sea level	The Armstrong limit: ambient pressure equals the vapour pressure of water at body temperature, so a pressure suit or pressurised cabin becomes a survival requirement rather than a comfort one

Two observations tie the table to the physics. First, **every boundary in it is a pressure boundary, not a temperature one** — the limits on human occupancy, on engine mass flow and on wing loading all follow the pressure profile, which is why the aviation-relevant structure of the atmosphere looks nothing like the temperature-based layer scheme above it. Second, the whole of that table lies within the lowest 20 km, which is to say within the troposphere and the bottom of the stratosphere. **Everything aviation does happens beneath the 20 km level, in a shell that holds about 95 % of the atmosphere's mass and almost none of its height.**

Troposphere (0 – ~11 km / 0 – ~36,000 ft)

The troposphere is the lowest and most important layer for aviation. It extends from the Earth's surface to the **tropopause**, which is at approximately 11 km (36,000 ft) altitude in the ISA. The word "troposphere" comes from the Greek *tropos*, meaning "turning" or "mixing," reflecting the turbulent, weather-producing nature of this layer.

Key characteristics of the troposphere:

- **Contains approximately 75 % of the total mass of the atmosphere** — the air is densest near the surface and becomes progressively thinner with altitude.
- **Temperature decreases with altitude** at an average rate of approximately **6.5 °C per 1,000 m** (about 2 °C per 1,000 ft). This rate of temperature change is called the **Environmental Lapse Rate (ELR)**. The reason temperature decreases is that the ground absorbs solar radiation and heats the air from below — the further you go from this heat source, the cooler it gets.
- **All significant weather occurs in the troposphere** — clouds, rain, snow, thunderstorms, turbulence, icing, and wind shear are all tropospheric phenomena. This is because the troposphere contains virtually all of the atmosphere's water vapour and has the vertical temperature profile that drives convective air currents.
- **Virtually all commercial and general aviation takes place here** — or in the very lowest part of the stratosphere. Turboprop aircraft typically cruise at 20,000–30,000 ft, while jet airliners cruise at 30,000–43,000 ft.
- **Wind speed generally increases with altitude** — surface friction slows winds near the ground. At higher levels, winds can be very strong, especially in the **jet stream** (narrow bands of very high-speed wind near the tropopause, often 100–200 knots).

Two features of the troposphere deserve more than a bullet, because they generate a disproportionate share of the operational problems an engineer sees recorded in the technical log.

The friction layer and the wind gradient

The lowest part of the troposphere — typically the first 2,000 ft or so above the surface, though it varies widely with terrain, surface heating and wind strength — is dominated by contact with the ground. Surface roughness drags on the moving air, and that drag is communicated upward by turbulent mixing, so the wind speed increases steadily with height through this layer and the wind direction backs towards the surface in the northern hemisphere as friction weakens the balance between pressure gradient and Coriolis deflection.

The result is a **wind gradient**: a change of wind speed with height, present on virtually every approach and landing, and strongest when the wind is strong and the air is stable. Its effect on an aircraft descending through it is a direct consequence of inertia. The aircraft's momentum is carried in the earth frame; the air's velocity changes as the aircraft descends into slower-moving air. For the very short time it takes the aircraft to re-establish its trimmed airspeed relative to the new air, the **indicated airspeed falls**, and with it the lift.

Direction worked through, because it is easy to reverse. An aircraft is on final approach into a headwind of 30 kt at 500 ft, decreasing to 10 kt at the surface. As it descends it enters air that is moving towards it 20 kt more slowly than the air it has just left. Its speed over the ground cannot change instantly, so for a moment its speed *through the air* is 20 kt lower than it was.

Lower indicated airspeed at a constant attitude means less lift, so the aircraft sinks below the glidepath — the classic **undershoot**. The correct response is an immediate increase in power to restore the airspeed, and the hazard is greatest close to the ground where there is neither height nor time to recover.

The reverse case, a *decreasing tailwind* on the approach, produces a momentary *increase* in indicated airspeed and a tendency to float or overshoot. In both cases it is the **change** in wind that matters. A steady 30 kt headwind all the way to the ground produces no such effect at all.

The tropopause is coldest over the equator, not over the poles

The equatorial tropopause is the *higher* of the two, and it is also the *colder* of the two, which sounds contradictory until the mechanism is laid out. The surface at the equator is hot and the surface at the poles is cold, so the equatorial temperature profile starts much warmer. But the equatorial troposphere is roughly twice as deep — about 16 to 18 km against about 8 km — and the lapse rate operates over that whole depth. Applying a lapse rate of the order of 6.5 °C/km to twice the height removes far more temperature than the warm start had to give away.

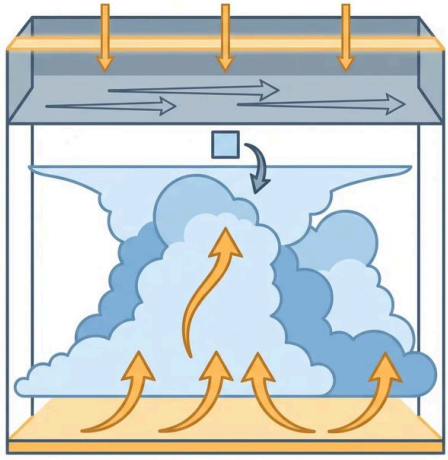
Worked through with round numbers: a surface temperature of about +27 °C falling over 17 km reaches roughly –83 °C, whereas a polar surface temperature of about –20 °C falling over 8 km reaches only about –72 °C. Real values vary with season and with the day, but the ordering is consistent and is well established: **the coldest air in the lower atmosphere is found at the tropical tropopause, not over the poles**. A crew flying a long tropical sector at high level, and an engineer investigating a fuel-temperature or a cold-soak-related defect, both meet the consequence of that fact.

Why "most of the mass" and "all of the weather" go together

The two headline properties of the troposphere are not independent facts to be memorised separately — each causes the other. The layer holds about three quarters of the atmosphere's mass because it is at the bottom

and everything above presses down on it. It holds virtually all of the water vapour because water vapour is evaporated from the surface and because the temperature falls steeply enough with height that very little survives to the tropopause. And it has weather because it is heated from below, which is precisely the arrangement that makes air unstable: warm, less dense air underneath colder, denser air will overturn.

Turn that last point around and the stratosphere follows for free. There the heating is from above, so temperature stops falling with height — constant through the lower stratosphere, rising above about 20 km. A parcel displaced upward still cools as it expands, but the air it arrives in is no colder than the air it left, so it ends up colder and denser than its surroundings and sinks back. The arrangement is stable — vertical motion is suppressed and there is no convective weather. **The single question "where is this layer heated from?" predicts both the temperature gradient and the weather behaviour of every layer in the atmosphere.**



Air heated from below overturns and makes weather; air heated from above is stable and stays clear.

Tropopause (~11 km / ~36,000 ft)

The **tropopause** is the boundary between the troposphere and the stratosphere. It is not a sharp line but rather a transition zone where the temperature lapse rate decreases to zero — temperature stops falling and either remains constant or begins to increase.

The tropopause is **not at the same altitude everywhere**:

- **At the poles:** approximately **8 km** (26,000 ft) — lower because polar air is colder and denser, so it compresses the troposphere
- **At the equator:** approximately **16–18 km** (52,000–59,000 ft) — higher because tropical air is warmer and expands the troposphere upward
- **At mid-latitudes (ISA standard):** approximately **11 km** (36,089 ft)
- The tropopause also varies with **season** (higher in summer, lower in winter) and with **weather systems** (higher in high-pressure systems, lower in low-pressure systems)

The tropopause is significant for aviation because it often marks the upper boundary of weather — flying above the tropopause generally means smooth, clear air. The jet stream is found near the tropopause, particularly where there are sharp breaks or discontinuities in its height.

How the tropopause is formally located

Because the real temperature profile is ragged rather than a clean pair of straight lines, meteorologists use a threshold definition rather than a corner: the tropopause is taken to be the **lowest level at which the lapse rate falls to 2 °C per kilometre or less, provided the average lapse rate stays at or below that value through the next 2 km above it**. The second clause exists to prevent a shallow inversion inside the troposphere from being mistaken for the top of it. In the standard atmosphere the problem does not arise, because the model simply declares the tropopause to be at 11,000 m by definition.

Keep the model and the real atmosphere separate. The tropopause is *defined* as 11,000 m (36,089 ft) in the standard atmosphere and is a fixed number there. In the real atmosphere it is a physically located surface that moves with latitude, season and the weather, and is frequently nowhere near 36,000 ft. Both statements are correct; questions specifying "in the ISA" want the first, and questions about polar or equatorial flying want the second.

What changes for the aircraft when it crosses the tropopause

The tropopause is not merely a meteorological curiosity; several quantities the aircraft depends on change their behaviour there, and the changes are examinable:

- **The temperature stops falling.** In the standard atmosphere, the temperature holds at –56.5 °C (216.65 K) from the tropopause up to about 20 km, which is roughly 65,000 ft. Above that it starts to increase again, at about 0.3 °C per 1,000 ft — a rate roughly one sixth of the tropospheric lapse rate and in the opposite sense.
- **The speed of sound stops falling.** This is the single most useful consequence for an engineer and it follows from the previous point. The local speed of sound depends on *absolute temperature alone* — not on pressure, not on density, and not on altitude except through temperature. It falls with height in the troposphere only because the temperature does. Above the tropopause the temperature is constant, so the speed of sound is constant too, at about 295 m/s (573 kt) against 340 m/s (661 kt) at sea level.
- **The relationship between Mach number and true airspeed therefore stabilises.** Below the tropopause, climbing at a constant Mach number means the true airspeed falls continuously as the air cools. Above it, a constant Mach number is a constant true airspeed, because the reference speed is no longer changing.
- **The pressure does NOT do anything special.** It continues to decay exponentially straight through the tropopause and on up. The exponential character of the pressure fall is a property of the whole atmosphere and has nothing to do with the layer boundaries, so no discontinuity of any kind appears in the pressure curve at 36,089 ft.
- **The weather generally stops.** The stable stratospheric air above the tropopause caps convection, which is why the anvil of a mature thunderstorm spreads out horizontally instead of continuing to grow upward — it has run into a lid it cannot penetrate.

Local speed of sound: $a = \sqrt{\gamma R T}$

Where $\gamma = 1.4$ for air, $R = 287.05\ \mathrm{J/(kg \cdot K)}$ and T is the absolute temperature in kelvin. At 288.15 K this gives $a = 340.3$ m/s; at 216.65 K it gives $a = 295.1$ m/s.

A frequently reversed claim. "The speed of sound decreases with altitude" is true in the troposphere and **false above the tropopause**, where the temperature is constant and the speed of sound is constant with it. The safe form of the statement, correct everywhere, is that the speed of sound depends only on the absolute temperature of the air. Pressure and density do not appear in the expression above, and changing either one at constant temperature does not change the speed of sound at all.

The tropopause break and clear air turbulence

Because the tropopause height differs sharply between adjacent air masses — particularly between the deep tropical troposphere and the shallow polar one — the surface is not continuous. Where a warm, high tropopause meets a cold, low one, there is a step known as a **tropopause break**. The strong horizontal temperature gradient across such a break drives a correspondingly strong wind, which is where the jet streams are found, and the shear at the edges of that wind is the principal source of clear air turbulence. The tropopause break is therefore not an academic feature: it is the physical location where the two most significant high-level hazards to a cruising aircraft both live, and it is dealt with further in the section on atmospheric phenomena below.

Stratosphere (~11 – 50 km)

The stratosphere extends from the tropopause to the **stratopause** at about 50 km. The name comes from the Latin *stratum*, meaning "layer," because the air here forms stable, horizontal layers with very little vertical mixing.

- **Temperature stops falling and then rises with altitude** in the stratosphere. In the lower stratosphere (from 11 km to about 20 km), temperature remains approximately constant at –56.5 °C (in the ISA). Above 20 km, it begins to rise, reaching about 0 °C at the stratopause (50 km).
- **The reason for this temperature increase is the ozone layer** — a region of high ozone (O_3) concentration centred at about 25 km altitude. Ozone absorbs ultraviolet (UV) radiation from the sun, converting it to heat. This heating warms the upper stratosphere.
- **Very little weather** — because temperature stops falling and then rises with altitude, the stratosphere is extremely stable. Warm air over cold air suppresses vertical motion, so there is virtually no convection, no clouds (except rare nacreous/polar stratospheric clouds), and no turbulence.
- **Some high-altitude aircraft operate here** — Concorde cruised at about 60,000 ft (18 km), and military reconnaissance aircraft like the U-2 and SR-71 routinely flew in the stratosphere.

Expressed in the units used by the altimeter, the stratosphere runs from about 36,000 ft to about 164,000 ft — a band roughly four times as deep as the whole of the troposphere beneath it, yet containing only about a fifth of the mass.

The temperature profile in detail

The standard atmosphere models the stratosphere as three straight-line segments rather than one, and knowing where the corners are removes most of the ambiguity from questions about temperature above the tropopause:

Band	Equivalent in feet	Rate of change with height	Temperature at the top of the band
11 km to 20 km	36,089 ft to about 65,600 ft	**Zero — isothermal**	–56.5 °C (216.65 K)
20 km to 32 km	about 65,600 ft to about 105,000 ft	+1.0 °C per km (about +0.3 °C per 1,000 ft)	–44.5 °C
32 km to 47 km	about 105,000 ft to about 154,000 ft	+2.8 °C per km (about +0.85 °C per 1,000 ft)	–2.5 °C
47 km to 51 km	about 154,000 ft to about 167,000 ft	Zero — isothermal again	–2.5 °C, the stratopause value

Two figures from that table are worth carrying into the examination. The first is the **65,600 ft break**, which is where the isothermal band ends. Questions phrased "above 65,800 ft, temperature..." are asking about the segment immediately above it, and the answer is an increase of about **0.3 °C per 1,000 ft** — a number that follows directly from converting 1.0 °C per kilometre into feet, since 1,000 ft is 0.3048 km. The second is that the warming does not carry the temperature far above freezing even at the stratopause; it climbs back to within a few degrees of 0 °C and no further.

Reading the three bands as one sentence: above the tropopause the temperature holds steady for about 30,000 ft, then rises — gently at first and then about three times faster — until it levels off again just below freezing at the stratopause. Every stratospheric temperature question is answered by identifying which of those three regimes the stated altitude falls in.

The ozone layer and where the heat actually goes

Ozone is produced and destroyed continuously in the stratosphere by ultraviolet radiation. Short-wavelength ultraviolet splits an ordinary oxygen molecule into two free oxygen atoms; each of those atoms attaches to another oxygen molecule to form ozone. Longer-wavelength ultraviolet then breaks the ozone apart again, and the cycle repeats. **Every step of that cycle converts radiant energy into molecular kinetic energy — that is, into heat.** The ozone is not consumed on balance; it acts as a catalyst that turns ultraviolet radiation into warmth.

Ozone is present in a broad band, commonly quoted as extending from roughly 15 km to 35 km with the greatest concentration in the middle of it. But the temperature does not peak where the ozone concentration peaks — it goes on rising to the stratopause at 50 km, well above the ozone maximum. The reason is a competition between two quantities that vary in opposite directions with height:

- **Ultraviolet intensity is greatest at the top** and is progressively absorbed on the way down, so the energy available diminishes with decreasing altitude.
- **Air density is greatest at the bottom**, so at low levels the absorbed energy is shared among far more molecules.

Temperature responds to energy absorbed *per unit mass*, so the strongest heating per kilogram occurs high up, where the radiation is still strong and there is very little air to warm. This is why the warmest point of the

stratosphere is its ceiling, and it is a clean illustration of a principle that recurs throughout this note: **a temperature is set by energy per unit mass, not by total energy**. The same principle explains the apparent paradox of the thermosphere, described below.

Why the stratosphere is aerodynamically attractive and physically hostile

The absence of vertical motion makes the stratosphere ideal cruising air: smooth, cloud-free and predictable. Against that, everything else about it is difficult:

- **There is almost no air.** At 60,000 ft the density is about 9 % of the sea-level value, so a wing must fly very fast, or be very large, to generate a given lift. The margin between the low-speed stall and the high-speed compressibility limit narrows as altitude increases, and it is the shrinking of that margin, rather than any single limit, that ultimately caps the cruising altitude of a subsonic transport.
- **Ozone itself is a contaminant.** It is chemically aggressive, attacks rubber and some elastomers, and is toxic to occupants in the concentrations found at high stratospheric levels. Aircraft certified for sustained flight at those levels have to address it — typically by catalytic conversion of ozone in the bleed air supply, since ozone decomposes readily back to ordinary oxygen when heated or passed over a suitable catalyst.
- **The air is extremely dry.** The cold trap at the tropopause removes almost all the water vapour, so stratospheric humidity is very low. Nacreous, or polar stratospheric, clouds form only in the exceptionally cold polar winter stratosphere and are the rare exception that proves the point.

Mesosphere (~50 – 85 km)

The mesosphere extends from the stratopause to the **mesopause** at about 80–85 km. Temperature **decreases** again with altitude in this layer, reaching the coldest temperatures in the entire atmosphere — as low as **–90 °C** at the mesopause. Meteors burn up in the mesosphere due to friction with the (still present but very thin) air molecules. This layer is above the operating altitude of all aircraft but below orbital altitude, making it difficult to study directly — it is sometimes called the "ignorosphere."

Why the mesosphere cools with height

The mesosphere sits between two layers that both warm with altitude, and its reversal is caused by the disappearance of the heat source rather than by the appearance of a cooling mechanism. Above the stratopause the ozone concentration has fallen away, so there is nothing left to absorb ultraviolet radiation efficiently. At the same time carbon dioxide in the layer radiates infrared energy directly to space, and there is no dense atmosphere above it to return any of that radiation. The layer is therefore losing heat at the top and receiving very little, which is exactly the arrangement that produces a falling temperature with height.

The consequence is that the **coldest temperatures found anywhere in the atmosphere occur at the mesopause** — colder than the tropical tropopause, colder than the polar surface in winter. Ice can form there in spite of the almost total absence of water, producing the highest clouds in the atmosphere, the noctilucent clouds, visible only when the sun is below the observer's horizon but still illuminating cloud 80 km up.

Meteors, and why the mesosphere is hard to study

The mesosphere is dense enough to decelerate an object arriving at tens of kilometres per second, and it is the layer in which almost all meteors are destroyed. The heating is aerodynamic rather than frictional in the everyday sense: the air ahead of a hypersonic object is violently compressed, and it is that compression, together with the shock system it creates, that raises the temperature enough to vaporise the object.

The same combination of properties makes the layer awkward to measure. It is far too thin to support winged flight — there is around one thousandth of the sea-level air density at its base and less than one hundred-thousandth at its top — yet still dense enough that a satellite in a circular orbit there would not survive a single revolution. It can be reached only by sounding rockets passing through it, which is the origin of its unofficial nickname among atmospheric scientists.

Placing the mesosphere in an examination answer. Three facts fix it: it lies immediately above the stratopause; the temperature *decreases* with height through it, so it behaves like the troposphere in that respect and unlike the two layers on either side of it; and its top, the mesopause, is the coldest point in the atmosphere. The extent is commonly given as roughly 50 km to 80–85 km, sources differing over exactly where to place the upper boundary because the mesopause height itself varies with season and latitude.

Thermosphere (~85 – 600 km)

The thermosphere is the outermost significant layer. Temperature **increases rapidly** due to absorption of extremely energetic solar radiation by the sparse gas molecules. Temperatures can reach **1,000 °C or more**, but because the air is so thin, these temperatures would not feel "hot" in the conventional sense — there are simply too few molecules to transfer significant heat. The International Space Station orbits within the thermosphere at about 400 km altitude. The **aurora borealis** (northern lights) and **aurora australis** (southern lights) occur in this layer.

The apparent paradox of a layer at 1,000 °C that would not feel hot is worth resolving properly, because it is the clearest available illustration of the difference between temperature and heat content set out earlier in this note. Temperature states how fast the individual molecules are moving; heat transfer to an object depends on how many of them strike it per second. In the thermosphere the molecules are extraordinarily energetic and extraordinarily scarce, so the first quantity is enormous and the second is negligible. A thermometer placed there would in fact read *cold*, because it would radiate energy away to space faster than the rare molecular impacts could deliver it. The quantity being quoted is the **kinetic temperature** — a statement about molecular speed, and a perfectly valid one — not a prediction of what the layer would do to a solid body placed in it.

Thin as it is, the thermosphere is not aerodynamically inert. It exerts enough drag to decay the orbit of a low-orbiting satellite over a period of years, which is why such satellites carry propellant for periodic re-boost and why they eventually re-enter. That drag is also variable: solar activity heats the thermosphere and makes it expand, raising the density at any given altitude, so satellite decay rates rise measurably at solar maximum. The thermosphere is the only layer whose thickness changes substantially over an eleven-year cycle.

The Exosphere

Above the thermosphere lies the **exosphere**, the outermost layer of the atmosphere and the region in which the atmosphere ceases to behave as a gas at all. Its base — the thermopause, also called the exobase — is conventionally placed at about **500 km**, though it moves between roughly 500 km and 1,000 km with the state of the sun, for the expansion reason just described.

What distinguishes the exosphere is not its composition or its temperature but the **mean free path**: the average distance a molecule travels before colliding with another one. Lower down, that distance is microscopic and the enormous number of collisions is what makes a body of air behave as a continuous fluid with a pressure, a density and a temperature. In the exosphere the mean free path becomes comparable with, and then far greater than, the scale of the atmosphere itself. A molecule there follows a long ballistic arc under gravity and is far more likely to complete it uninterrupted than to hit anything.

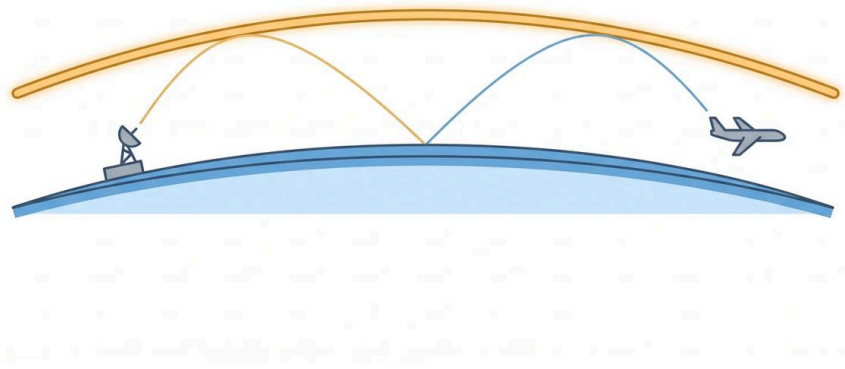
The consequence gives the layer its name and its significance: a molecule travelling upward fast enough, and not colliding with anything, simply **escapes into space**. Because escape depends on speed and lighter molecules travel faster at a given temperature, hydrogen and helium escape preferentially while nitrogen and oxygen are effectively retained. The exosphere is where the Earth loses its atmosphere, one molecule at a time, and its outer limit is not a surface but a gradual fading into the interplanetary medium at several tens of thousands of kilometres.

Two "outermost" statements, both correct, and how to tell which one a question wants. The **thermosphere** is the outermost layer that still has enough gas to matter physically — it is where the aurorae occur, where the International Space Station orbits, and the last layer in which anything behaves aerodynamically. The **exosphere**, beginning at about 500 km, is the outermost layer of the atmosphere in the complete sequence, and it is defined by molecular escape. A question listing the layers in order, or asking where molecules escape to space, wants the exosphere; a question about the highest layer of any operational consequence wants the thermosphere.

The Ionosphere

The **ionosphere** is not a fifth layer in the temperature sequence but a property of a region that spans several of them. Solar ultraviolet and X-radiation strip electrons from the sparse gas molecules, producing a shell of free electrons and positive ions extending from roughly **60 km to about 1,000 km** — that is, from the upper mesosphere, through the whole thermosphere, and into the base of the exosphere.

Its importance to aviation is entirely a radio one, and it is substantial. Free electrons refract radio waves, and for waves in the high-frequency band the refraction is strong enough to bend a signal launched obliquely upward back towards the ground. Repeated bounces between the ionosphere and the earth's surface allow HF communication over thousands of miles, which is why HF remains the primary long-range voice link on oceanic and remote routes where no ground VHF station is in range.



HF signals bend back from the ionosphere and bounce off the surface, reaching aircraft beyond VHF range.

The same medium has two further consequences an engineer meets in practice:

- **The ionosphere is variable, and HF performance varies with it.** Ionisation depends on solar radiation, so it is stronger by day than by night and stronger at solar maximum. The usable frequency for a given path therefore changes with the time of day and the season, which is why HF frequencies are allocated in day and night pairs and why an HF set that works perfectly on a functional check on the ground may be unusable on a particular route at a particular hour. That is a propagation limitation, not a defect.
- **It delays satellite navigation signals.** The free electrons slow the signal's modulation as it passes through, introducing a ranging error that is one of the largest error sources in satellite navigation. Dual-frequency receivers measure and remove it directly by comparing the delay at two frequencies; single-frequency receivers rely on a broadcast model, and augmentation systems transmit measured corrections.

International Standard Atmosphere (ISA)

The real atmosphere is constantly changing — it varies with geographic location, time of day, season, and weather patterns. On a hot summer day in Dubai, sea-level conditions are radically different from those on a cold winter day in Tromsø, Norway. This variability creates a problem: how can we compare aircraft performance data, calibrate instruments, or design aircraft when the atmospheric conditions keep changing?

The solution is the **International Standard Atmosphere (ISA)** — a fixed, internationally agreed reference model of the atmosphere. The ISA was originally defined by ICAO (International Civil Aviation Organization) and is also published in ISO 2533. It does *not* represent the actual atmosphere at any particular place or time — it is a **theoretical model** used as a common reference baseline.

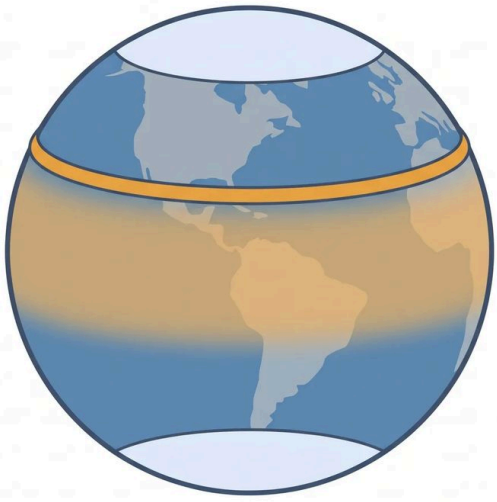
What the ISA actually fixes, and what it is referenced to

Three points about the model are examined repeatedly and are easy to state loosely.

First, the ISA defines a whole profile, not a single set of surface numbers. It specifies temperature, pressure and density at mean sea level *and* the rule by which each of them changes with altitude, all the way up. An answer that describes the ISA as "the standard sea-level temperature of 15 °C" has named one value

out of the several dozen the model contains, and has missed the variation with height that is the entire reason the model is useful for anything above the ground.

Second, the ISA is referenced to 45 degrees latitude. Not to the equator, where the surface is warmer and the tropopause far higher, and not to the poles. Forty-five degrees is a mid-latitude compromise, and the sea-level values — 15 °C, 1013.25 hPa, 1.225 kg/m³ — are representative annual mean conditions there. The choice of latitude also fixes the value of gravitational acceleration built into the model, 9.80665 m/s², which is very close to the true value at 45 degrees.



The ISA is referenced to 45 degrees latitude, a mid-latitude compromise between warm equator and cold poles.

Third, the ISA is used worldwide precisely because it is not local. An engine performance figure, a take-off distance or an altimeter calibration means nothing unless the atmosphere it was obtained in is specified. Rather than publish data for every possible atmosphere, the industry publishes it for *one* and corrects to the conditions of the day. A question offering "the atmosphere which can be used worldwide to provide comparable performance results" against alternatives describing a particular place or a particular season is testing exactly this: the ISA's purpose is comparability, and every other property it has follows from that purpose.

The ISA never predicts the weather and is not intended to. The actual atmosphere matches it only by coincidence and usually not even then. Any answer describing the ISA as a forecast, or as "the conditions observed at a given place and time", has inverted the model's whole purpose — it is a fixed yardstick, and a yardstick that changed with the weather would be useless.

The assumptions the model is built on

The ISA is a physical model with explicit simplifications, and knowing them explains both its accuracy and its limits:

- **The air is dry.** Humidity is taken as zero throughout. This is why the standard density of 1.225 kg/m³ is a *dry*-air figure, and why humidity has to be handled separately as a correction rather than appearing anywhere in the standard tables.

- **The air is a perfect gas** of constant composition, with a fixed mean molar mass of 28.96 g/mol and hence a fixed specific gas constant of 287.05 J/(kg·K). This holds throughout the homosphere, which is well above the range the model is used in.
- **The atmosphere is in hydrostatic equilibrium** — there is no vertical motion, and at every level the pressure exactly supports the weight of the air above. This is what makes the pressure at any altitude calculable from the temperature profile alone.
- **Gravitational acceleration is constant** at 9.80665 m/s². In reality it falls slightly with height, and the model absorbs that by expressing altitudes as *geopotential* rather than geometric height. The two differ by about 60 ft at the tropopause and by under 100 ft at 45,000 ft, less than 0.3 % of the height in either case, so for practical purposes the distinction can be ignored — but it is the reason the tropopause is quoted as exactly 11,000 m, a defined geopotential altitude, rather than a measured one.
- **The temperature profile is piecewise linear**, with the lapse rate constant within each band and changing abruptly at the boundaries. The real atmosphere has no such corners.
- **There is no wind**, and the air is still.

None of these is true of the real atmosphere on any given day. That is not a defect in the model; it is what makes the model a *standard*. The full model as published extends to about 80 km, far above any aircraft; aviation uses only its lowest 15 km or so.

ISA Sea-Level Conditions

The ISA defines the following values at **mean sea level (MSL)**:

Property	ISA Value	Equivalent Values
Temperature	15 °C	288.15 K / 59 °F
Pressure	1013.25 hPa (mb)	29.92 inHg / 760 mmHg / 14.696 psi / 101,325 Pa / 1 atm
Density	1.225 kg/m³	0.002377 slugs/ft³
Temperature lapse rate	–6.5 °C per 1,000 m	≈ –1.98 °C per 1,000 ft
Tropopause altitude	11,000 m	36,089 ft
Tropopause temperature	–56.5 °C	216.65 K
Speed of sound at MSL	340.3 m/s	661.5 knots / 1,116 ft/s
Gravitational acceleration	9.80665 m/s²	32.174 ft/s²

Memory aid for the ISA baseline: Remember "15-1013-1.225" — that is 15 °C temperature, 1013.25 hPa pressure, 1.225 kg/m³ density. These three numbers are the foundation of the ISA and appear frequently in exam questions.

Where the sea-level numbers come from

Only two of the values in the table above are independent. Temperature and pressure are *defined*; everything else in the column is *derived* from them, and being able to reproduce the derivations is more useful than

memorising the results.

The standard density is not an independent choice. Apply the equation of state to the two defined values:

$$\rho = \frac{p}{RT} = \frac{101325}{287.05 \times 288.15} = \frac{101325}{82713} = 1.225 \text{ kg/m}^3$$

The published 1.225 kg/m³ is exactly what the defined pressure and temperature produce; it was not measured and then adopted.

The standard speed of sound likewise follows from the temperature alone:

$$a = \sqrt{\gamma RT} = \sqrt{1.4 \times 287.05 \times 288.15} = \sqrt{115799} = 340.3 \text{ m/s}$$

Converting: $340.3 \div 0.5144 = 661.5$ kt, and $340.3 \div 0.3048 = 1116$ ft/s — both of which appear in the table above. Note that neither the pressure nor the density enters the calculation.

Why the pressure is quoted in so many different units

The equivalences listed in the table are all the same physical pressure expressed against different reference quantities, and each has a physical origin worth knowing:

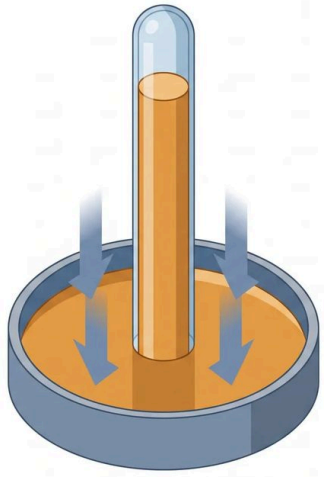
- **760 mmHg** is the length of the column of mercury that standard atmospheric pressure will support. The arithmetic is a direct application of pressure as force per unit area: mercury has a density of about 13,595 kg/m³, so a column 0.76 m tall exerts $13595 \times 9.80665 \times 0.76 = 101{,}325$ Pa at its base. That is the original definition of one atmosphere and the reason mercury barometers are calibrated in length units at all.
- **29.92 in Hg** is the same column measured in inches: $760 \div 25.4 = 29.92$.
- **1013.25 hPa** is the SI expression of the same pressure. The hectopascal and the millibar are *identical* — both are 100 pascals — so a conversion between them is a change of name, not of value. Meteorology adopted the hectopascal to become SI-consistent while keeping every existing chart's numbers unchanged.
- **14.696 psi** is the same pressure in imperial engineering units, which is where most aircraft system gauges live.
- **1 atm** and **1.01325 bar** are simply names for the whole quantity.

Order-of-magnitude checks that catch the common distractors. Sea-level pressure is *about a thousand* millibars, *about thirty* inches of mercury and *about fifteen* pounds per square inch. An option offering 1.013 mb is a thousand times too small; 14.7 kPa is a hundred times too small (14.7 psi is 101 kPa, not 14.7 kPa); 100 mb is a tenth of an atmosphere, roughly the pressure at 53,000 ft; and 10 bar is ten atmospheres, a pressure found in a hydraulic accumulator rather than the sky. Each of these appears as an option in the bank, and each is identifiable without knowing the exact standard value.

The instrument behind the number: barometers

A **barometer** is an instrument that measures atmospheric pressure — that and nothing else. It does not measure density, and it does not measure temperature, though its reading is used to *infer* altitude and, with a

temperature measurement, density. Two constructions are met:



A mercury barometer: air pressing on the open dish holds up a mercury column beneath a vacuum.

- **The mercury barometer** balances the atmosphere against a column of mercury in an evacuated tube. Because the space above the mercury is a vacuum, it reads **absolute** pressure directly, which is why atmospheric pressures are quoted as absolute values.
- **The aneroid barometer** uses a sealed, evacuated, thin-walled metal capsule that flexes as the surrounding pressure changes, its deflection driving a pointer through a mechanical linkage. It needs no liquid, tolerates vibration and can be made small, which is why every pressure altimeter and every vertical speed indicator ever fitted to an aircraft is an aneroid instrument at heart. The altimeter is, quite literally, a barometer with a scale marked in feet instead of hectopascals.

ISA Temperature Lapse Rate

The ISA assumes a **constant temperature lapse rate of 6.5 °C per 1,000 metres** (approximately 2 °C per 1,000 feet) throughout the troposphere. This means that for every 1,000 m you climb, the temperature drops by 6.5 °C. The lapse rate continues until the tropopause is reached at 11,000 m (36,089 ft), where the temperature has fallen to **–56.5 °C**. Above the tropopause, in the ISA, the temperature remains constant at –56.5 °C throughout the lower stratosphere (up to 20 km).

ISA temperature at any altitude below the tropopause:

$T_{^\circ C} = 15 - (6.5 \times h_{km})$

Or equivalently: $T_{^\circ C} = 15 - (2 \times h_{thousands\ of\ ft})$

Where T is temperature in °C, h_{km} is altitude in kilometres, and $h_{thousands\ of\ ft}$ is altitude in thousands of feet.

Example 1: What is the ISA temperature at 25,000 ft?

$T = 15 - (2 \times 25) = 15 - 50 = -35\ ^\circ C$

Example 2: What is the ISA temperature at 40,000 ft?

40,000 ft is above the tropopause (36,089 ft), so the ISA temperature is constant at $-56.5\ ^{\circ}C$.

(We do NOT continue to apply the lapse rate above the tropopause.)

The word "lapse rate", and the three numbers that are the same number

Used without qualification, **lapse rate always means a rate of change of temperature with height**. Pressure and density change with altitude too, and change faster, but neither has ever been called a lapse rate. If a question asks what quantity a lapse rate refers to, the answer is temperature.

The standard tropospheric lapse rate is quoted three different ways and the three are arithmetically identical, differing only in units:

Form	Where you meet it	Conversion
6.5 °C per 1,000 m	The SI definition of the model; also written 6.5 K/km	The defining value
1.98 °C per 1,000 ft	Performance and altimetry work in feet	$6.5 \times 0.3048 = 1.9812$
2 °C per 1,000 ft	Mental arithmetic and most examination questions	1.98 rounded

The 1 % error introduced by rounding 1.98 to 2 is worth about 0.7 °C at the tropopause, which is immaterial for anything an examination will ask. Use 2 °C per 1,000 ft to reason, and recognise 1.98 as the same rate when it is offered as an option. Watch for two traps in the options: a rate quoted in **Fahrenheit** rather than Celsius is a different quantity even when the number is the same, and 4 °C per 1,000 ft is roughly double the true value.

The complete ISA temperature profile

The troposphere is only the first of the model's bands. The full profile, which is what the published tables are computed from, is:

Base of band	In feet	Base temperature	Lapse rate	Behaviour
0 km (sea level)	0 ft	+15.0 °C (288.15 K)	–6.5 °C/km	Temperature falls
11 km	36,089 ft	–56.5 °C (216.65 K)	0	Isothermal
20 km	65,617 ft	–56.5 °C	+1.0 °C/km	Temperature rises slowly
32 km	104,987 ft	–44.5 °C	+2.8 °C/km	Temperature rises faster
47 km	154,199 ft	–2.5 °C	0	Isothermal
51 km	167,323 ft	–2.5 °C	–2.8 °C/km	Temperature falls

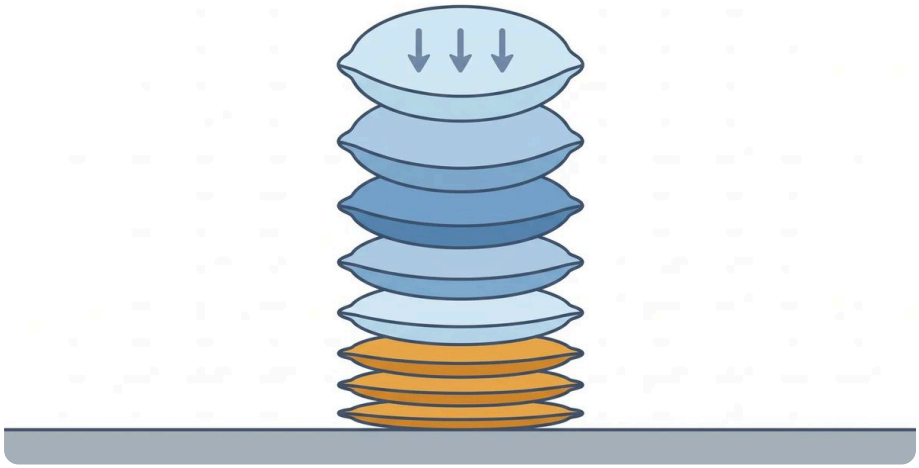
Everything above the second row is outside the operating range of any aircraft in service, so for practical work only the first two bands matter. They are, however, the source of a distinction the examination does test: below 36,089 ft use the lapse rate; between 36,089 ft and about 65,600 ft the temperature is a constant –56.5 °C; above that it begins to increase again.

Linear temperature, exponential pressure

A point of vocabulary that is worth being exact about, because it is asked directly. In the standard atmosphere:

- **Temperature varies LINEARLY with altitude** within each band. The lapse rate is a constant number of degrees per thousand feet, so a graph of temperature against altitude is a straight line in the troposphere. This is a deliberate choice of the model, made because it is simple and because it fits the real average profile well.
- **Pressure and density vary EXPONENTIALLY with altitude.** Neither is a straight line, and neither could be: each layer of air is compressed by the weight of everything above it, so the same height increment removes far more pressure low down, where the air is dense, than high up where it is thin.

So a question asking whether the ISA lapse rate is linear, exponential or non-linear is asking about *temperature*, and the answer is linear — even though the very same model makes pressure exponential. The two statements sit side by side without contradiction because they describe different quantities.



Each layer of air is squashed by the weight above it, so pressure falls fastest close to the ground.

Worked examples using the 2 °C rule. In every case start at +15 °C and subtract 2 °C for each 1,000 ft, stopping at the tropopause:

- 10,000 ft: $15 - (2 \times 10) = -5\ ^{\circ}\mathrm{C}$ (the exact figure using 1.98 is –4.8 °C)
- 20,000 ft: $15 - (2 \times 20) = -25\ ^{\circ}\mathrm{C}$ (exactly –24.6 °C)
- 30,000 ft: $15 - (2 \times 30) = -45\ ^{\circ}\mathrm{C}$ (exactly –44.4 °C)
- FL350: $15 - (2 \times 35) = -55\ ^{\circ}\mathrm{C}$ (exactly –54.3 °C, and still below the tropopause, so the lapse rate does apply)
- FL450: above the tropopause, so the answer is –56.5 °C and the lapse rate is **not** applied

The rounding always makes the two-degree answer slightly colder than the exact one, and the gap grows with altitude — 0.2 °C at 10,000 ft, 0.6 °C at 30,000 ft, 0.7 °C at FL350. It never becomes large enough to change which option is correct.

A frequent error is applying the lapse rate above the tropopause. Continuing the 2 °C rule to FL450 gives –75 °C, which is colder than the standard atmosphere ever becomes anywhere below the mesosphere. Before applying the rule, check the altitude against 36,089 ft; above it, the standard temperature is a constant until roughly 65,600 ft.

The standard atmosphere as a table

The model evaluated at intervals is what an engineer normally works from. The table below extends the pressure table given later in this note with the temperature, density, density ratio and speed of sound at each level. It covers the whole range in which aircraft operate, and it is worth reading across a row as well as down a column: at every altitude the density ratio is the pressure ratio divided by the temperature ratio, and the speed of sound is fixed by the temperature alone.

Altitude (ft)	Temperature (°C)	Pressure (hPa)	Density (kg/m³)	Density ratio σ	Speed of sound (kt)
0	+15.0	1013.2	1.225	1.000	661.5
2,000	+11.0	942.1	1.155	0.943	656.9
5,000	+5.1	843.1	1.056	0.862	650.0
8,000	–0.8	752.6	0.963	0.786	643.0
10,000	–4.8	696.8	0.905	0.738	638.3
20,000	–24.6	465.6	0.653	0.533	614.3
25,000	–34.5	376.0	0.549	0.448	601.9
30,000	–44.4	300.9	0.458	0.374	589.3
35,000	–54.3	238.4	0.380	0.310	576.4
36,089	–56.5	226.3	0.364	0.297	573.6
40,000	–56.5	187.5	0.302	0.246	573.6
45,000	–56.5	147.5	0.237	0.194	573.6
50,000	–56.5	116.0	0.186	0.152	573.6
60,000	–56.5	71.7	0.115	0.094	573.6
65,000	–56.5	56.4	0.091	0.074	573.6

Four features of the table repay attention, and each of them puts something the note states in prose into a form that can be checked at a glance:

- **The temperature column stops changing at 36,089 ft** and every row below that in the table carries the same –56.5 °C. So does the speed-of-sound column, for the same reason — the two are locked together.
- **The pressure column never stops falling.** It halves between sea level and 18,000 ft, halves again by about 36,000 ft, and goes on halving with no regard whatever for the tropopause.

- **The density ratio falls more slowly than the pressure ratio below the tropopause and at the same rate above it.** Compare the 30,000 ft row, where the pressure has fallen to 29.7 % and the density only to 37.4 %, with the interval from 40,000 ft to 60,000 ft, where pressure falls by a factor of 2.62 and density by the same 2.62. Below the tropopause the cooling props the density up; above it, with the temperature fixed, it cannot.
- **Half the atmosphere is below 18,000 ft.** That altitude is not repeated here because the pressure table later in this note carries it, at 50 % of the sea-level value — and as established above, that percentage is the fraction of the atmosphere's mass still lying above.

Why the tropopause pressure is sometimes quoted as 226 hPa and sometimes as 227. Both are the same standard atmosphere read at slightly different heights. The tropopause is defined at exactly **11,000 m**, which is **36,089 ft**, and the pressure there is 226.3 hPa. Tables that round the tropopause to a flat 36,000 ft are reading the model 89 ft lower down, and at that level one hectopascal is worth about 92 ft — so 89 ft is worth almost exactly 1 hPa, and the figure comes out at 227.3. The two numbers differ by the rounding of the altitude and by nothing else, and the arithmetic that connects them is the same pressure-per-foot relationship developed later in this note.

Interpolating between rows. Temperature may be interpolated linearly with confidence, because the model itself is linear in each band. Pressure and density may not, strictly, because they are exponential — but over a 2,000 or 5,000 ft interval the curvature is small and linear interpolation is accurate to well within one per cent, which is better than the atmosphere itself will oblige you with on any given day. For anything requiring more, use the equation of state with the interpolated temperature and a properly computed pressure.

ISA Deviation

In the real world, the actual atmospheric temperature rarely matches the ISA exactly. The difference between the actual (observed) temperature and the ISA temperature at the same altitude is called the **ISA deviation**, expressed as **ISA ± X °C**.

$$\text{ISA deviation} = T_{actual} - T_{ISA}$$

Example 3: The actual temperature at 20,000 ft is –15 °C. What is the ISA deviation?

ISA temperature at 20,000 ft = $15 - (2 \times 20) = -25\ ^\circ C$

ISA deviation = $-15 - (-25) = +10\ ^\circ C$

This is reported as **ISA + 10** — it is 10 °C warmer than ISA.

Example 4: The actual temperature at 10,000 ft is –12 °C. What is the ISA deviation?

ISA temperature at 10,000 ft = $15 - (2 \times 10) = -5\ ^\circ C$

ISA deviation = $-12 - (-5) = -7\ ^\circ C$

This is reported as **ISA – 7** — it is 7 °C colder than ISA.

ISA deviation is important because aircraft performance data (take-off distances, climb rates, engine thrust ratings) are published for ISA conditions. When the actual conditions differ, corrections must be applied. A positive ISA deviation (warmer than standard) means **reduced performance** — lower air density, less engine thrust, longer take-off distances, and reduced climb rates.

Two errors that account for most wrong answers on ISA deviation

ISA deviation is a subtraction, and both of the ways it goes wrong are failures to perform that subtraction rather than failures of physics.

Error one — quoting the actual temperature as the deviation. The actual sea-level temperature is 30 °C. What is the ISA deviation?

The standard sea-level temperature is 15 °C, so

$30 - 15 = +15\ ^\circ\mathrm{C}$, that is **ISA + 15**.

Not ISA + 30. The deviation is a *difference* from standard, and the two coincide only when the standard value happens to be zero, which it never is at sea level.

Error two — forgetting to compute the standard value for the altitude. The actual temperature at FL240 is –40 °C. What is the deviation?

First find the standard temperature *at that altitude*: $15 - (2 \times 24) = -33\ ^\circ\mathrm{C}$.

Then subtract: $-40 - (-33) = -7\ ^\circ\mathrm{C}$, that is **ISA – 7**.

Comparing –40 °C against the sea-level standard of 15 °C instead would have given ISA – 55, which is nonsense: it would describe an atmosphere far colder than any ever recorded. The check for this error is arithmetical rather than climatological. Using the sea-level standard in place of the standard for the altitude shifts the answer by exactly the difference between the two, which is 2 °C for every 1,000 ft — 48 °C at FL240, and 70 °C at FL350. Size alone will not settle it, because real deviations do run large: the regional table later in this note gives ISA + 15 to ISA + 30 near the tropical surface and ISA – 30 and beyond in polar winter, and Problem 4 at the end of this note works a genuine ISA – 36. What settles it is recomputing the standard value for the stated altitude, which takes one multiplication.

Signs, in words. A *positive* deviation means the air is **warmer** than standard, therefore **less dense** than standard, therefore performance is **worse** than the book figure. A *negative* deviation means colder, denser and better than book. Warm is always the bad direction for performance, and that single association resolves a large number of questions without any arithmetic at all.

Relative density, and the three standard ratios

Performance engineering rarely works with absolute values of pressure, density and temperature. It works with each one expressed as a **ratio to its own standard sea-level value**, because ratios are dimensionless, are the same in every unit system, and combine cleanly. Three ratios are defined, and each carries a Greek letter that appears throughout performance documentation:

Symbol	Name	Definition	Value at sea level in ISA	Behaviour with increasing altitude
σ (sigma)	Density ratio, also called relative density	Actual density ÷ 1.225 kg/m³	1.000	Falls
δ (delta)	Pressure ratio	Actual pressure ÷ 1013.25 hPa	1.000	Falls
θ (theta)	Temperature ratio	Actual absolute temperature ÷ 288.15 K	1.000	Falls in the troposphere, then constant

Three points about them are examined. **All three are dimensionless**, being ratios of like quantities. **Theta must be formed from absolute temperatures**: a ratio of Celsius values is meaningless and can be negative or infinite, so 288.15 K is always the denominator. And **all three are equal to 1 at sea level in standard conditions and less than 1 above it**, so in the standard atmosphere a pressure ratio greater than 1 is impossible — there is nowhere above sea level where the standard pressure exceeds the standard sea-level value. On a real day the ratio can exceed 1 near sea level, because the actual pressure does whenever the QNH is above 1013.25 hPa: a QNH of 1030 hPa at a sea-level aerodrome gives a pressure ratio of 1.017.

They are not independent. Dividing the equation of state at altitude by the same equation at sea level gives a relationship worth remembering because it lets any one ratio be recovered from the other two:

$$\sigma = \frac{\delta}{\theta}$$

Density ratio equals pressure ratio divided by temperature ratio.

Altitude	θ (temperature ratio)	δ (pressure ratio)	σ (density ratio)	$1/\sqrt{\sigma}$
Sea level	1.000	1.000	1.000	1.000
10,000 ft	0.931	0.688	0.738	1.164
20,000 ft	0.862	0.460	0.533	1.370
30,000 ft	0.794	0.297	0.374	1.635
35,000 ft	0.759	0.235	0.310	1.796
40,000 ft	0.752	0.185	0.246	2.015

Check the identity on the 30,000 ft row: $\delta/\theta = 0.297/0.794 = 0.374$, which is the tabulated density ratio. The check works on every row — allow a unit in the last printed digit, since every figure in the table is itself rounded — and is a quick way of confirming that a value has been read from the correct column of a set of tables.

The final column is the quantity that converts equivalent airspeed to true airspeed, and it is the reason the density ratio is the most used of the three. It is also worth noticing what the theta column does at the bottom: its fall almost stops between 35,000 ft and 40,000 ft, dropping only 0.007 against 0.035 over the previous

5,000 ft. The reason is that only the first 1,089 ft of that interval lies below the tropopause at 36,089 ft; above the tropopause the standard temperature is constant, so theta stops changing altogether. Delta and sigma keep falling, because pressure and density keep falling; only the temperature-derived ratio flattens.

Standard atmosphere tables and how they are used

The published ISA tables are simply this model evaluated at intervals of altitude, tabulating temperature, pressure, density, the three ratios and the local speed of sound. An engineer uses them in one of two directions:

- **Forwards**, to find the standard conditions at a stated altitude — the reference against which a measured value is compared, and the source of the ISA temperature used in every deviation calculation.
- **Backwards**, to find the altitude at which a stated condition would occur. This is exactly what pressure altitude and density altitude are: the tables read in reverse, from a pressure or a density to the altitude that matches it.

The reverse direction is the more important one conceptually, because it is what the aircraft's own instruments do. A pressure altimeter measures a pressure and looks up the ISA altitude that corresponds to it; that is the whole of its operating principle, and it explains why every discrepancy between the altimeter and the real world can be traced to a difference between the real atmosphere and the model.

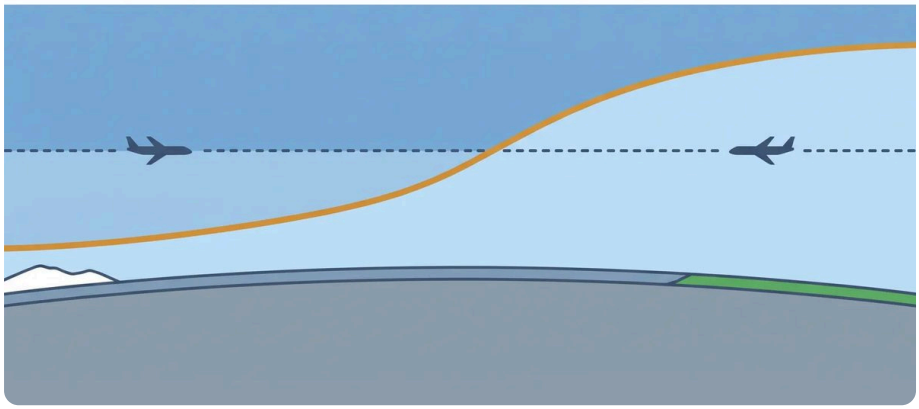
How far the real atmosphere departs from the model

Knowing the standard atmosphere is only half of the skill; knowing how far reality strays from it is the other half, because that is what determines whether a correction is a formality or the dominant term in a calculation. The departures are systematic rather than random, and they follow the general circulation described later in this note:

Region and season	Sea-level temperature against ISA	Tropopause height	Tropopause temperature	Practical consequence
Tropical	Warmer, commonly ISA + 15 to ISA + 30 near the surface	High — around 16 to 18 km	Very cold, well below the standard –56.5 °C	Severe take-off performance penalties; the lapse rate applies to a much greater height, so cruise levels are still in the troposphere and still cooling
Mid-latitude, summer	Warmer than standard	Near or a little above the standard 11 km	Near standard	The standard atmosphere is a reasonable approximation; corrections are modest
Mid-latitude, winter	Colder than standard	Near or a little below 11 km	Near standard	Altimeter temperature error becomes significant; performance is better than book

Region and season	Sea-level temperature against ISA	Tropopause height	Tropopause temperature	Practical consequence
Polar, winter	Far colder, ISA – 30 and beyond at the surface	Low — around 8 km	Warmer than the tropical tropopause	Large altimeter temperature errors and mandatory cold-temperature corrections; cruise levels are well into the stratosphere

Two of those rows contain results that look wrong until the mechanism is recalled. The tropical tropopause is the *coldest*, because the lapse rate operates over nearly twice the height even though it starts from a much warmer surface. And in polar winter an aircraft at a normal cruising flight level may be well above the tropopause, in genuinely isothermal air, while the same flight level over the tropics is still in the troposphere with the temperature continuing to fall. **The flight level is the same; the layer of the atmosphere it lies in is not.**



The same flight level lies above the low polar tropopause but below the high tropical one.

The one departure that must never be handled by rule of thumb. In extreme cold the altimeter temperature error becomes large enough to consume the entire obstacle clearance margin of an instrument approach, and the direction is the dangerous one: the altimeter over-reads, so the aircraft is lower than indicated. Below the aerodrome temperatures at which corrections become mandatory, published minima are corrected from a table rather than estimated, and the 4 % per 10 °C rule described below is a way of understanding the size of the effect rather than a way of computing an approach minimum.

Application of ISA to Aerodynamics

The ISA is not just a theoretical curiosity — it is the practical foundation upon which aircraft are designed, certified, and operated. Here are the major applications:

A useful way to hold the four applications below together is to notice that every one of them is the same operation performed on a different quantity. In each case a real measurement is made in a real atmosphere;

the standard atmosphere is then used to convert it into a form that can be compared with something. Performance data is corrected to standard conditions so that two aircraft, or the same aircraft on two days, can be compared. An altimeter converts a measured pressure into the altitude that pressure would have in the standard atmosphere. An airspeed indicator converts a measured pressure difference into the speed that difference would represent at standard sea-level density. And an engine's measured parameters are corrected to standard conditions before being judged against a limit. **The standard atmosphere is a common currency, and all four applications are exchange operations into and out of it.**

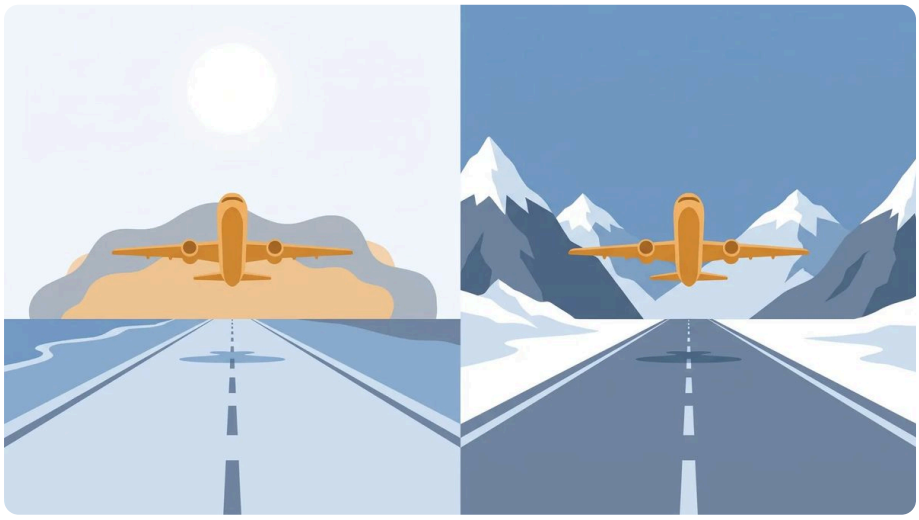
1. Aircraft Performance Calculations

All performance data in the Aircraft Flight Manual (AFM) — take-off distances, landing distances, climb gradients, cruise performance, ceiling — are **calculated and published for ISA conditions**. When the actual conditions differ, the pilot (or performance engineer) must apply corrections. For example:

- On a **hot day (ISA+20)**, the air density is significantly lower than standard. An aircraft that needs 1,800 m of runway at ISA conditions might need 2,200 m or more at ISA+20.
- At a **high-altitude airport** (e.g., Addis Ababa at 7,625 ft), the lower pressure means lower density, requiring longer take-off runs and reducing climb performance.
- When both high temperature AND high altitude combine (e.g., a hot day at a high-altitude airport), the performance penalties are **compounded**, potentially making operations unsafe if not properly accounted for.

Density altitude is the single variable that carries all of this

Field elevation, ambient pressure, temperature and humidity all affect performance, but they do not affect it in four independent ways. They act through *one* quantity — air density — and performance work therefore reduces them to a single number, the density altitude. Once the density altitude is known, the elevation and the temperature that produced it cease to matter: an aircraft at a sea-level airfield on a very hot day and an aircraft at a mountain airfield on a cold day will perform identically if their density altitudes are the same.



A hot sea-level airfield and a cold mountain airfield with the same density altitude give the same take-off.

Worked example — a hot day at a high airfield. An airfield lies at 5,000 ft elevation. The QNH is 1013 hPa, so the pressure altitude is also 5,000 ft. The outside air temperature is +30 °C.

Standard temperature at 5,000 ft: $15 - (2 \times 5) = +5\ ^\circ\mathrm{C}$.

ISA deviation: $30 - 5 = +25\ ^\circ\mathrm{C}$, that is ISA + 25.

Density altitude: $5000 + (25 \times 120) = 5000 + 3000 = 8{,}000$ ft.

The aircraft will accelerate, lift off, climb and (if it is piston-powered) produce power as though it were at 8,000 ft in standard conditions. A full calculation from the equation of state gives about 7,800 ft, so the 120 ft-per-degree rule over-estimates by roughly 200 ft here — it errs on the pessimistic side, which is the right direction for a rule of thumb used for performance planning.

Why the take-off penalty is worse than the density loss alone suggests

It is tempting to assume that if the density has fallen by 25 %, the take-off run must be about 25 % longer. It is worse than that, because low density penalises the take-off **twice over**, and the two penalties multiply rather than add:

- **The aircraft must reach a higher TRUE airspeed.** Lift-off occurs at a fixed indicated airspeed, because the wing responds to dynamic pressure. But a fixed indicated airspeed in thin air means a higher true airspeed, in proportion to $1/\sqrt{\sigma}$. The kinetic energy that has to be built up on the runway therefore rises in proportion to $1/\sigma$.
- **The thrust available to build it is smaller.** Thrust falls roughly in proportion to density, so the acceleration available is reduced by a factor of about σ as well.

Energy required up by $1/\sigma$, and the rate of supplying it down by σ, gives a ground roll that scales roughly with $1/\sigma^2$ — a squared sensitivity, not a linear one. That is why performance charts for hot-and-high conditions are so much steeper than intuition suggests, and it is the quantitative reason the compounding effect described above is dangerous. Charted figures are usually worse still than this simple scaling predicts, because rolling friction does not fall with density, the engine's installed thrust does not fall exactly in proportion to it, and the aircraft spends longer in the region where drag is highest.

Density altitude is invisible from the cockpit and from the ramp. Nothing in the aircraft displays it, no gauge moves, and the air does not look or feel any different. A pilot who has operated from a particular runway a hundred times in temperate conditions has no sensory warning that the same runway on a hot afternoon at a high elevation is a materially different problem. This is why density altitude is computed rather than judged, and it is the reason accident reports involving take-off performance so frequently list it as a factor.

The rest of the flight, and why the effects do not cancel

Low density does not merely delay the take-off; it degrades every subsequent phase, and it is worth being explicit about the direction in each case:

Phase	Effect of reduced density	Mechanism
Take-off run	Substantially **longer**	Higher true lift-off speed against reduced thrust, as above

Phase	Effect of reduced density	Mechanism
Initial climb	Rate and gradient **reduced**	Rate of climb depends on excess power; thrust falls while drag at a given indicated speed does not
Obstacle clearance	Margin **reduced** twice over	The aircraft is both further down the runway at lift-off and climbing more shallowly afterwards
Cruise	True airspeed **higher** for the same indicated speed	The one effect that works in the operator's favour, and the reason jets cruise high
Ceiling	**Lower**	Thrust falls faster with altitude than drag does, so the excess runs out sooner
Approach and landing	True airspeed at touchdown **higher**; landing roll **longer**	Approach is flown at a fixed indicated speed, so the true speed and the ground speed are higher
Braking	Energy to be absorbed **higher**	Kinetic energy goes with the square of the true ground speed, so a modest speed increase is a large energy increase

Only the cruise row is favourable, and it is the reason the whole business is worth tolerating: an aircraft cruising at 35,000 ft is flying through air of roughly 31 % sea-level density, which lets it achieve a true airspeed of nearly 1.8 times its indicated speed for the same dynamic pressure and therefore for very nearly the same drag. High-altitude cruise is a deliberate exploitation of exactly the property that makes take-off from a high airfield difficult.

Where the published performance figures come from

Performance data does not appear in a flight manual as a set of measurements taken on the day of a test flight. It is measured under whatever conditions the test happened to encounter, reduced to standard conditions using exactly the ratios described earlier in this note, and then expanded back out into charts covering the full range of conditions the aircraft is approved for. Two points follow that are worth understanding when reading such a chart:

- **The chart axes are pressure altitude and temperature, not elevation and weather.** This is deliberate. Pressure altitude and temperature between them determine the density, and density is what the aircraft responds to. Entering a chart with a field elevation instead of a pressure altitude produces an error in the direction of optimism whenever the QNH is below standard, which is precisely when the conditions are already poor.
- **Certification requires demonstration across a range of atmospheres**, not merely at the standard one. Hot-day and cold-day conditions are addressed explicitly, because an aircraft's limitations are frequently set by the extremes rather than by the standard case — engine temperature limits in hot conditions, and systems and structural considerations in very cold ones.

Flat rating: why thrust does not always follow the day

An engine's thrust would, on the reasoning above, fall steadily as the ambient temperature rises, because the mass flow falls. Many engines nevertheless deliver a *constant* take-off thrust across a range of ambient temperatures, and it is worth being clear why, because it looks at first like a contradiction of the physics.

The engine is not being made to produce more thrust than the air allows; it is being **restricted below its capability at low temperatures**. The rated thrust is chosen at some reference temperature, and at any lower temperature the engine could produce more than that but is held back by its control system. Only above that reference temperature — the corner point, or flat-rating temperature — does the thrust begin to fall away with temperature in the manner the mass-flow argument predicts.

The benefit is not primarily to the operator's convenience. Holding the thrust constant means the turbine entry temperature does not have to be pushed to its limit on a cold day to reach a rating that the aircraft does not need, so the engine spends most of its life a long way below its temperature limits. That is a direct extension of hot-section life, and it is why a flat-rated engine is deliberately operated below its potential.

The maintenance consequence. Below the corner point, take-off thrust is constant and a hotter day produces a *higher* turbine temperature for the same thrust, because the engine must work harder to reach the same target in thinner air. Above the corner point, thrust falls and the temperature is held. This is why an engine's exhaust gas temperature margin — the gap between the temperature it reaches at full rated thrust and the certified limit — is measured and trended against ambient conditions, and why a margin measured on a cold morning cannot be compared directly with one measured on a hot afternoon without correcting both to the same reference.

2. Altimeter Calibration and Pressure Altitude

A **pressure altimeter** is essentially a calibrated barometer — it measures atmospheric pressure and converts it to an altitude reading using the ISA pressure-altitude relationship. If the actual atmospheric conditions match the ISA exactly, the altimeter reads correctly. In reality, they never match exactly, so corrections are needed:

- **QNH setting** — the altimeter is adjusted to show altitude above *mean sea level* by setting the actual sea-level pressure on the altimeter's subscale. This corrects for the difference between the actual sea-level pressure and the ISA standard of 1013.25 hPa.
- **QFE setting** — the altimeter is adjusted to show *zero* on the ground at the airfield (height above the airfield).
- **Standard pressure setting (1013.25 hPa)** — above the **transition altitude**, all aircraft set their altimeters to 1013.25 hPa. The resulting reading is called **pressure altitude** or **flight level**. This ensures vertical separation between aircraft regardless of local pressure variations.
- **Temperature error** — the altimeter corrects for pressure but *not* for temperature. In cold air (ISA–), the aircraft is **lower than the altimeter indicates** (because cold, dense air compresses the pressure levels closer together). This is summarised by the rule: **"From high to low (temperature), look out below."**

The altimeter is a standard-atmosphere computer

It is worth stating plainly what the instrument does, because every altimeter error follows from it. A pressure altimeter senses one quantity, the ambient static pressure, and displays another, an altitude. The conversion between them is not a measurement — it is a **look-up in the standard atmosphere**, built into the instrument's mechanism as the shape of its aneroid capsule and the geometry of its linkage. The altimeter answers the question "at what altitude would this pressure be found in the standard atmosphere?", and it answers it correctly every time. What it cannot know is whether today's atmosphere is the standard one.

The subscale does not change that. Turning the subscale shifts the *datum* from which the standard profile is applied — it tells the instrument what pressure to treat as the zero of the scale — but above that datum the instrument still assumes standard temperatures and the standard lapse rate. This is why setting a correct QNH removes the error caused by non-standard *pressure* at sea level but does nothing whatever about non-standard *temperature*, which is dealt with next.

Temperature error: the mechanism, and its direction

Pressure falls with height at a rate set by the density of the air, since each metre of climb sheds the weight of one metre of air column. Cold air is dense, so in cold air the pressure falls **more** rapidly per foot of climb than the standard atmosphere assumes. Work one case through:

Direction check on a concrete case. An aircraft has QNH set correctly and departs an airfield in air much colder than standard. It climbs to a *true* height of 1,000 ft above the field.

Over that 1,000 ft, the dense cold air has produced a larger pressure drop than the standard atmosphere would produce over the same 1,000 ft. The altimeter, which knows only the pressure drop and converts it using the standard scale, therefore concludes that the aircraft has climbed *more* than 1,000 ft.

The altimeter **over-reads**. The aircraft is **lower than indicated**. In warm air the whole argument runs the other way: the pressure drops more slowly, the altimeter under-reads, and the aircraft is higher than indicated — which is the safe direction and therefore the one nobody worries about.

The magnitude is commonly approximated as **4 % of the height above the setting datum for each 10 °C of ISA deviation**. On an instrument approach with a 2,000 ft platform altitude in air 20 °C colder than standard, that is about 160 ft of over-reading — enough to matter on an approach in mountainous terrain, which is why cold-temperature corrections are applied to published minima below specified aerodrome temperatures.

"From high to low, look out below" covers two separate errors, and both are true.

- **High temperature to low temperature:** flying into colder air without any change of setting, the altimeter over-reads and the aircraft is lower than indicated — the case worked through above.
- **High pressure to low pressure:** flying into an area of lower QNH while leaving the old, higher setting on the subscale, the altimeter again over-reads and the aircraft is again lower than indicated.

The two mechanisms are unrelated — one is a lapse-rate error, the other a datum error — but they produce the same hazard, and the mnemonic deliberately covers both.

A correctly reading altimeter is not the same as a correct altitude. In non-standard conditions the instrument is not faulty and no maintenance action will improve it; it is faithfully reporting the pressure it senses in the only terms it has. A reported discrepancy of about 400 ft, recorded 4,000 ft above the datum in air 25 °C below standard, is exactly what the 4 % rule predicts — that is physics being chased, not a fault. The functional test that distinguishes the two is a comparison against a known pressure applied by a pitot-static test set, not a comparison against a known height.

3. Airspeed Indicator Calibration

The airspeed indicator (ASI) measures the difference between **total pressure** (from the pitot tube) and **static pressure** (from the static port). This difference is called **dynamic pressure**, and it is related to airspeed by:

$$q = \frac{1}{2}\rho V^2$$

Where q = dynamic pressure (Pa), ρ = air density (kg/m³), V = true airspeed (m/s)

The ASI is **calibrated assuming ISA sea-level density** (1.225 kg/m³). The speed it displays is called **Indicated Airspeed (IAS)**. At sea level in ISA conditions, IAS equals True Airspeed (TAS). At altitude, where the actual density is lower than 1.225 kg/m³, the IAS under-reads compared to TAS. The various airspeed definitions are:

Airspeed	Abbreviation	Definition
Indicated Airspeed	IAS	The reading on the airspeed indicator (uncorrected)
Calibrated Airspeed	CAS	IAS corrected for instrument and position errors
Equivalent Airspeed	EAS	CAS corrected for compressibility effects (significant above ~250 kt and at high altitude)
True Airspeed	TAS	EAS corrected for density — the actual speed of the aircraft through the air mass. TAS is always higher than IAS at altitude.

Rule of thumb: TAS increases above IAS by approximately **2 % per 1,000 ft** of altitude. So at 30,000 ft, TAS is roughly 60 % higher than IAS. An aircraft indicating 250 kt IAS at FL300 is actually flying at about 400 kt TAS.

The three pressures the air data system works with

The equation above contains the airspeed, but the instrument cannot measure a speed. What it can measure are pressures, and three distinct ones are involved. Keeping them separate is the foundation of everything the pitot-static system does:

Pressure	Also called	What it is	Where it is sensed	Behaviour with altitude and speed
Static pressure p_s	Ambient pressure	The pressure of the undisturbed air, exerted **equally in all directions**, arising from the weight of the air column above	Static ports, set flush in a region of undisturbed flow	**Decreases** with altitude; ideally unaffected by the aircraft's speed
Dynamic pressure q	—	The pressure equivalent of the kinetic energy of the moving air, $\frac{1}{2}\rho V^2$	Not sensed directly — obtained by subtraction	**Increases** with the *square* of speed; **decreases** with density and hence with altitude

Pressure	Also called	What it is	Where it is sensed	Behaviour with altitude and speed
Total pressure p_t	Pitot pressure, stagnation pressure, ram pressure	The pressure reached when the moving air is brought completely to rest	Pitot tube, facing directly into the airflow	The sum of the other two

$p_t = p_s + q = p_s + \frac{1}{2}\rho V^2$
and therefore $q = p_t - p_s$

The second form is the one the airspeed indicator is built around. The instrument has the pitot line on one side of a capsule and the static line on the other, so the capsule deflects according to the *difference* between them, which is the dynamic pressure. **An airspeed indicator is a differential pressure gauge with a speed scale.** The static pressure itself — the much larger of the two quantities — cancels out entirely and never appears in the reading.

Get the sign of the total-pressure relationship the right way round. Total pressure is static *plus* dynamic. Bringing moving air to rest converts its kinetic energy into pressure, so the pitot tube always senses *more* than the surrounding static pressure, never less. An expression of the form $p_t = p_s - q$ would make the pitot pressure fall below ambient as the aircraft accelerated, which would mean air being sucked out of a forward-facing tube. In still air, with the aircraft stationary, q is zero and the pitot and static pressures are equal — which is exactly why an airspeed indicator reads zero on the ground.

Why indicated airspeed is the speed that is flown

An obvious question is why the aircraft is flown on a reading that is known to be wrong at altitude, when a true airspeed is available from the air data computer. The answer is that indicated airspeed is not an approximation to true airspeed at all — it is a direct measure of the quantity the airframe actually responds to.

Every aerodynamic force on the aircraft is proportional to dynamic pressure. Lift is $C_L \cdot q \cdot S$; drag is $C_D \cdot q \cdot S$; the hinge moment on a control surface, the load on a flap and the bending moment on a wing all scale with q in the same way. Since indicated airspeed is calibrated directly from q, it follows that:

- The **stalling speed** for a given weight and configuration is the same indicated airspeed at every altitude, even though the true airspeed at the stall may be half as much again at 30,000 ft as at sea level.
- The **flap and landing-gear limiting speeds**, and the manoeuvring speed, are structural limits set by q and are therefore quoted as indicated airspeeds.
- The **never-exceed speed** at low altitude is likewise an indicated speed, because it is a structural and flutter limit.

Flying on indicated airspeed therefore gives the pilot one set of numbers that is valid at every altitude and weight combination. True airspeed matters for navigation, for fuel planning and for the Mach number; it does not tell the pilot how close the wing is to stalling.

Where the exception lies. Compressibility does not scale with dynamic pressure but with Mach number, so at high altitude the limiting speed ceases to be an indicated airspeed and becomes a Mach number instead. This is why airspeed limits are published as an indicated speed below a certain altitude and as a Mach number above it, and why the two are equal at one crossover altitude.

The airspeed ladder in full

Working from the instrument outward, four corrections separate the reading from the aircraft's real speed through the air. The order is fixed, and each correction addresses one specific defect:

Step	Speed	Correction applied to reach it	What that correction removes	Size of the correction
1	**IAS**	— the raw reading	—	—
2	**CAS**	Instrument error and **position error**	Imperfections in the mechanism, and the fact that the static port cannot be placed anywhere the local pressure is exactly ambient	Small in the cruise; largest at high angle of attack, low speed and with flaps or gear extended
3	**EAS**	**Compressibility** correction	The pitot tube compresses the air it stops, so it senses slightly more than $\frac{1}{2}\rho V^2$; the error grows with speed and altitude	Negligible below about 200 kt at low level; significant above roughly 250 kt and at high altitude, and always *subtractive*
4	**TAS**	**Density** correction, $\mathrm{TAS} = \mathrm{EAS}/\sqrt{\sigma}$	The instrument's assumption of sea-level density	Zero at sea level in ISA; the largest correction of the four at altitude

A mnemonic for the order: I-C-E-T — Indicated, Calibrated, Equivalent, True — corrected in that sequence for **I**nstrument and position, **C**ompressibility, then density. Applying the density correction to IAS instead of to EAS is the usual shortcut, and it is acceptable at low speed and moderate altitude precisely because the first two corrections are then small.

Worked example — converting EAS to TAS. An aircraft is at 20,000 ft in standard conditions, showing 200 kt with instrument, position and compressibility corrections already applied — so 200 kt EAS. What is the true airspeed?

The density ratio at 20,000 ft is $\sigma = 0.533$, so

$$\mathrm{TAS} = \frac{\mathrm{EAS}}{\sqrt{\sigma}} = \frac{200}{\sqrt{0.533}} = \frac{200}{0.730} = 274 \text{ kt}.$$

Direction check: the air is thinner than at sea level, so the aircraft must travel *faster* through it to produce the same dynamic pressure and the same instrument reading. TAS must therefore exceed EAS, and 274 kt is duly greater than 200 kt. Had the arithmetic produced a smaller number, the calculation would have been multiplying by $\sqrt{\sigma}$ instead of dividing by it.

Worked example — the relationship at a FIXED altitude. An aircraft's true airspeed is halved. What happens to its indicated airspeed?

At a fixed altitude the density, and therefore σ, does not change. The relationship $\mathrm{EAS} = \mathrm{TAS} \times \sqrt{\sigma}$ is then a simple proportionality with a constant multiplier, so halving the true airspeed **halves** the indicated airspeed.

Note carefully that the dynamic pressure does *not* halve — it falls to a quarter, because q goes with V^2. But the airspeed indicator's scale is deliberately non-linear so that it displays the speed rather than the pressure, which restores the direct proportionality between the reading and the speed. Confusing the behaviour of the pressure with the behaviour of the scale is what produces the "reduced to a quarter" distractor.

How large the density correction gets, and where the rule of thumb fails

The two-per-cent-per-thousand-feet rule quoted above is a linear approximation to a relationship that is not linear, and it is worth knowing where it stops working. Comparing the rule against the exact factor $1/\sqrt{\sigma}$:

Altitude	Rule of thumb: TAS exceeds EAS by	Exact: TAS exceeds EAS by	Verdict
10,000 ft	20 %	16 %	Rule slightly over-states
20,000 ft	40 %	37 %	Good
30,000 ft	60 %	64 %	Good
40,000 ft	80 %	102 %	Rule badly under-states

The rule is at its best through the middle of the range, which is where it is normally applied, and it degrades above about FL350 as the exponential decay of density outruns the straight line. At 40,000 ft the true airspeed is very nearly *double* the equivalent airspeed — the density ratio has fallen to about a quarter, and the square root of a quarter is a half. Use the rule for mental estimates in the cruise band and the exact expression above it.

Ground speed sits at the end of the chain

One further step separates true airspeed from the speed that matters for navigation. True airspeed is measured relative to the air mass; the air mass is itself moving; so:

Ground speed = True airspeed ± the wind component along track

(subtract a headwind component, add a tailwind component)

The complete chain from instrument to map is therefore **IAS → CAS → EAS → TAS → ground speed**, with the first three steps correcting instrument and air properties and only the last one involving the wind. An engineer investigating an airspeed complaint needs to know which of those five quantities the crew were looking at, because a discrepancy between the ground speed shown by the navigation system and the airspeed shown on the pitot-static instrument is normally the wind doing exactly what it is supposed to do.

The wind triangle: what the moving air mass does to navigation

The relationship between true airspeed, ground speed and wind was stated above for the simple case of a wind blowing directly along the aircraft's track. In general the wind is at an angle to it, and the three quantities then combine as vectors rather than as numbers. The construction is called the **wind triangle** and its three sides are:

- **The air vector** — the aircraft's *heading* (the direction it is pointing) and its *true airspeed*. This is what the aircraft does relative to the air mass, and it is entirely under the pilot's control.
- **The wind vector** — the direction the wind is blowing from, and its speed. This is what the air mass does relative to the ground, and it is not under anyone's control.
- **The ground vector** — the aircraft's *track* (the path actually made good over the ground) and its *ground speed*. This is the sum of the other two, and it is what navigation is concerned with.

Two consequences are worth stating precisely because the vocabulary is easy to blur. **Heading and track are different directions**, and the angle between them is the *drift*: an aircraft flying in a crosswind must point upwind of its intended track in order to make that track good, and the amount by which it must do so is the drift angle. And **true airspeed and ground speed are different speeds**: only in perfectly still air are they equal, and only with a wind exactly along the track do they differ by the whole of the wind speed.

The aerodynamics never see the wind. The wing, the pitot tube, the engine and the airspeed indicator all respond to the air vector alone. Lift, drag, thrust, stalling speed and indicated airspeed are entirely unaffected by a steady wind of any strength or direction. What the wind changes is where the aircraft ends up and how long it takes to get there — a navigation and fuel problem, not an aerodynamic one. The single exception, worth repeating because it is the one that hurts, is a *changing* wind: shear and the wind gradient produce genuine, transient aerodynamic effects because the aircraft's inertia prevents it from adopting the new air-mass velocity instantly.

The same distinction resolves the persistent question of why a round trip into and out of a wind takes longer than the same trip in still air, even though the wind helps as much on one leg as it hinders on the other. The help and the hindrance are equal in *speed*, but the aircraft spends longer on the slow leg than on the fast one, so the slow leg is weighted more heavily in the total time. The wind's effect on speed cancels; its effect on time does not.

Worked example. An aircraft cruises at 300 kt true airspeed and flies 600 NM out and 600 NM back along the same line, in a steady 60 kt wind blowing along that line.

Outbound, into the wind: ground speed $300 - 60 = 240$ kt, so the leg takes $600/240 = 2.50$ hours.

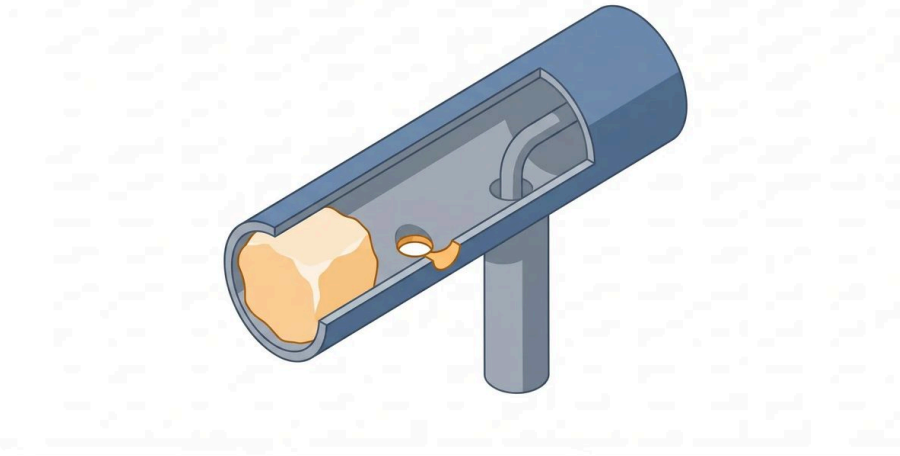
Homebound, with the wind: ground speed $300 + 60 = 360$ kt, so the leg takes $600/360 = 1.67$ hours.

Total 4.17 hours, against $1200/300 = 4.00$ hours in still air — 10 minutes longer despite the wind having been exactly as helpful as it was unhelpful. The stronger the wind relative to the true airspeed, the worse the penalty becomes, and at a wind equal to the true airspeed the aircraft would never complete the outbound leg at all.

What happens when one of the three pressures goes wrong

Because the airspeed indicator displays a *difference*, an error in either of the two pressures feeding it produces an airspeed error — and the direction of that error depends on which line is affected and on what the aircraft is doing at the time. All of the following can be worked out from $q = p_t - p_s$ alone, without memorising a single case, and doing so is far more reliable than trying to recall a table.

Blocked pitot tube, with the drain hole still clear. The ram inlet is obstructed but the pressure trapped behind it can escape through the drain, so the pitot line pressure bleeds down until it equals the static pressure. Then $p_t - p_s = 0$, and the airspeed indicator falls towards **zero** regardless of how fast the aircraft is actually flying. The altimeter and vertical speed indicator, which use only the static line, are unaffected.



A pitot tube with its ram inlet iced over but its drain hole clear, so trapped pressure can escape.

Blocked pitot tube, with the drain hole also blocked. Now the pitot pressure is *trapped* at whatever value it had when the blockage formed. The static pressure continues to change with altitude, so the difference the instrument measures changes with altitude and not with speed:

- **In a climb**, the static pressure falls while the trapped pitot pressure does not, so the difference *grows* and the airspeed indicator **over-reads** — it appears to accelerate as the aircraft climbs.
- **In a descent**, the static pressure rises, the difference shrinks, and the instrument **under-reads**.
- **In level flight**, the reading is frozen and does not respond to genuine speed changes at all.

An airspeed indicator behaving like an altimeter — increasing steadily in a climb at constant power and attitude — is the signature of this failure, and it is a dangerous one, because the natural response to an apparently increasing airspeed in the climb is to raise the nose further.

Blocked static ports. The static pressure is now trapped at its value when the blockage formed, and three instruments are affected at once:

- The **altimeter freezes** at the altitude of the blockage, since it has nothing else to respond to.
- The **vertical speed indicator reads zero**, for the same reason.
- The **airspeed indicator** now errs in the *opposite* sense to the blocked-pitot case. In a climb the true static pressure falls but the trapped value does not, so the instrument computes a difference that is too small and **under-reads**. In a descent it **over-reads**.

Three instruments failing together, with the altimeter stuck, points at the static system; a single airspeed anomaly with a healthy altimeter points at the pitot system. That distinction is the first diagnostic step in any air-data complaint.

Alternate static sources and position error

Because a blocked static system disables three instruments at once, many aircraft provide an **alternate static source**, usually drawing from inside an unpressurised part of the airframe. The alternate source is a compromise rather than an equivalent, and the reason lies in aerodynamics: air flowing around the fuselage accelerates, and where flow accelerates its static pressure falls. The pressure inside the cabin or an equipment bay is therefore normally a little *lower* than true ambient. A lower sensed static pressure makes the altimeter believe it is higher and makes the airspeed indicator compute a larger dynamic pressure, so with the alternate source selected both instruments typically read slightly **high**. The magnitude is installation-specific, which is why an aircraft flight manual carries a correction table for alternate static operation rather than a single figure.

The same physics is the origin of **position error**, which is present even when the system is perfectly serviceable. There is no point on an aircraft where the local static pressure is exactly the free-stream value at every speed and every angle of attack; the designer chooses the least-bad location and the residual difference is calibrated out. Position error is therefore largest where the flow field changes most: at high angles of attack, at low speed, and with flaps and landing gear extended. This is precisely the correction that turns indicated airspeed into calibrated airspeed, and it explains why the correction is negligible in the cruise and significant on the approach.

Water, ice and the practical care of the system

Two atmospheric properties dealt with earlier in this note come back here as maintenance issues:

- **Water.** Humid air drawn into pitot and static lines condenses when the aircraft climbs into colder air, and liquid water trapped in a line will move under acceleration and alter the pressure transmitted. Systems are therefore routed with a continuous downward slope to drain points, and those drains are opened at the intervals the maintenance schedule specifies. A slug of water in a static line is a classic cause of intermittent, attitude-dependent altimeter and airspeed errors that cannot be reproduced on the ground.
- **Ice.** A pitot tube is a forward-facing opening in air that frequently contains supercooled water, so it is heated as a matter of course, and static ports in icing-prone locations are heated too. Loss of pitot heat in visible moisture at low temperature produces exactly the blocked-pitot symptoms described above, and does so within minutes.

Why a pitot-static leak check is a pressure check, not a flight check. The instruments are calibrated against the standard atmosphere, so the only meaningful test is to apply known pressures with a pitot-static test set and compare the indications against the values the standard atmosphere assigns to those pressures. Comparing an altimeter against a known aerodrome elevation on a non-standard day tests the weather as much as the instrument; comparing it against an applied pressure tests only the instrument. And because the system is checked for *leaks* as well as accuracy, the applied pressure must hold within a specified tolerance over a specified time — a slow leak in a static line behaves exactly like a partial blockage in flight.

4. Engine Performance

Both jet engines and piston engines depend on the **mass of air** they can ingest per unit time. This mass flow rate depends directly on air density. Since density decreases with altitude (and with increasing temperature), engine performance degrades at altitude and in hot conditions:

- **Jet engines** — thrust decreases approximately in proportion to the decrease in density. A turbofan engine producing 100 kN at sea level ISA might produce only about 30 kN at 35,000 ft.
- **Piston engines** — power output decreases with altitude unless the engine is supercharged or turbocharged. A naturally aspirated piston engine loses approximately 3–4 % of its power for every 1,000 ft of altitude gain.
- All engine ratings in the Type Certificate Data Sheet (TCDS) are referenced to ISA conditions. When testing engines during maintenance, the measured performance must be corrected to ISA conditions for comparison with the published limits.

Mass flow is the quantity that matters

The reason both engine families behave the same way is that both are governed by the mass of air passing through them per second:

$$\dot{m} = \rho A V$$

Where $\dot{m}$ = mass flow (kg/s), ρ = air density (kg/m³), A = intake area (m²), V = velocity of the air entering (m/s).

The intake area is fixed by the hardware and the velocity is fixed by the flight condition and the compressor, so at a given flight condition the mass flow follows the density directly. A turbine engine at altitude is still swallowing the same *volume* of air per second as it did at sea level; what it is short of is the mass inside that volume. Since thrust is produced by accelerating a mass of air rearwards, less mass means less thrust.

Two effects the simple density rule leaves out

Thrust falling in proportion to density is a good first approximation and the right one for an examination, but two refinements explain what is actually seen on the flight deck and on a test cell:

- **Ram effect works in the engine's favour.** At speed, the air is partially compressed simply by being brought towards rest in the intake, so the pressure and density at the compressor face are higher than ambient. The faster the aircraft, the greater this recovery, so at cruise Mach numbers the thrust does not

fall quite as far as the ambient density alone would suggest. At low forward speed there is almost no ram recovery, which is why the static thrust penalty on a hot, high runway is the harshest case an engine sees.

- **Above the tropopause, thrust falls away faster.** Below the tropopause the density decline is softened by the falling temperature: density is pressure divided by temperature, and a falling denominator partly offsets a falling numerator. Above the tropopause the temperature stops falling, so that offset disappears and the density falls in strict proportion to the pressure. The density loss per kilometre is measurably steeper above 11 km than below it, and that is one of the reasons a subsonic transport's practical ceiling sits close to the tropopause rather than far above it.

Why the piston engine loses more than density alone predicts

Scaling by density alone would suggest a naturally aspirated piston engine losing about 2.5 to 3 % of its power for each of the first few thousand feet. The figures quoted in practice are larger, and the reason is that only the *power produced* scales with density, while a substantial part of the *power consumed inside the engine* does not. Friction in the bearings and rings, the work of driving the camshaft and accessories, and the pumping losses of drawing air past a throttle are all set by the engine's speed and geometry, not by the air density. As the gross power falls with altitude, those fixed internal losses become a larger fraction of it, so the useful output at the propeller shaft falls faster than the density does.

Supercharging and turbocharging attack the problem at its root by restoring the induction pressure to a sea-level value. The compressor is driven either mechanically from the crankshaft or by a turbine in the exhaust, and it raises the manifold pressure back to its sea-level figure. This works up to a **critical altitude** — the level above which the compressor can no longer maintain the target manifold pressure, either because it has reached its own speed limit or because the wastegate is fully closed. Above the critical altitude a boosted engine loses power in the same way an unboosted one does, just from a much higher starting altitude.

Correcting measured engine parameters to standard conditions

The reason the pressure and temperature ratios introduced earlier are worth knowing is that they are what engine performance is actually reduced against. A measured engine parameter recorded on the day cannot be compared with a published limit unless both refer to the same atmosphere, so measurements are divided by the appropriate ratio to give a **corrected** (or referred) parameter that is independent of the conditions:

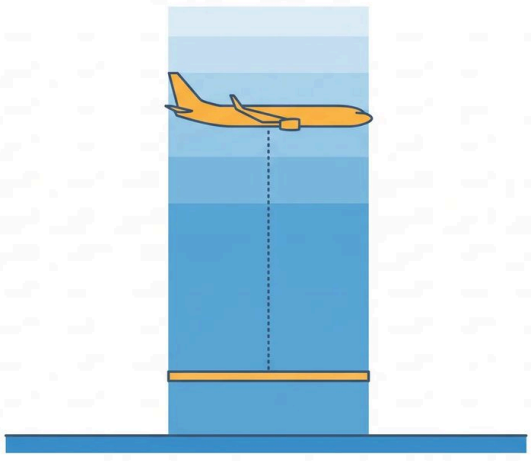
Measured quantity	Corrected form	Why the ratio takes that form
Rotor speed, N	$N/\sqrt{\theta}$	Aerodynamic behaviour of a compressor blade depends on the ratio of blade speed to the local speed of sound, and the speed of sound goes with the square root of absolute temperature
Thrust or power	F/δ	Thrust is generated by pressure forces, which scale with the pressure level at the engine face
Air mass flow	$\dot{m}\sqrt{\theta}/\delta$	Combines the density term through δ with the velocity term through $\sqrt{\theta}$
Turbine gas temperature	T/θ	Temperature rise through the engine is proportional to the inlet temperature

This is the mechanism behind the practical instruction to correct ground-run figures before judging them. Two runs of the same healthy engine on two different days will give different raw readings and identical corrected ones; a genuine deterioration shows up as a change in the corrected value. Comparing raw numbers between a cold morning and a hot afternoon compares the weather, not the engine.

5. Pressurisation and the Physiological Atmosphere

The fifth application is the one that keeps the occupants alive, and it depends on the same pressure profile as everything above. As established in the section on composition, the proportion of oxygen in the air is constant with altitude and it is the *partial pressure* of that oxygen that falls. A pressurised cabin exists to hold the total pressure — and therefore the oxygen partial pressure — at a level the body can work with, while the aircraft flies where the air is thin enough to be efficient.

Cabin altitude is the standard way of expressing that. It is not a physical dimension of the cabin, and it is not the aircraft's altitude. It is the *altitude in the standard atmosphere at which the pressure now inside the cabin would be found* — the same reverse table look-up that defines pressure altitude, applied to the pressure on the inside of the fuselage instead of the outside. An aircraft cruising at 39,000 ft with a cabin pressure equal to the standard pressure at 7,000 ft is said to have a cabin altitude of 7,000 ft.



Cabin altitude is the lower level in the standard atmosphere where the air matches the cabin's pressure.

Quantity	What it refers to	Typical cruise value
Aircraft (pressure) altitude	The ambient pressure outside	The cruising flight level
Cabin altitude	The pressure inside the cabin	Well below the flight level, and limited by certification
Cabin differential pressure	The *difference* between inside and outside	The structural design case for the fuselage

The certification requirement for large aeroplanes limits the maximum cabin altitude to **8,000 ft** at the aircraft's maximum operating altitude in normal operation. That figure is a compromise between two competing

demands, and both directions must be understood:

- **A lower cabin altitude is better for the occupants** — more oxygen partial pressure, less fatigue, less gas expansion in the body's cavities. At 8,000 ft the oxygen partial pressure is about three quarters of its sea-level value, which is tolerable for healthy occupants without supplemental oxygen.
- **A lower cabin altitude is worse for the structure** — it means a larger differential pressure across the fuselage, which is a pressure vessel loaded in tension every flight. Differential pressure sets the skin thickness, the frame and stringer sizing and, because the loading cycles once per flight, the fatigue life of the whole cabin. Maintaining sea-level pressure at cruise altitude would demand a differential the airframe could not carry at acceptable weight.

The consequence of losing that protection is **hypoxia** — insufficient oxygen reaching the tissues. Because it is the partial pressure that falls, and because it does so smoothly, hypoxia produces impaired judgement before it produces any obvious physical distress, which is what makes it dangerous. The time available to act after a rapid decompression shortens sharply with altitude: it is measured in minutes in the mid twenties of thousands of feet, and in seconds above about 40,000 ft. That is the reasoning behind the requirement for immediate crew oxygen and an expeditious descent.

Two altitudes worth remembering as thresholds. At **18,000 ft** the total pressure has fallen to half its sea-level value, so each breath delivers half the oxygen molecules; supplemental oxygen is essential for sustained operation anywhere near it. At about **63,000 ft** — the Armstrong limit — the ambient pressure has fallen to roughly 62 hPa, which is the vapour pressure of water at body temperature. Above that altitude, body fluids would boil at body temperature, and no quantity of supplemental oxygen is of any use: only a sealed pressure suit or a pressurised cabin will do.

Other physiological effects of the atmosphere

Hypoxia, dealt with above, is the most immediate consequence of the pressure profile for the occupants, but it is not the only one. Three further effects follow directly from the properties this note has established, and all three are Boyle's law or the humidity relationships acting on the human body:

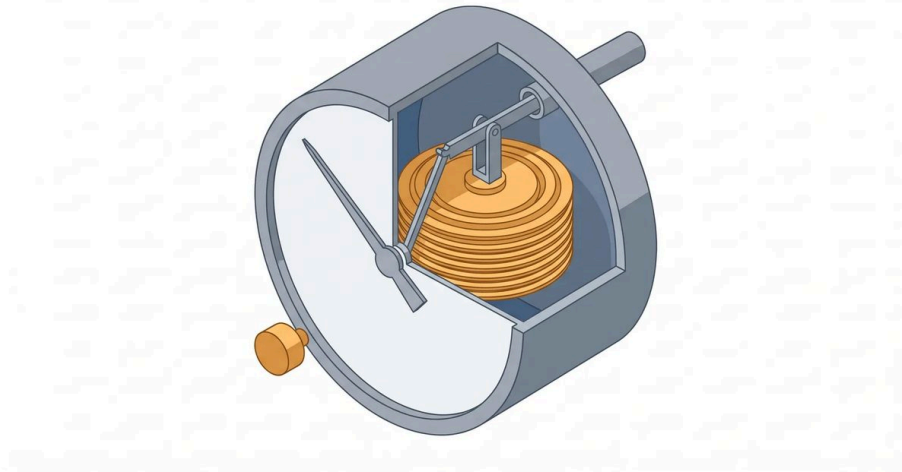
- **Trapped gas expansion.** Gas in the middle ear, the sinuses, the teeth and the digestive tract is at approximately ambient pressure on the ground. As the cabin altitude increases, the ambient pressure falls and that trapped gas expands in inverse proportion — Boyle's law exactly as set out earlier. The discomfort is worst where the gas cannot escape freely, which is why a blocked eustachian tube during a descent, when the pressure is *rising* and the ear must admit gas rather than vent it, is far more painful than the same blockage during a climb.
- **Decompression sickness.** Nitrogen dissolves in body tissues in proportion to its partial pressure, following the same law that governs any gas dissolving in a liquid. A rapid reduction in ambient pressure reduces that partial pressure and the dissolved nitrogen can come out of solution as bubbles — the same phenomenon as opening a carbonated drink. The risk becomes significant at cabin altitudes above roughly 18,000 ft and rises sharply with altitude, and it is much increased by recent diving, because the diver has already been breathing nitrogen at greater than atmospheric pressure and starts the flight with an elevated tissue load.

- **Dehydration.** The air at cruising altitude holds almost no water, for the saturation reason established earlier: at –55 °C the saturation vapour pressure is a tiny fraction of its sea-level value. Cabin air drawn from that source and warmed to a comfortable temperature has a very low relative humidity, because warming raises the amount the air *could* hold while adding nothing to what it does hold. The consequences — dry eyes, dry airways, dehydration over a long sector — are a direct and unavoidable consequence of where the air comes from.

One further effect is not a gas-law consequence at all but is nonetheless an altitude effect. **Cosmic radiation** from the sun and from beyond the solar system is attenuated by the mass of atmosphere above, and that mass, as established earlier, is measured by the pressure. The dose rate therefore rises steadily with altitude — at typical cruising levels it is greater than at the surface by an order of magnitude or more — and it also rises with latitude, because the earth's magnetic field deflects charged particles towards the poles and offers least shielding there. The combination means that polar routes at high flight levels carry the highest exposure, which is why crew radiation dose is assessed and monitored as an occupational matter, and why solar activity is relevant to route planning at the highest latitudes.

Altimetry: Pressure Settings and Altitude Datums

An altimeter measures pressure and displays a height, and a height is meaningless without a datum. Everything in this section follows from that single observation: there are several useful datums, the pilot selects between them on the subscale, and the word "altitude" means a different physical quantity depending on which one is selected. This is where a large share of examination questions in this paragraph live, and almost all of them are answered by identifying the datum.



Inside the altimeter, a capsule senses only pressure; the subscale knob sets which pressure reads as zero.

The subscale: what turning it actually does

The altimeter's aneroid capsule responds to the ambient static pressure and nothing else. Turning the subscale knob does not change what the capsule senses; it mechanically shifts the relationship between the capsule's deflection and the pointer, which is equivalent to telling the instrument **which pressure to display**

as zero feet. Everything above that pressure is then indicated using the standard atmosphere's pressure-height relationship.

Three settings are standardised internationally under the ICAO **Q-code** system, and there are exactly three of them:

Code	Pressure set on the subscale	What the altimeter then indicates	Reading when on the ground at the aerodrome	Where it is used
QNH	The local pressure reduced to **mean sea level**	**Altitude** — height above mean sea level	The aerodrome **elevation**	Below the transition altitude, for terrain and obstacle clearance
QFE	The actual pressure at the **aerodrome reference point**	**Height** — height above that aerodrome	**Zero**	Circuit flying and some approaches, chiefly at military and training aerodromes
QNE	The **standard** pressure, 1013.25 hPa (29.92 in Hg)	**Pressure altitude**, expressed as a flight level	Whatever pressure altitude the field happens to have that day — rarely a round number and rarely the elevation	At and above the transition altitude, for vertical separation between aircraft

Three settings, three readings, and a one-line test for each. Ask what the instrument would read sitting on the runway. **QNH gives the aerodrome elevation**, because the datum is the sea. **QFE gives zero**, because the datum is the aerodrome itself. **QNE gives neither**, because the datum is a fixed pressure that has nothing to do with this aerodrome. Any option that describes an altimeter reading zero on the ground is describing QFE; any option that describes it reading the field elevation is describing QNH; any option quoting 1013.25 hPa is describing QNE. The three are mutually exclusive and no other Q-code is used for altimetry.

Altitude, height and flight level — three words that are not synonyms

The three settings give rise to three technical terms which aviation uses with precision even though everyday English treats them as interchangeable:

- **Altitude** is measured from **mean sea level**. It is what the altimeter indicates on QNH, and it is the quantity charts use for terrain, obstacle and minimum safe altitudes.
- **Height** is measured from a specified **datum on the surface**, normally the aerodrome. It is what the altimeter indicates on QFE.
- **Flight level** is a **pressure altitude** flown with 1013.25 hPa set, expressed in hundreds of feet with the leading zeros retained. FL350 means a pressure altitude of 35,000 ft, not necessarily a true altitude of 35,000 ft.

The critical property of a flight level is that **every aircraft using it is referred to the same fixed datum**. Two aircraft assigned FL340 and FL350 are separated by 1,000 ft of pressure difference regardless of what the local sea-level pressure happens to be, because both altimeters are ignoring the local pressure entirely. If both

were instead flying on local QNH settings obtained from different sources, a difference between those settings would translate directly into a reduction of the separation between them. **Flight levels do not guarantee true altitude; they guarantee that everyone is wrong by the same amount**, which is exactly what vertical separation requires.

Which is why a flight level is not a height above the ground. Over high terrain a flight level may be far closer to the surface than its number suggests, and in cold air it is lower still — the temperature error described earlier applies to a flight level exactly as it does to any other pressure-derived altitude. The transition altitude exists precisely so that flight levels are used only where terrain clearance has stopped being the governing concern.

Transition altitude, transition level and the layer between them

Since terrain clearance needs a sea-level datum and traffic separation needs a common pressure datum, the airspace is divided at a defined level and the two systems are used one above the other:

Term	Definition	Altimeter setting	Changed when
Transition altitude	The altitude at or below which vertical position is controlled by reference to **altitudes** (QNH)	QNH below it	Climbing through it, the setting is changed to 1013.25
Transition level	The lowest **flight level** available for use above the transition altitude	1013.25 at and above it	Descending through it, the setting is changed to QNH
Transition layer	The airspace between the transition altitude and the transition level	—	Not used for level cruising flight
Altimeter setting region	The airspace **below** the transition altitude, where altimeters are set to local QNH	QNH	—

The transition altitude is **not a single worldwide number**. It is published for each state or each aerodrome and varies widely — from a few thousand feet in parts of Europe, where terrain is low and airspace is congested, up to considerably higher figures elsewhere. A commonly quoted figure of 18,000 ft belongs to one particular national system and is offered as a distractor precisely because it is so often assumed to be universal.

The transition level, by contrast, is not fixed even at a single aerodrome: it moves with the pressure of the day, because it is the lowest flight level that provides the required buffer above the transition altitude, and a low QNH pushes the flight levels physically downward.

Pressure altitude, and how to work it out without an altimeter

Pressure altitude is the altitude indicated with 1013.25 hPa set on the subscale — equivalently, the altitude in the standard atmosphere at which the ambient pressure equals the pressure now being sensed. It is the input to every performance chart, because performance depends on density and density is computed from pressure altitude and temperature.

On the ground the pressure altitude can be found without touching the altimeter, using the sea-level rate of pressure change with height, which is derived later in this note:

$$\text{Pressure altitude} = \text{Field elevation} + [(1013 - \text{QNH in hPa}) \times 27]$$

The 27 is the number of feet per hectopascal near sea level. Many operators use 30 instead, which is the rounder rule of thumb; the difference is small at ordinary pressure deviations.

Worked example — low pressure. An aerodrome is at 2,000 ft elevation and the QNH is 993 hPa.

$1013 - 993 = 20$ hPa below standard, so

$2000 + (20 \times 27) = 2000 + 540 = 2{,}540$ ft pressure altitude.

Direction check: the pressure is *lower* than standard, which means the standard atmosphere would have to be climbed *higher* before reaching this pressure. The pressure altitude must therefore be greater than the elevation, and 2,540 ft duly is.

Worked example — high pressure. The same aerodrome with the QNH at 1030 hPa.

$1013 - 1030 = -17$ hPa, so

$2000 + (-17 \times 27) = 2000 - 459 = 1{,}541$ ft pressure altitude.

On a high-pressure day the aircraft performs as though the aerodrome were 460 ft *lower* than it is — a small bonus, and the reverse of the penalty in the previous example. Note that the elevation of the aerodrome has not changed by a single foot in either case; what has changed is the pressure environment the aircraft and its engines are operating in.

The five altitudes an engineer must be able to distinguish

Aviation uses the word "altitude" for five different quantities. They are defined against four different datums, and two of the five are computed rather than read from an instrument; the exam tests whether they can be separated:

Altitude	Definition	Datum	How it is obtained	When it equals the others
Indicated altitude	The value read directly off the altimeter with the local QNH set	Mean sea level, via the local pressure	Read off the instrument	Equals true altitude only when the temperature profile matches the ISA
Pressure altitude	The altitude indicated with 1013.25 hPa set	The standard pressure datum	Read off the instrument, or computed from elevation and QNH	Equals indicated altitude only when the QNH happens to be 1013.25

Altitude	Definition	Datum	How it is obtained	When it equals the others
Density altitude	Pressure altitude corrected for non-standard temperature: the ISA altitude having the same air density as the present conditions	The standard atmosphere's density profile	Computed — never displayed	Equals pressure altitude only when the temperature is exactly standard
True altitude	The actual geometric height above mean sea level	Mean sea level	Derived from indicated altitude with a temperature correction	Equals indicated altitude only in ISA conditions
Absolute altitude	The height above the **terrain directly below**	The ground, wherever it happens to be	Measured by radio altimeter	Equals true altitude only over the sea

How to tell them apart from a question stem. If the description mentions the local altimeter setting, it is **indicated** altitude. If it mentions 1013.25 hPa or 29.92 in Hg, it is **pressure** altitude. If it mentions a correction for temperature and refers to air density or performance, it is **density** altitude. If it says "actual height above sea level", it is **true** altitude. If it says "above the ground" or "varies with the terrain", it is **absolute** altitude. The word "absolute" here means "above the surface", which is the opposite of the intuition that would make it the most fundamental of the five.

Worked example — all five at once. An aircraft is over terrain of 3,000 ft elevation. Its altimeter, set to the local QNH of 1003 hPa, indicates 10,000 ft. The outside air temperature is –15 °C.

- **Indicated altitude:** 10,000 ft — the reading.
- **Pressure altitude:** $10000 + [(1013 - 1003) \times 27] = 10{,}270$ ft.
- **Density altitude:** standard temperature at 10,270 ft is about $15 - (2 \times 10.27) = -5.5\ ^{\circ}\mathrm{C}$; the actual is –15 °C, so the deviation is about –9.5 °C and the density altitude is roughly $10270 - (9.5 \times 120) = 9{,}130$ ft. The air is *colder* than standard, so it is denser, so the density altitude is *below* the pressure altitude.
- **True altitude:** the air is colder than standard, so the altimeter over-reads and the aircraft is lower than 10,000 ft. Applying the 4 % per 10 °C rule to the 10,000 ft above the QNH datum gives roughly 380 ft of over-reading, so the true altitude is about 9,600 ft.
- **Absolute altitude:** true altitude minus terrain elevation, about $9600 - 3000 = 6{,}600$ ft above the ground.

Five different numbers, all correct, all describing the same aircraft at the same instant. Which one is wanted depends entirely on the question being asked: separation wants the pressure altitude, terrain clearance wants the true altitude, performance wants the density altitude, and the radio altimeter is the only instrument reading the last of them.

Altimetry accuracy and reduced vertical separation

Everything above establishes that a flight level is a pressure surface rather than a height, and that every aircraft on the same flight level is therefore wrong by the same amount. That argument has a limit: it holds only

to the extent that every aircraft's altimetry system converts pressure to indicated altitude in the *same* way. Where two aircraft's systems disagree, the disagreement subtracts directly from the vertical separation between them.

This is why the vertical separation minimum above the higher flight levels was historically 2,000 ft rather than 1,000 ft: the errors grow with altitude, for the reason given in the section on pressure variation — at 40,000 ft a hectopascal is worth about 111 ft, against 27 ft at sea level, so a given pressure-measurement error produces four times the height error. Reducing that minimum to 1,000 ft, in the band from about FL290 to FL410, required the altimetry system itself to be qualified rather than merely serviceable. The consequences for maintenance are specific and are the reason the subject appears in a Part-66 syllabus at all:



Skin near a static port is controlled: dents, repairs, roughness or paint build-up alter the sensed pressure.

- **The whole altimetry system is qualified, not just the instruments.** The static ports, their exact location and surface finish, the pipework, the air data computers and the altitude reporting to the transponder are all part of the approved configuration.
- **Skin condition near the static ports is controlled.** A repair, a dent, a mismatched doubler, a patch of surface roughness or even paint build-up in the vicinity of a static port alters the local flow and therefore the sensed static pressure. Work in those zones is subject to specific surface-condition limits, and it is one of the few places where paint thickness is an airworthiness matter.
- **Periodic verification is required**, comparing what the system reports against a known applied pressure, because a slow drift produces no symptom the crew would notice — the aircraft simply flies slightly off its assigned level.

Why the errors that matter here are not the ones discussed earlier. Temperature error, non-standard pressure and the rest affect every aircraft on a flight level identically, so they do not erode separation between two aircraft. What erodes separation is *differential* error — one aircraft's system reading differently from another's for the same ambient pressure. That is entirely a matter of the individual installation, which is why the requirement is on the aircraft rather than on the atmosphere.

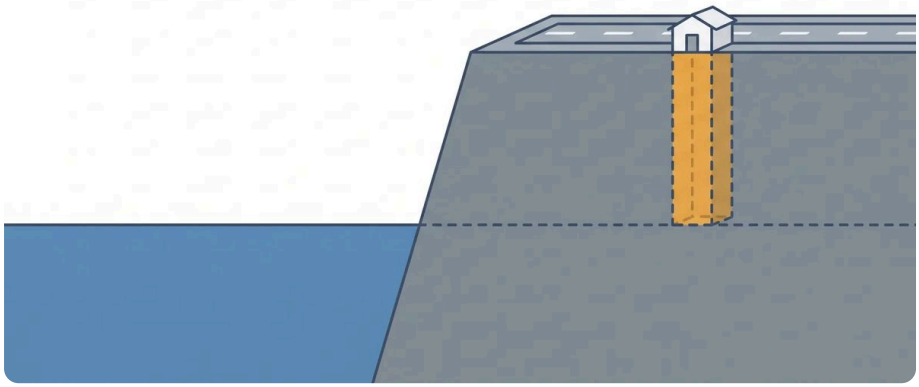
Measuring and Reporting the Atmosphere

Every correction described so far — ISA deviation, pressure altitude, density altitude, altimeter setting — requires a measurement of the real atmosphere to work from. Those measurements come from a well-defined network, and knowing where each number originates explains several things about its accuracy and its area of validity.

Surface observations and the derivation of QNH

An aerodrome measures the pressure at its own barometer, which is at the aerodrome's own elevation. That reading, uncorrected, is the aerodrome pressure — the quantity set on the subscale as QFE. QNH cannot be measured directly at all, because it is the pressure that *would* exist at sea level under that aerodrome, and at any aerodrome above sea level there is no such place to put an instrument.

QNH is therefore **computed** from the measured aerodrome pressure by adding the pressure of a notional column of air extending from the aerodrome down to sea level. Two consequences follow directly:



QNH adds a notional air column from the aerodrome down to sea level, where no barometer can stand.

- **QNH is a derived quantity and depends on an assumption about the notional column.** The standard practice is to assume standard-atmosphere temperatures within it, which is why the difference between QNH and QFE at a given aerodrome is very close to a fixed number of hectopascals and changes only slightly with the weather.
- **QNH has an area of validity.** Because pressure varies horizontally, the QNH computed at one aerodrome is not valid a hundred miles away. This is the reason for regional pressure settings covering defined areas, and for the requirement to obtain a new setting when entering a new region.

Aerodrome observations are encoded and disseminated in routine reports and forecasts, and passed to arriving and departing aircraft through the automatic terminal information service. The quantities in them are exactly the ones this note has dealt with: wind, visibility, cloud, temperature, dew point and QNH. Reading a report as a set of atmospheric measurements rather than as a code makes its content predictable — the

temperature and dew point together give the humidity and the likelihood of fog, and the QNH gives the pressure altitude of the runway.

Measuring the vertical profile

The environmental lapse rate, the tropopause height, the position of inversions and the strength of the jet stream are all properties of the *profile*, and none can be obtained from a surface observation. They come from two sources:

- **Radiosondes.** A small instrument package carried aloft by a balloon, transmitting temperature, humidity and pressure continuously as it ascends, with the wind derived from its tracked drift. Launches are made at synchronised times from a worldwide network, and the resulting soundings are what the environmental lapse rate is actually measured from. A radiosonde ascent typically continues into the stratosphere until the balloon bursts, so it profiles the whole of the aviation range and beyond.
- **Aircraft themselves.** Modern transports report air data automatically — pressure altitude, static air temperature and wind derived from the difference between air data and inertial or satellite navigation — producing a dense set of profiles along every busy route. This is why forecast accuracy at cruising levels is far better on well-travelled routes than over empty ocean.

Why the measured profile matters to an engineer as well as a forecaster. The temperature reported at a flight level determines the ISA deviation, and the ISA deviation determines whether a set of cruise engine parameters is within limits. An engine appearing to run hot at a particular flight level on a particular day may be doing nothing more than reporting an ISA + 15 atmosphere honestly. Correcting to standard conditions, as described earlier in this note, is what turns the observation into evidence.

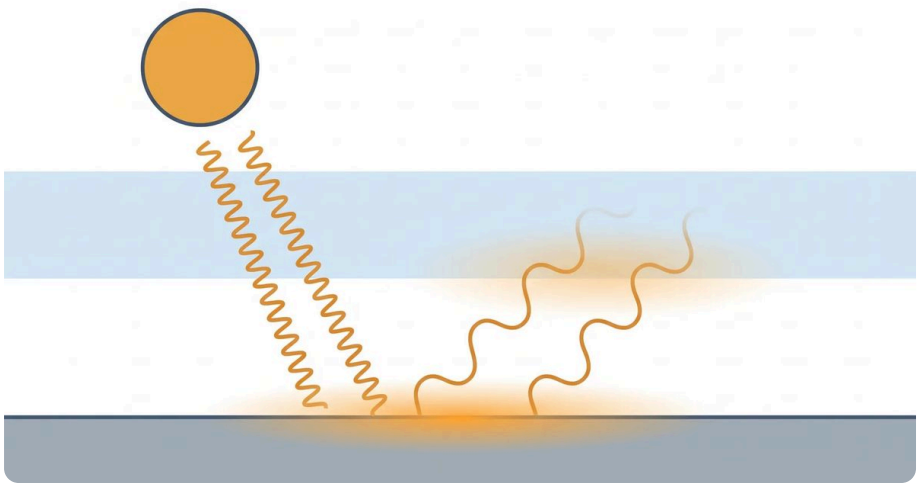
Temperature Variation with Altitude — Detailed

The decrease in temperature with altitude in the troposphere occurs because the primary heat source for the lower atmosphere is the **Earth's surface**, not direct solar radiation. The sun's energy passes through the atmosphere relatively freely and warms the ground. The ground then re-radiates this energy as infrared (heat) radiation and also heats the air in direct contact with it through **conduction**. This warm surface air then rises by **convection**, carrying heat upward. However, as air rises, it **expands** (because atmospheric pressure decreases with altitude), and this expansion causes the air to **cool** (adiabatic cooling). The net result is that the further from the ground, the cooler the air — this is the temperature lapse rate.

The **actual (environmental) lapse rate** varies with location and conditions — it can be greater than, equal to, or less than the ISA standard of 6.5 °C/km. In some situations, a **temperature inversion** occurs where temperature actually *increases* with altitude over a limited height band. Inversions are significant because they suppress vertical air movement, trapping pollutants and moisture below them, and can cause wind shear at their boundaries.

The radiation balance behind the lapse rate

The reason the surface is the heat source, rather than the air itself, comes down to the wavelength of the radiation involved. Solar radiation arrives mostly as short-wavelength visible and near-infrared energy, and the atmosphere is largely *transparent* to it — nitrogen, oxygen and argon absorb almost none of it, so most of it reaches the ground. The ground warms, and being at a few hundred kelvin rather than a few thousand it re-emits that energy as long-wavelength infrared. The atmosphere is *not* transparent to those wavelengths, because water vapour and carbon dioxide absorb them strongly. The air is therefore heated from below and from within by the surface's outgoing radiation, and hardly at all by the incoming sunlight passing through it.



Sunlight passes through the air to warm the ground; the ground's infrared is absorbed by the air.

All three modes of heat transfer are involved, but on very different scales:

- **Conduction** matters only in the first few millimetres. Air is a poor conductor — that is exactly why it is used as an insulator in double glazing and lagging — so conduction transfers heat from the ground into the air layer touching it and no further.
- **Convection** carries the heat upward. Air warmed at the surface becomes less dense than its surroundings and rises, and this bulk movement of warm air is what distributes surface heat through the depth of the troposphere.
- **Radiation** carries the energy in and out of the system as a whole, and is the only mechanism by which the earth can lose heat to space.

Why the actual lapse rate is never quite the standard one

The standard 6.5 °C/km is a long-term global average, and the atmosphere at any moment departs from it in predictable ways:

- **Time of day.** Strong surface heating on a sunny afternoon steepens the lapse rate near the ground, sometimes far beyond the standard value. Overnight radiative cooling of the surface does the opposite and frequently reverses it entirely near the ground.
- **Surface type.** Dry land heats and cools quickly; water changes temperature slowly. The lapse rate over a runway on a summer afternoon and over the sea a few miles away can be quite different, which is one

origin of coastal turbulence and sea breezes.

- **Cloud and moisture.** Cloud cover reduces both incoming solar radiation by day and outgoing infrared by night, damping the daily swing at the surface and flattening the lapse rate.
- **Air-mass history.** Air that has travelled over a warm surface is heated from below and its lapse rate steepens; air moving over a colder surface is cooled from below and its lapse rate flattens or reverses.

This is why the environmental lapse rate has to be measured rather than assumed, and why the difference between the measured value and the adiabatic rates — not its absolute size — is what determines the weather. That comparison is the subject of the next section.

Measuring air temperature from a moving aircraft

One practical complication belongs here, because it affects every temperature the flight deck and the maintenance engineer actually see. A temperature probe on an aircraft moving at several hundred knots does not measure the temperature of the undisturbed air. The air brought to rest against the probe has had its kinetic energy converted into heat, exactly as the pitot tube converts it into pressure, so the probe reads *warmer* than the air it is flying through. Two quantities therefore have to be distinguished:

- **Static air temperature (SAT)**, also called outside air temperature — the true temperature of the undisturbed air. This is the quantity that appears in the ISA tables, that determines the speed of sound and that is used for every performance and density calculation.
- **Total air temperature (TAT)**, also called ram air temperature — the temperature the probe actually senses, which is the static temperature plus the **ram rise** caused by bringing the air to rest.

For air brought fully to rest, $\frac{T_{total}}{T_{static}} = 1 + 0.2\,M^2$

Where M is the Mach number, and both temperatures are absolute. Real probes recover slightly less than the full rise, so an installation-specific recovery factor slightly below 1 is applied.

Worked example — how large is the ram rise? An aircraft cruises at Mach 0.80 where the static air temperature is –50 °C.

Convert to absolute: $-50 + 273.15 = 223.15$ K.

$T_{total} = 223.15 \times (1 + 0.2 \times 0.80^2) = 223.15 \times (1 + 0.128) = 223.15 \times 1.128 = 251.7$ K

which is $251.7 - 273.15 = -21.4\,^{\circ}\mathrm{C}$. The ram rise is nearly **29 °C** — far too large to ignore.

Direction check: stopping moving air converts kinetic energy into thermal energy, so the probe must read *warmer* than the free air, never cooler, and the rise must grow with speed. Both hold: at Mach 0.4 the same air would give a rise of about 7 °C, roughly a quarter of the Mach 0.8 value, as the M^2 term requires.

The air data computer performs this correction continuously, computing the static air temperature from the measured total temperature and the Mach number, which it in turn derives from the pitot and static pressures. Two consequences follow for maintenance. The displayed outside air temperature at cruise is a *computed* value, not a measured one, so an error in the pitot-static system can appear as a temperature indication fault. And on the ground, with the aircraft stationary, the Mach number is zero and the ram rise vanishes — which is

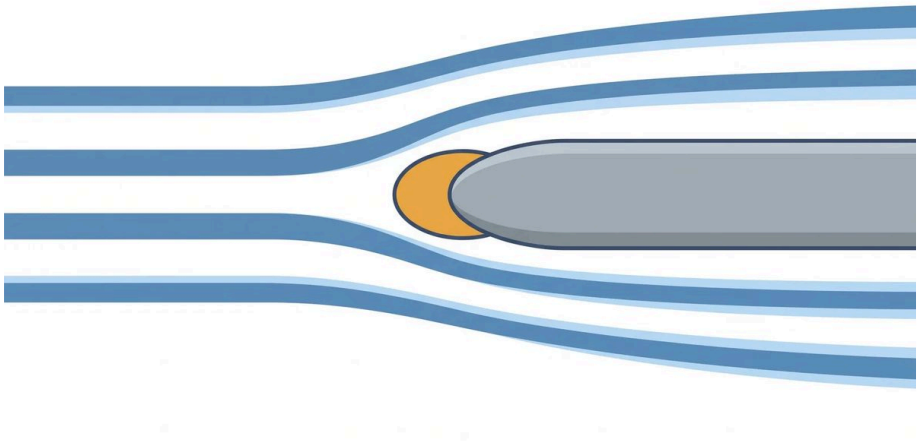
why a static air temperature indication can be checked directly against an ambient thermometer on the ramp but not in flight.

Measuring atmospheric temperature

Given how much of this note depends on knowing the air temperature, it is worth knowing how that measurement is actually made and where it can mislead. Three sensing principles are met on aircraft, and each has a characteristic behaviour:

Sensor	Principle	Characteristics	Typical application
Resistance bulb (platinum or nickel element)	Electrical resistance of a metal rises predictably with temperature	Accurate, stable and repeatable over a moderate range; needs an excitation supply and is sensitive to lead resistance	Outside air temperature and total air temperature probes; oil and fuel temperature
Thermocouple	A junction of two dissimilar metals generates a small voltage proportional to the temperature difference between it and a reference junction	Withstands very high temperatures, needs no excitation, but produces a small signal and requires reference-junction compensation	Exhaust and turbine gas temperature, cylinder head temperature — not usually ambient air
Bimetallic element	Two bonded metals of different expansion coefficients bend as the temperature changes	Simple, self-contained, requires no power; modest accuracy	Direct-reading ambient thermometers and simple indicators

The measurement problem in flight is not the sensor but the airflow, and it has already been dealt with above: a probe in a moving airstream reads the total rather than the static temperature, because the air it senses has been brought partly or wholly to rest and heated in the process. A total air temperature probe is therefore designed around that fact rather than against it. It is shaped to bring the air to rest in a controlled and repeatable way so that the recovery is a known fraction of the full ram rise, and it is heated to prevent icing — with the heating deliberately arranged so that it does not warm the sensing element and corrupt the reading.



Air brought to rest at the probe face is heated, so the probe senses total rather than static temperature.

Two checks that distinguish an instrument fault from physics. On the ground with the aircraft stationary, the ram rise is zero, so the displayed static air temperature should agree closely with an ambient thermometer in the shade — that comparison tests the sensor. In flight, the difference between total and static air temperature should follow the Mach number in the way set out earlier; a total air temperature that fails to rise with speed suggests the probe is not seeing the airflow it should, and a static air temperature that disagrees with the value implied by the Mach number suggests the computation rather than the probe. In both cases the discriminator is that the atmosphere is a poor conspirator: real air temperature varies smoothly and slowly, while a sensor fault does not.

Lapse Rates, Stability and Atmospheric Phenomena

The previous section dealt with how the temperature of the atmosphere varies with height. This one deals with the consequences of that variation, because the vertical temperature profile is what decides whether the air is calm or violent, whether it holds cloud or clears, and where the hazards a cruising aircraft meets are found. Nearly all of it follows from comparing two rates of temperature change that are easily confused and mean quite different things.

Two kinds of lapse rate, and why they must never be mixed up

	Environmental lapse rate (ELR)	Adiabatic lapse rate
What it describes	The temperature of the *surrounding air* at different heights, at one moment	The temperature change of *one parcel of air* as it moves vertically
How it would be measured	By a thermometer carried up through the atmosphere on a balloon	By a thermometer inside a rising parcel, following it
Is it fixed?	No — it varies with place, time of day, season and air mass	Yes — it is set by physics and is very nearly a constant

	Environmental lapse rate (ELR)	Adiabatic lapse rate
Typical value	About 6.5 °C/km on average, but anything from strongly negative to well over 10 °C/km	About 9.8 °C/km dry; roughly half that when saturated

The word **adiabatic** means "without heat exchange with the surroundings". An air parcel is a poor conductor and moves faster than heat can diffuse across its boundary, so to a very good approximation it exchanges no heat with the air around it. Its temperature nevertheless changes as it moves vertically, and the reason is purely mechanical:

- A parcel that **rises** moves into lower pressure, so it **expands**. Expanding means doing work on the surrounding air, and with no heat coming in to replace it, that work is paid for out of the parcel's own internal energy. Its temperature **falls**.
- A parcel that **descends** moves into higher pressure, so it is **compressed**. Work is done on it, and with no heat escaping, its temperature **rises**.

Note the direction carefully: rising air cools and descending air warms, and neither has anything to do with the temperature of the air it is passing through. This is the same physical process that makes the outlet of a compressor hot and the outlet of a turbine or an expansion valve cold, and it is why an air-cycle machine can produce cold air with no refrigerant at all.

The dry and saturated adiabatic rates

Rate	Applies when	Approximate value	Why it has that value
Dry adiabatic lapse rate (DALR)	The parcel is unsaturated — relative humidity below 100 %, no condensation occurring	**About 3 °C per 1,000 ft** (9.8 °C per 1,000 m)	Determined entirely by gravity and the specific heat of air; it is a physical constant, not a measurement
Saturated adiabatic lapse rate (SALR)	The parcel is saturated and condensation is occurring as it rises	**About 1.5 °C per 1,000 ft**, varying with temperature and pressure	Latent heat released by condensation partly offsets the cooling of expansion

The saturated rate is the smaller of the two, and the mechanism is worth stating explicitly because the direction is easy to reverse. A saturated parcel is doing two things at once as it rises: cooling by expansion, exactly as a dry parcel does, and *releasing* latent heat as its water vapour condenses. The released heat warms the parcel, so the net cooling is less than it would be if the parcel were dry. **Condensation therefore makes a rising parcel cool more slowly, not faster.** The saturated rate is also not a fixed number, because how much water condenses per degree of cooling depends on how much vapour the air held to begin with — which is why the saturated rate is close to the dry one in very cold air, where there is almost no water to condense, and much smaller in warm tropical air.

Three lapse rates, three numbers, in the same units. The dry adiabatic rate is about **3 °C per 1,000 ft**; the ISA standard environmental rate is about **2 °C per 1,000 ft**; the saturated adiabatic rate is about **1.5 °C per 1,000 ft**. Note the ordering: the dry adiabatic rate is the steepest of the three and the standard atmosphere sits between the two adiabatic rates. That last fact means the standard atmosphere is *conditionally unstable* — stable against a parcel that stays unsaturated, unstable against one in which condensation has begun — which is the next topic.

Stability: comparing the two rates

Whether a displaced parcel returns to where it started or keeps going is decided entirely by comparing how fast it cools against how fast the air around it gets colder with height. Work the two cases through:

Case 1 — a shallow environmental lapse rate (stable). The environmental lapse rate is 1 °C per 1,000 ft. A dry parcel is displaced 1,000 ft upward.

- The parcel cools at the dry adiabatic rate: 3 °C colder than it was.
- The surrounding air at the new level is only 1 °C colder than at the old level.
- The parcel is therefore **2 °C colder than its surroundings**, hence denser, hence heavier.

It sinks back to where it came from. The atmosphere resists the displacement: it is **stable**.

Case 2 — a steep environmental lapse rate (unstable). The environmental lapse rate is 4 °C per 1,000 ft. The same parcel is displaced 1,000 ft upward.

- The parcel again cools 3 °C.
- The surrounding air at the new level is 4 °C colder.
- The parcel is therefore **1 °C warmer than its surroundings**, hence less dense, hence buoyant.

It keeps rising, and the further it goes the greater the discrepancy becomes. The atmosphere amplifies the displacement: it is **unstable**.

The general rule follows directly, and covers three cases:

Condition	Comparison	Behaviour of displaced air
Absolutely stable	ELR is *less* than the SALR	Returns to its original level whether saturated or not
Conditionally unstable	ELR lies *between* the SALR and the DALR	Stable while unsaturated; becomes unstable once condensation begins
Absolutely unstable	ELR is *greater* than the DALR	Continues rising in either condition

The middle row explains a great deal of real weather. Air that is stable near the surface may become vigorously unstable once it has been lifted far enough to saturate, because from that moment it cools at the smaller saturated rate while the environment continues to cool at its own steeper one. This is why a morning of harmless fair-weather cumulus can turn into an afternoon of thunderstorms with no change in the

environmental profile at all — the trigger is the parcel reaching its condensation level, at which point the latent heat mechanism described earlier takes over as an energy source.

What stability means for the aircraft

	Stable air	Unstable air
Ride	Smooth — vertical motion is suppressed	Turbulent — up- and down-draughts throughout the convective layer
Cloud type	Layered: stratus, stratocumulus, and none at all in very dry stable air	Heaped and vertically developed: cumulus, towering cumulus, cumulonimbus
Precipitation	Steady and prolonged: drizzle, continuous light rain	Showery and intense, often with hail and lightning
Visibility below cloud	**Poor** — smoke, haze and pollutants are trapped by the lack of mixing	**Good** — vertical mixing disperses contaminants through a deep layer
Icing risk	Extensive in area but generally lighter	Localised but potentially severe, with large supercooled droplets in vigorous updraughts

The visibility row is the one that surprises people, because the smooth, quiet, stable day is the one with the worst visibility. Stable air suppresses the vertical mixing that would otherwise dilute surface pollution through the depth of the troposphere, so everything emitted at the surface stays there. This is precisely the mechanism by which an inversion traps pollutants, and it is why industrial smog episodes occur on calm, stable, anticyclonic days rather than windy ones.

Temperature inversions

An **inversion** is a layer in which the temperature *increases* with height instead of decreasing — the reverse of the normal state, which is where the name comes from. Because it puts warm, light air on top of cold, dense air, it is the most stable arrangement the atmosphere can adopt, and it acts as a lid on everything beneath it. Three ways of producing one are commonly examined, and they are distinguished by their *mechanism*, not by their appearance:

Type	How it forms	Where it sits	Characteristic consequences
Radiation inversion	On a clear, calm night the ground radiates its heat away to space and cools rapidly; the air in contact with it is chilled from below	At the surface, typically a few hundred feet deep	Radiation fog, dew and frost; near-calm surface winds with a sharp wind shear at the inversion top
Subsidence inversion	A large mass of air sinks in a high-pressure system and is **compressed and warmed adiabatically** as it descends into higher pressure	Aloft, often a few thousand feet up, and it can be very extensive	Caps convection, so cumulus flattens against it; traps haze into a distinct visible layer with a sharp top

Type	How it forms	Where it sits	Characteristic consequences
Frontal inversion	Warm air, being less dense, is forced to ride up *over* a wedge of colder air at a front	Along the sloping frontal surface, so its height varies across the front	Layered cloud and prolonged precipitation; the freezing rain hazard, where rain falling from the warm layer supercools in the cold air beneath

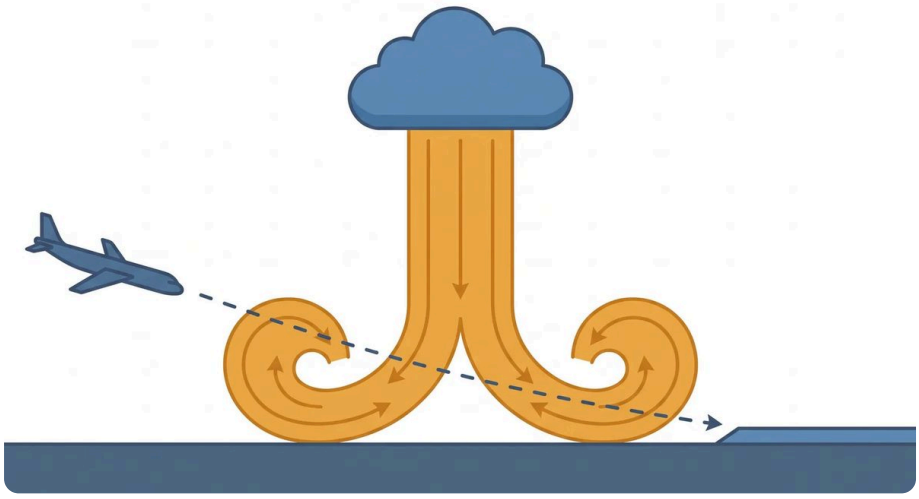
The subsidence case deserves attention because its mechanism is the adiabatic process running in the descending direction, and it is worth confirming the arrow: air sinking into higher pressure is compressed, compression does work on it, and with no heat exchange its temperature rises. Sinking air is **warming** air. A whole layer subsiding therefore ends up warmer than the air beneath it, which has not descended, and an inversion is created aloft with no heating from any external source at all.

An inversion produces wind shear at its boundaries. Because the layer suppresses mixing, the air above it and the air below it are mechanically decoupled and can have quite different velocities. An aircraft climbing or descending through an inversion can therefore meet a rapid change of wind speed and direction over a very small height change — smooth air on both sides with a distinct jolt between them. Nocturnal low-level jets above a radiation inversion are a classic case, and one that is met on exactly the sort of clear, calm night on which no turbulence is expected.

Wind shear

Wind shear is any change of wind velocity — speed, direction, or both — over a short distance. It can be vertical, as in the inversion case above and in the wind gradient described in the section on the troposphere, or horizontal, as at a front or on the flank of a thunderstorm. What makes it hazardous is not the wind itself but the *rate of change*, because the aircraft's inertia prevents it from adjusting instantly and its airspeed changes as a result.

Two points frequently appear as distractors and both are false. Wind shear is **not** confined to thunderstorms — fronts, inversions, terrain, sea breezes and even large buildings all produce it — and it is **not** the same thing as a steady strong wind, which produces no aerodynamic effect at all once the aircraft has settled into it. The most severe form, the microburst, is a concentrated downdraught that spreads out radially on reaching the ground, so an aircraft flying through one meets a strong headwind, then a violent downdraught, then a strong tailwind within a few seconds — a performance-decreasing sequence at the worst possible moment.



A microburst spreads out at the ground, giving a headwind, then a downdraught, then a tailwind.

Jet streams and clear air turbulence

A **jet stream** is a narrow, fast-flowing ribbon of wind, typically a few thousand feet deep and a few hundred miles wide, found near the tropopause with core speeds commonly between **100 and 200 knots** and occasionally far more. It forms where two air masses of markedly different temperature lie side by side, because a horizontal temperature gradient causes the wind to strengthen with height — the greater the temperature contrast, the stronger the wind aloft. The strongest contrasts are at the boundaries between polar and tropical air, and that is where the principal jet streams are found, running broadly west to east and meandering north and south in great waves.

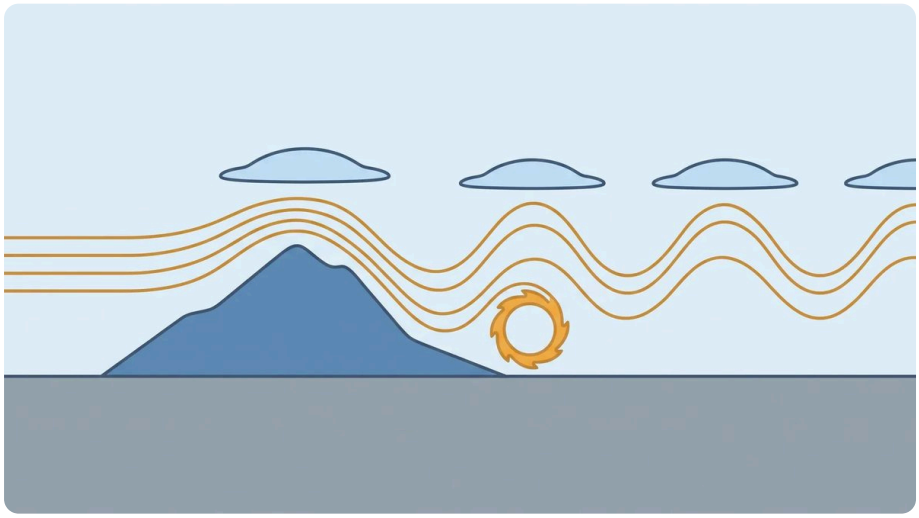
Two operational consequences follow. A jet stream is a very large fuel and time lever: a 150 kt tailwind added to a 480 kt true airspeed makes a decisive difference to a transatlantic sector, which is why eastbound and westbound routings across an ocean are often hundreds of miles apart. And the same feature is the principal source of **clear air turbulence**.

CAT is turbulence at high altitude produced by the wind shear at the *edges* of a jet stream, where the core's speed falls away to that of the surrounding air over a relatively short distance. It has three properties that together make it a genuine hazard:

- **It gives no visible warning.** The air at those altitudes is far too dry to form cloud, so there is nothing to see and nothing for weather radar to detect — airborne weather radar returns come from precipitation, and CAT contains none.
- **It can be severe.** The description "clear air" refers only to the absence of cloud, not to the intensity, and CAT encounters have caused serious injuries to unrestrained occupants and, occasionally, structural loads requiring inspection.
- **It is hard to forecast precisely.** Its location can be predicted only as a region of probability, which is why the operational defence is the seat-belt sign and the reporting of encounters rather than avoidance by observation.

Mountain waves

When a stable airstream meets a mountain range more or less at right angles, the air is forced up over the ridge. Being stable, it is negatively buoyant on the far side and sinks back down — then overshoots, rises again, and continues to oscillate downwind in a train of **standing waves**. The waves are stationary relative to the ground even though the air is flowing rapidly through them, which is why they are also called standing or lee waves.



Stable air forced over a ridge oscillates downwind in standing waves, with lenticular cloud and rotor below.

The conditions required are specific: a stable layer, a wind of reasonable strength blowing roughly perpendicular to the ridge, and wind that does not change direction greatly with height. Given those, the effects extend far beyond the terrain that caused them:

- **Strong vertical currents** in the wave crests and troughs, capable of exceeding the climb or descent performance of an aircraft, so that a constant power and attitude produce an uncommanded climb or descent.
- **Rotor** — violently turbulent horizontal roll circulations beneath the wave crests, close to the terrain, which produce the most severe turbulence associated with the phenomenon.
- **Great vertical extent.** Mountain waves are not a low-level phenomenon confined to the hills. They routinely propagate into the upper troposphere and can reach the stratosphere, so an aircraft at normal cruising levels far downwind of a range is well within their reach.
- **Lenticular clouds**, smooth lens-shaped clouds marking the crests where the rising air has saturated, are one of several visual indications, and they appear only when the air is moist enough. The others belong to the same flow: the rotor carries its own rotor cloud, ragged and visibly turbulent; a cap cloud sits on the summit while a banner cloud streams downwind from it; and dust or snow raised off the surface shows the strength of the low-level wind. Their absence proves nothing: a severe wave event can occur with little or no visual sign at all.

The Coriolis effect and the shape of the general circulation

Air is driven from high pressure towards low pressure by the pressure gradient, but it does not travel in a straight line between them, because the earth is rotating beneath it. The **Coriolis effect** is the apparent deflection this produces when the motion is viewed from the rotating surface. Its direction is fixed by the hemisphere and must be quoted correctly:

- In the **northern hemisphere**, moving air is deflected to the **right** of its direction of travel.
- In the **southern hemisphere**, moving air is deflected to the **left**.
- The effect is zero at the equator and greatest at the poles, and it grows with the speed of the air.

The consequence is that air does not flow straight into a low-pressure centre and fill it. Instead the deflection turns the flow until it is running very nearly *along* the isobars, balanced between the pressure gradient pulling it inward and the Coriolis effect pushing it outward. Winds therefore **circulate around** pressure systems rather than through them: anticlockwise around a low and clockwise around a high in the northern hemisphere, and the reverse in the southern. It is this balance, not the pressure gradient alone, that gives weather systems their characteristic rotation, and the same effect acting on the great convective cells of the tropics is what turns them into the steady trade winds rather than a simple flow from equator to pole.

Buys Ballot's law, which is the same fact in a usable form: in the northern hemisphere, stand with your back to the wind and the low pressure is on your left. In the southern hemisphere, it is on your right. It follows directly from the deflection rules above and is a quick way of recovering them if the direction of the Coriolis deflection has been forgotten.

Two points of scope are worth making. The Coriolis effect acts on **any** freely moving mass over the earth, atmosphere and ocean alike — it is not restricted to ocean currents, and it is certainly not absent from the atmosphere. But it is significant only over large distances and long times, so it plays no part in the airflow over a wing, in a hangar, or in any other small-scale motion.

How cloud forms: four ways of cooling air to its dew point

Cloud is condensed water, and condensation requires the air to be cooled to its dew point. Almost all atmospheric cooling is achieved by lifting the air, because rising air expands and cools adiabatically without any heat needing to be removed from it. The four lifting mechanisms account between them for essentially all cloud:

Mechanism	What lifts the air	Characteristic cloud	Where it is met
Convection	Surface heating makes a parcel buoyant, and it rises of its own accord	Cumulus, growing to cumulonimbus in unstable air	Over land on sunny afternoons; over warm sea in cold air masses
Orographic lifting	Terrain forces the airstream upward	Cloud capping and streaming from high ground; lenticular cloud in wave crests	Upwind slopes of hills and mountains

Mechanism	What lifts the air	Characteristic cloud	Where it is met
Frontal lifting	Warm air is undercut by, or rides over, colder air at a front	Layered cloud over a wide area at a warm front; vigorous convective cloud at a cold front	Along frontal systems, often hundreds of miles wide
Convergence	Air flowing into a region from several directions has nowhere to go but up	Extensive cloud and precipitation	Low-pressure centres; the inter-tropical convergence zone

A fifth process needs no lifting at all: air in contact with a surface colder than its own dew point is chilled directly, and the result is fog rather than cloud. The distinction between fog and cloud is only one of height — fog is cloud in contact with the ground — and the same three cooling routes account for the main types:

- **Radiation fog** forms overnight when the ground cools by radiation under a clear sky and chills the air above it. It needs light wind: dead calm gives dew or frost with no mixing to spread the cooling upward, while strong wind mixes the layer too deeply to saturate. It usually clears with surface heating after sunrise.
- **Advection fog** forms when warm, moist air moves horizontally over a colder surface — typically warm maritime air moving over a cold sea or over snow. Unlike radiation fog it can form and persist in strong winds and can last for days, which makes it the more operationally troublesome of the two.
- **Steam fog** forms when very cold air moves over much warmer water; evaporation saturates the cold air almost immediately. It is shallow and usually short-lived.

Supercooled water and airframe icing

Water in the atmosphere does not reliably freeze at 0 °C. Small droplets suspended in clean air frequently remain liquid well below freezing — they are then **supercooled** — because freezing requires a nucleus to start on and a small, pure droplet may not encounter one. Supercooled droplets are common down to about –20 °C and are found, less often, considerably colder than that.

The consequence is the whole of airframe icing. A supercooled droplet is in a metastable state, and striking an aircraft surface provides exactly the disturbance it needs: it freezes on impact. Three forms result, and which one appears depends on the droplet size and the temperature, because those two together decide how much of the droplet freezes instantly and how much runs back before freezing:

Type	Forms in	Appearance and mechanism	Why it matters
Rime ice	Small droplets, lower temperatures	Opaque, white and granular. Each droplet freezes almost instantly on impact, trapping air between the frozen droplets	Builds forward from leading edges, spoiling the aerofoil shape; comparatively brittle and light
Clear (glaze) ice	Large droplets, temperatures close to freezing	Transparent, dense and smooth. Only part of each droplet freezes on impact; the rest runs back before freezing	Heavier, harder to shed, and can form *behind* the protected area — the run-back that gives it its danger
Hoar frost	On a cold airframe, on the ground or after rapid descent	Feathery crystals deposited directly from vapour to solid, without a liquid stage	Increases surface roughness over large areas; obscures windows; must be removed before flight

The three conditions for airframe icing must all be present at once, and the absence of any one of them prevents it: **visible moisture** (cloud, rain, drizzle or wet snow); an **airframe surface temperature at or below 0 °C**; and **droplets that are liquid** when they arrive. This is why flight in cloud at –40 °C is often ice-free — the water there is already ice crystals, which bounce off rather than adhere — and why the worst accumulation is usually found in the band from about 0 °C to –15 °C, where large supercooled droplets are most abundant. It also explains why an aircraft descending rapidly from a cold cruise into warm moist air can acquire ice on a cold-soaked airframe in air that is well above freezing.

Precipitation, and one hazard it produces

Cloud droplets are far too small to fall out of the sky at any useful rate; they must first grow by a factor of many thousands in volume. Two processes achieve this: collision and coalescence between droplets of different sizes in warm cloud, and, in cloud containing both ice crystals and supercooled water, the growth of the ice crystals at the expense of the droplets. The second is far the more effective, which is why most precipitation in temperate latitudes begins its life as ice even when it reaches the ground as rain.

One product of a frontal inversion deserves specific mention because it is among the most dangerous icing conditions there is. **Freezing rain** occurs where rain falls from a warm layer aloft into a sub-freezing layer beneath it — the temperature profile of a warm front. The drops are large, they are supercooled by the time they arrive, and they arrive in a location where an aircraft is likely to be flying an approach. Large supercooled drops produce extensive clear ice with substantial run-back, potentially beyond the protected surfaces, and the accumulation rate can exceed the capability of the ice protection system. The presence of freezing rain implies warmer air above, so the standard escape is to climb rather than to descend — the reverse of the usual instinct.

Thunderstorms: the whole chapter in one phenomenon

A thunderstorm is the atmosphere's instability, latent heat and moisture mechanisms operating together at full strength, and it is worth tracing because every element of it has already appeared in this note. Three ingredients are required: **moisture**, **instability** and a **lifting trigger**. Given all three, the sequence is:

- **Developing stage.** The trigger lifts moist air to its condensation level. Above that level the parcel cools at the saturated rather than the dry rate, so it stays warmer than its surroundings and accelerates upward. The cloud is dominated by updraughts throughout.
- **Mature stage.** Precipitation begins to fall, and falling precipitation drags air down with it while evaporating into it and cooling it — producing a downdraught alongside the continuing updraught. This coexistence is what makes the mature stage the most violent: severe turbulence from the shear between the two, hail, lightning, and a spreading gust front where the downdraught reaches the ground. The updraught continues until it reaches the tropopause, where the stable stratospheric air above stops it and the cloud spreads horizontally into the characteristic anvil.
- **Dissipating stage.** The downdraught eventually cuts off the supply of warm moist air to the updraught, the storm loses its energy source, and precipitation and turbulence gradually decay.

Note that the anvil top is a direct demonstration of the layer structure described earlier: the storm is capped at the tropopause because that is where the environmental profile changes from unstable to stable, and the

cloud can penetrate a short distance into the stratosphere on momentum alone before spreading out beneath the lid.

Induction icing: ice formed by the engine itself

The airframe icing described above requires the ambient air to be at or below freezing. Induction icing does not, and the reason is a direct application of the adiabatic and latent-heat processes set out earlier in this note. Air being drawn into a piston engine's induction system undergoes two cooling processes at once:

- **Pressure drop.** Accelerating the air through a venturi and past a partly closed throttle drops its static pressure, and expanding air cools.
- **Fuel vaporisation.** Liquid fuel sprayed into that airstream evaporates, and evaporation absorbs latent heat — from the air. This is the larger of the two effects.

Together these can drop the temperature inside a carburettor by a great deal — enough that ice can form there when the outside air is well above freezing, in conditions that look entirely benign. The requirement is simply that the air contains enough moisture and that the cooling is sufficient to take it below freezing, so the classic conditions are a **warm, humid day at low power**, with the throttle partly closed so the pressure drop across it is greatest. A descent at low power on a humid summer afternoon is the textbook case, and the fact that it is warm and clear is exactly why the hazard is so often not anticipated.

> **The symptom sequence differs between propeller types and is worth knowing.** With a *fixed-pitch* propeller, ice restricting the induction system reduces power, so the engine and propeller slow and the rpm falls. With a *constant-speed* propeller, the governor holds the rpm by fining the blade pitch, so the rpm does **not** fall — the loss shows instead as a fall in manifold pressure. In both cases the engine eventually runs rough and stops if the accumulation continues. Carburettor heat is the countermeasure, and it works by admitting air warmed by the exhaust system, which both melts existing ice and keeps the mixture above freezing; using it reduces the density of the induction air and therefore costs power, which is the price of the protection.

Turbine engines are not immune, although the mechanism is different. There is no carburettor and no fuel evaporating in the intake, but the intake itself accelerates air, and the resulting pressure drop cools it — so a turbine intake can be several degrees colder than the ambient air and can accumulate ice at ambient temperatures slightly above freezing. Ice shed from an intake lip into the compressor is a foreign-object event, which is why intake anti-icing is normally applied on the basis of temperature *and* visible moisture rather than waiting for evidence of accumulation, and why the trigger temperature is set a few degrees above zero rather than at it.

Air masses: why a body of air behaves the way it does

An **air mass** is a large body of air — hundreds or thousands of miles across — whose temperature and humidity are broadly uniform horizontally. It acquires those properties by sitting over a source region long enough to come into equilibrium with it, and it then carries them away when it moves. Two questions settle everything about how it will behave, and both have already been answered in general terms in this note: *what did it pick up at its source*, and *what is happening to it now*.

Source characteristic	Label	What the air mass brings
Ocean	**Maritime**	High moisture content, moderate temperatures
Large land mass	**Continental**	Low moisture content, extreme temperatures — very hot in summer, very cold in winter
High latitude	**Polar** or **Arctic**	Cold, and in the arctic case very cold and very dry
Low latitude	**Tropical**	Warm, and when maritime also very moist

The two labels combine, so a maritime polar air mass is cool and moist, a continental tropical one is hot and dry, and so on. That is the source. The second question — what is happening to it now — is where the stability material above does the work:

- **An air mass moving over a surface WARMER than itself is heated from below.** Heating the bottom of a layer steepens its lapse rate, so it becomes **less stable**. Convection develops, cumulus forms, showers become possible, turbulence increases — and visibility beneath the cloud is *good*, because the vertical mixing disperses haze through a deep layer. Cold air arriving over a warm sea in winter is the archetype.
- **An air mass moving over a surface COLDER than itself is cooled from below.** Cooling the bottom flattens the lapse rate, so it becomes **more stable**. Vertical motion is suppressed, any cloud is layered, and visibility beneath is *poor* because pollutants and moisture are trapped near the surface. If the cooling takes the air below its dew point, advection fog results. Warm moist air arriving over a cold sea or over snow is the archetype.

The rule in one line: warm air over a cold surface is stabilised and gives fog, stratus, drizzle and poor visibility; cold air over a warm surface is destabilised and gives cumulus, showers, turbulence and good visibility. Note that the more comfortable-sounding of the two — the smooth stable one — is the one that closes aerodromes.

Fronts: the boundary between two air masses

A **front** is the sloping surface where two air masses of different temperature meet. They do not mix readily, because the density difference between them resists it; instead the colder, denser air stays beneath and the warmer, less dense air is displaced above. The front is therefore always a *sloping* surface, not a vertical wall, and the weather associated with it is the weather of air being lifted along that slope.

	Warm front	Cold front
What is advancing	Warm air, riding up over retreating cold air	Cold air, undercutting and lifting the warm air ahead of it
Slope	Shallow — typically of the order of 1 in 150	Steep — typically of the order of 1 in 50
Consequence of the slope	Lifting is gentle and spread over a great distance, so cloud arrives many hours ahead of the surface front	Lifting is abrupt and concentrated, so the weather is narrow and violent

	Warm front	Cold front
Cloud sequence	Cirrus, then cirrostratus, altostratus, and finally nimbostratus at the surface front — a descending, thickening layer	Towering cumulus and cumulonimbus in a narrow band, often as a squall line
Precipitation	Continuous and prolonged, becoming heavier as the front approaches	Heavy, showery and short-lived, frequently with hail and thunder
Icing risk	Widespread and persistent; freezing rain where rain falls from the warm air into sub-freezing air beneath	Severe but localised, in the convective cloud
After passage	Temperature rises, cloud lowers to stratus, visibility often poor in the warm sector	Temperature falls sharply, pressure rises, showers with bright intervals and excellent visibility
Wind change (northern hemisphere)	Veers — changes clockwise	Veers sharply — a marked clockwise change with the passage

An **occlusion** forms when a cold front, which travels faster, catches up with the warm front ahead of it and lifts the whole of the warm sector clear of the ground. The result combines the characteristics of both: extensive layered cloud from the warm-front process together with embedded convective cloud from the cold-front process, which is why an occlusion is associated with prolonged, mixed and often awkward weather rather than with a clean sequence.

Every front carries a wind shear, by definition. Two air masses with different densities and different velocities are separated by a sloping surface, so an aircraft crossing that surface passes from one wind field into another. On a slowly moving warm front the change may be gentle; at an active cold front it can be abrupt, and the combination of shear, turbulence and heavy precipitation in a narrow band is why cold-front crossings are planned around rather than flown through.

Why the tropopause slopes across a front

One consequence ties the fronts back to the layer structure described earlier. Warm air masses have deep troposphere and high tropopauses; cold ones have shallow troposphere and low tropopauses. Where the two meet at a front, the tropopause is therefore stepped rather than continuous — the tropopause break introduced above. The strong horizontal temperature contrast that defines the front at the surface is the same contrast that drives the jet stream aloft, which is why the polar front jet is found above the polar front rather than anywhere else. **The surface front, the tropopause break, the jet stream and the clear air turbulence at its edges are four aspects of one feature**, and an aircraft crossing a major frontal system at cruising level is crossing all of them at once.

The general circulation: why the winds blow where they do

The Coriolis effect explains why moving air turns. What sets it moving in the first place is the uneven heating of the earth: the sun's energy arrives nearly vertically at the equator and at a glancing angle near the poles, so

the tropics gain more heat than they radiate away and the polar regions lose more than they gain. The atmosphere and the oceans exist, from a physical point of view, to carry the surplus from one to the other.

A single overturning circulation from equator to pole would do the job if the earth did not rotate. Because it does, the Coriolis deflection breaks that single cell into three in each hemisphere:

Cell	Latitudes	Vertical motion	Surface wind produced	Characteristic weather
Hadley cell	Equator to about 30°	Rising at the equator, sinking near 30°	The **trade winds**, blowing towards the equator and deflected to become north-easterly in the northern hemisphere and south-easterly in the southern	Heavy convective rainfall at the equatorial convergence; the world's great deserts under the subsiding branch near 30°
Ferrel cell	About 30° to about 60°	Sinking near 30°, rising near 60°	The prevailing **westerlies** of the mid-latitudes	The travelling depressions and fronts of temperate latitudes
Polar cell	About 60° to the pole	Rising near 60°, sinking at the pole	The **polar easterlies**	Cold, dry, generally settled conditions over the polar high

Two features of that picture matter operationally. The boundary between the Ferrel and polar cells, near 60°, is the **polar front** — a permanent zone of contrast between cold polar air and warmer mid-latitude air, and the birthplace of most of the depressions that cross the temperate latitudes. And the subsiding branch near 30° is the **subtropical high-pressure belt**, where descending air warms adiabatically, its relative humidity collapses, and cloud is suppressed — a subsidence inversion operating on a planetary scale. The deserts sit there for exactly that reason.

The equatorial rising branch is the **inter-tropical convergence zone**, where the trade winds of the two hemispheres meet and have nowhere to go but up. It is a band of deep convection, heavy rain and vigorous thunderstorm activity that migrates north and south with the seasons, and it is the single most significant weather feature on tropical routes.

Local winds: the same physics at a smaller scale

Several winds an aircraft meets are produced not by the general circulation but by local temperature differences, and each is a small-scale version of a mechanism already described:

- **Sea breeze.** By day, land heats faster than water because its heat capacity is lower and the heating is concentrated at the surface. The air over the land warms, expands and rises, lowering the surface pressure there, and air flows in from the sea to replace it. The direction is therefore *from sea to land* in the afternoon. Convergence at the sea-breeze front frequently triggers cumulus development a few miles inland.
- **Land breeze.** At night the land cools faster than the sea, the situation reverses, and a weaker breeze blows *from land to sea*. It is weaker because the night-time temperature contrast is smaller than the daytime one.

- **Anabatic wind.** By day, a sunlit slope heats the air in contact with it, which becomes buoyant and flows *up* the slope. It is a gentle wind and a favourite of glider pilots.
- **Katabatic wind.** At night, a slope cools by radiation, the air in contact with it becomes cold and dense, and gravity drains it *down* the slope. Over a large ice sheet or a long valley the drainage can accumulate into a strong and persistent wind, and it can arrive at a valley aerodrome as a substantial and unexpected downslope flow on an otherwise calm night.
- **Foehn wind.** Air forced over a mountain range cools at the dry adiabatic rate until it saturates, then more slowly at the saturated rate while precipitating out its moisture on the windward side. Having lost that moisture, it descends the lee side at the *dry* adiabatic rate all the way down — faster than it cooled on the way up. It therefore arrives at the same altitude on the far side both **warmer and much drier** than it started. This is one of the clearest demonstrations that the two adiabatic rates are genuinely different numbers, and it produces the dramatic temperature rises and humidity collapses associated with such winds.

The foehn effect worked through with the standard rates. Take air at 15 °C at sea level forced over a 3,000 m ridge, saturating at 1,000 m. It cools at about 10 °C/km for the first kilometre, reaching 5 °C, then at about 5 °C/km for the remaining two kilometres, reaching –5 °C at the summit. Descending the lee side it is now unsaturated, so it warms at the dry rate of about 10 °C/km for the whole 3,000 m, arriving at sea level at about **+25 °C** — ten degrees warmer than it set out, with all of that difference having come from the latent heat released into it during the ascent.

Pressure patterns and the winds they produce

Horizontal pressure variation was introduced above as the reason QNH is a local quantity. Its other consequence is the wind, and the patterns it forms are the shapes seen on any surface weather chart. Pressure is plotted as **isobars** — lines joining points of equal sea-level pressure — and the resulting map contains a small vocabulary of features:

Feature	Shape on the chart	Vertical motion	Typical weather
Low (depression, cyclone)	Closed isobars around a pressure minimum	Rising air, converging at the surface	Cloud, precipitation, wind; fronts are found within depressions
High (anticyclone)	Closed isobars around a pressure maximum	Sinking air, diverging at the surface	Settled, often cloud-free; subsidence inversion aloft trapping haze; fog in winter
Trough	An elongated extension of low pressure	Rising	A band of unsettled weather, often with a front along it
Ridge	An elongated extension of high pressure	Sinking	A band of temporarily improved weather between depressions
Col	The neutral area between two highs and two lows	Weak and indeterminate	Light and variable winds; fog and low cloud in winter, thunderstorms in summer

Two rules connect the chart to the wind, and both follow from material already established. The **spacing of the isobars gives the wind speed**: closely spaced isobars mean a large pressure change over a short distance, a strong pressure gradient force, and therefore a strong wind. And the **orientation of the isobars gives the wind direction**, because the Coriolis deflection turns the flow until it runs nearly parallel to them rather than across them.

The parallel-flow result is not exact near the surface. Friction slows the air in the boundary layer, and slower air experiences less Coriolis deflection, so the balance shifts and the surface wind blows partly *across* the isobars towards the low pressure. Two practical consequences follow, both stated for the northern hemisphere:

- **The surface wind is backed relative to the wind just above the friction layer**, by roughly 10 to 20 degrees over the sea and by more over rough land, where the friction is greater. Climbing out of the friction layer, the wind therefore veers and strengthens — which is precisely the wind gradient described earlier, now with a direction change as well as a speed change.
- **Air spirals inward at the surface of a low and outward at the surface of a high.** That inflow is the convergence that forces air upward in a depression, and the outflow is the divergence that draws air downward in an anticyclone. The vertical motion column in the table is a consequence of friction acting on the horizontal flow, not a separate fact to be memorised.

Why an anticyclone means good weather and a depression means bad, in one sentence each. In a high, air is sinking; sinking air is compressed and warms adiabatically; warming air has its relative humidity reduced, so cloud evaporates rather than forms. In a low, air is rising; rising air expands and cools adiabatically; cooling air approaches saturation, so cloud forms and precipitation follows. The entire distinction is the adiabatic process running in the two directions.

Visibility and optical effects

Visibility is an atmospheric property in its own right, determined by how much of the light from a distant object survives its passage through the intervening air. Two mechanisms reduce it, and they behave differently:

- **Scattering by particles** — water droplets in fog, mist and cloud, and solid aerosols in haze, smoke, dust and sand. The particles redirect light out of the line of sight and also scatter ambient light into it, which is why a fog appears luminous rather than dark and why looking towards the sun through haze is worse than looking away from it.
- **Absorption**, which is comparatively minor in the visible band and matters more for the infrared and radio wavelengths discussed below.

The distinction between fog and mist is one of degree and is defined by the resulting visibility rather than by any difference of mechanism: both are suspended water droplets, and the threshold conventionally used is 1,000 m. Haze, by contrast, is composed of dry particles, and it is the characteristic obscuration of stable air, for the trapping reason set out above. This is why the smoothest flying days are frequently the ones with the poorest slant visibility on the approach, and why visibility looking vertically downward through a haze layer can be far better than the visibility looking obliquely through it — the oblique path traverses several times as much of the layer.

Two optical effects arise from refraction rather than scattering. Because the density of air varies with height, and the refractive index varies with density, light travelling near-horizontally through the atmosphere is bent slightly downward, following the curvature of the earth a little. The practical consequences are that objects at the horizon are visible slightly further away than pure geometry allows, and that this bending increases when there is a strong temperature gradient near the surface — which is the origin of mirages over a hot runway, where the intense surface heating creates a shallow layer of very low density that refracts light sharply upward.

Turbulence: one word for several different phenomena

Turbulence is irregular motion of the air, and an aircraft flying through it experiences rapid changes in the direction and speed of the relative airflow — which is to say rapid changes in angle of attack, and therefore in lift and in load factor. Several quite different atmospheric processes produce it, and each has already appeared in this note under its own heading:

Type	Cause	Where and when	Depth
Convective (thermal)	Buoyant parcels rising and cooler air descending in unstable air	Over land on sunny days; over warm water in cold air masses	Through the convective layer — usually the lowest few thousand feet, far deeper in vigorous conditions
Mechanical	Friction and obstruction: wind flowing over rough terrain, buildings, trees and hangars	Low level, on windy days; worse downwind of obstacles	Shallow — the friction layer
Orographic	Terrain forcing air upward, and the wave and rotor systems that result	Near and downwind of hills and mountains	Surface to the stratosphere in a well-developed wave
Shear	Adjacent layers of air moving at different velocities	At inversions, at fronts, and at the edges of jet streams — where it is called clear air turbulence	Confined to the sheared layer, which may be thin
Frontal	Lifting and the velocity change across the frontal surface	In the vicinity of an active front, especially a cold one	Through the depth of the frontal cloud

The intensities are reported in standard terms — light, moderate, severe and extreme — defined by the aircraft's response rather than by any measurement of the air, because the same air produces different responses in different aircraft. A light aircraft and a large transport crossing the same thermal will report differently and both will be right.

Why turbulence produces a load, and why there is a speed for it

The link back to the atmosphere is through dynamic pressure. A vertical gust arriving at an aircraft in level flight changes the direction of the relative airflow, and therefore changes the angle of attack by an amount that depends on the ratio of the gust velocity to the aircraft's own true airspeed. That change of angle of attack

produces a change of lift, and the change of lift divided by the aircraft's weight is the change in load factor. Two consequences follow, and both are directions worth stating explicitly:

- **For a given gust, the load increment grows in direct proportion to the indicated airspeed.** A vertical gust of a given strength changes the angle of attack by less at a higher speed, but the lift produced by each degree of angle of attack grows with the square of the speed, and the second effect wins — so flying faster through the same turbulence produces larger loads, not smaller ones. This is why aircraft have a recommended rough-air or turbulence-penetration speed, and why it is quoted as an indicated airspeed rather than a true one.
- **The recommended speed is not simply "as slow as possible", however**, because slowing down reduces the margin to the stall while the gust is trying to increase the angle of attack. The turbulence-penetration speed is the compromise between those two competing limits — low enough that gust loads stay within the structure's capability, high enough that a gust-induced angle-of-attack increase does not reach the stall. Naming both limits is the whole of the reasoning behind it.

Air density enters the same relationship, and this produces a result that is initially surprising. At high altitude the air is thin, so a gust of a given *true* vertical velocity produces a *smaller* load increment for the same indicated airspeed than it would at sea level. Turbulence at cruising level is genuinely gentler in terms of structural load than the same gust at low level would be — but the aircraft is also operating with much less margin between its low-speed and high-speed limits up there, so the operational significance of an encounter is not correspondingly reduced.

Atmospheric electricity

The atmosphere is not electrically neutral, and the charge it carries has two distinct consequences for aircraft: lightning, which is an external discharge between charge centres, and precipitation static, which is charge the aircraft acquires for itself.

Charge separation occurs in a vigorous cumulonimbus, where strong updraughts carry supercooled water and small ice crystals upward past larger, heavier ice pellets falling under gravity. Collisions between them transfer charge, and because the two populations are then carried in opposite directions the cloud ends up with the charges physically separated — broadly, a large positive region in the upper part of the cloud and a large negative region lower down. The potential difference between them builds until the insulating strength of the air between is exceeded, and the result is a lightning discharge, within the cloud, between clouds, or to the ground.

The requirement for strong updraughts and mixed-phase precipitation explains why lightning is essentially confined to cumulonimbus and effectively never occurs in stratiform cloud, however thick. It is the same instability, latent heat and vertical motion that produce every other thunderstorm hazard, expressed electrically.

An aircraft in flight is an isolated conductor, and that is the design assumption. A lightning strike attaches at an extremity — nose, wingtip, fin tip, or a trailing edge — passes through the structure and exits at another extremity, continuing the discharge on its way. The airframe is not damaged by carrying the current provided it presents a continuous, low-resistance path; the damage occurs where the current is forced to jump a gap or to pass through a poor joint, because that is where it deposits its energy. Two consequences follow for maintenance:

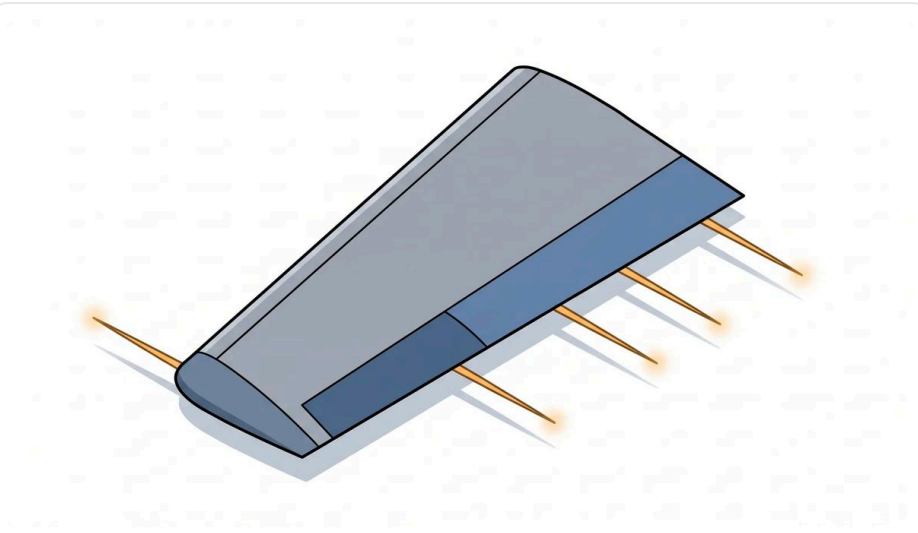
- **Bonding is a safety function, not a housekeeping one.** Every structural joint, every control surface and every removable panel needs an electrical path to the rest of the airframe, and bonding leads exist to guarantee one where the structural joint alone cannot. A missing or high-resistance bond turns a harmless through-current into a localised burn.
- **Composite structure needs a conductive layer added.** Carbon fibre conducts poorly compared with aluminium and glass fibre not at all, so lightning protection is built in as an embedded mesh or foil that provides the conductive path an aluminium skin would have provided by itself. Damage to that layer is damage to a system, and its repair is prescribed accordingly.

Precipitation static

Separately from lightning, an aircraft flying through precipitation, cloud, dust or ice crystals accumulates charge simply by colliding with the particles — a triboelectric effect, the same one that charges a comb drawn through hair. The airframe potential rises until the charge starts to leak away from the sharpest available points, and that discharge is what causes the trouble. It is a broadband radio-frequency noise source sitting directly on the airframe, and its symptoms are characteristic:

- Loud crackling or hissing on HF and on the lower-frequency navigation receivers, sometimes to the point of making them unusable.
- Erratic behaviour of automatic direction finding.
- In extreme cases a visible corona discharge from extremities and around windscreens at night, historically called St Elmo's fire.

Static discharge wicks are the countermeasure. They are deliberately sharp, high-resistance points fitted at the trailing extremities — wingtips, aileron and elevator trailing edges, fin and stabiliser tips — positioned in the aerodynamic wake so that the discharge leaves the aircraft downstream of the antennas rather than from a point near them. The resistance is deliberate: it limits the current in each discharge so that it produces less radio noise than an uncontrolled arc from a random point would. Their condition is therefore an airworthiness item even though nothing on the aircraft stops working when one is missing, and their resistance is checked rather than merely their presence, because a wick that has aged into a low-resistance path has stopped doing the job it was fitted for.



Sharp static wicks at the wingtip and aileron trailing edge let built-up charge leak away into the wake.

Pressure Variation with Altitude — Detailed

Atmospheric pressure at any point is caused by the **weight of all the air above** that point. As you climb higher, there is less air above you, so the pressure decreases. The key points about pressure variation are:

- **Pressure decrease is NOT linear** — it is *exponential*. Pressure drops rapidly at low altitudes (where the air is dense and heavy) and more slowly at high altitudes (where the air is thin and light). Near sea level, pressure drops by about 1 hPa for every 27 feet of altitude gain (commonly rounded to 30 ft as a rule of thumb). At 18,000 ft, the same 1 hPa drop requires about 50 feet of altitude change.
- **The pressure halving rule** — a useful approximation: pressure halves approximately every 18,000 ft (5.5 km). Starting from 1013 hPa at sea level: ~500 hPa at 18,000 ft, ~250 hPa at 36,000 ft, ~125 hPa at 54,000 ft. The rule is accurate for the first halving and progressively over-estimates above it — the true ISA pressure at 36,000 ft is 227 hPa, not 250 (see the table below).

Altitude	ISA Pressure (hPa)	% of Sea Level	Flight Level
Sea level (0 ft)	1013.25	100 %	—
5,000 ft (1,524 m)	843	83 %	FL050
10,000 ft (3,048 m)	697	69 %	FL100
18,000 ft (5,486 m)	507	50 %	FL180
25,000 ft (7,620 m)	376	37 %	FL250
30,000 ft (9,144 m)	301	30 %	FL300
35,000 ft (10,668 m)	238	24 %	FL350
36,089 ft (11,000 m)	227	22 %	FL360
40,000 ft (12,192 m)	188	19 %	FL400

The hydrostatic equation: where the whole pressure profile comes from

The statement that pressure is the weight of the air above can be written as an equation, and doing so turns a piece of physical reasoning into a calculation that produces every number in the table above.

$$\Delta p = -\rho\, g\, \Delta h$$

Where Δp is the change in pressure (Pa) over a height change Δh (m), ρ is the density of the air in that height band (kg/m³) and g is 9.80665 m/s². The minus sign records that climbing reduces the pressure.

Read this carefully and the exponential behaviour explains itself. **The rate at which pressure falls with height is set by the density of the air at that height** — and density itself falls as pressure falls. So the pressure gradient is steep where the air is dense and shallow where it is thin, which is precisely the shape of an exponential decay: a quantity whose rate of decrease is proportional to how much of it is left.

Worked example — deriving the "1 hPa per 27 ft" figure. At sea level the standard density is 1.225 kg/m³, so

$$\frac{\Delta p}{\Delta h} = \rho g = 1.225 \times 9.80665 = 12.01 \text{ Pa per metre}$$

One hectopascal is 100 Pa, so one hectopascal corresponds to

$$\frac{100}{12.01} = 8.33 \text{ m} = 27.3 \text{ ft}$$

which is the 27 ft quoted above.

The same calculation at 18,000 ft, where the standard density has fallen to 0.698 kg/m³:

$\rho g = 0.698 \times 9.80665 = 6.85 \text{ Pa per metre}$, so one hectopascal now takes $100/6.85 = 14.6 \text{ m} = 47.9 \text{ ft}$ — the "about 50 feet" quoted above. Neither figure was measured or memorised; both fall straight out of the density at the level in question.

Continuing the same calculation up the atmosphere gives a table that is worth having, because it shows just how far the relationship departs from the sea-level rule that altimetry is built on:

Altitude	ISA density (kg/m³)	Height change for 1 hPa
Sea level	1.225	27 ft
5,000 ft	1.056	32 ft
10,000 ft	0.905	37 ft
18,000 ft	0.698	48 ft
30,000 ft	0.458	73 ft
40,000 ft	0.302	111 ft

At 40,000 ft one hectopascal is worth four times as many feet as it is at sea level. This has a direct consequence for the altimeter: since the instrument converts pressure to height, the same small pressure error produces a much larger height error at altitude than near the ground. It is also why the pressure-altitude correction of 27 or 30 ft per hectopascal is applied only to the QNH setting, which is a sea-level quantity, and never to a pressure difference measured high up.

Scale height, and what "exponential" means numerically

Where the temperature is constant, the hydrostatic equation has an exact solution: the pressure falls by a constant *factor* for each equal increment of height. That increment is the **scale height**, and it is the height over which the pressure falls to 1/e — about 37 % — of its value:

$$H = \frac{RT}{g}$$

At the standard sea-level temperature of 288.15 K this gives $H = 8{,}435$ m, about 27,700 ft. In the isothermal stratosphere at 216.65 K it gives $H = 6{,}342$ m, about 20,800 ft.

The scale height is proportional to *temperature*, and that is the physically important part. **A warm atmosphere is a "taller" one**, in which pressure falls off more gradually with height; a cold atmosphere is compressed into a shallower depth and its pressure falls off more steeply. This is the same fact, expressed in a different vocabulary, as the altimeter temperature error discussed earlier: cold air packs the pressure levels closer together, so a given pressure is found at a lower true height.

The halving rule quoted above is this exponential in disguise. Halving takes 0.693 of a scale height — the natural logarithm of 2 — which at sea-level temperatures is about 19,200 ft, close to the 18,000 ft of the rule of thumb. The rule degrades with altitude for exactly the reason the scale height falls: the air aloft is colder, so the true halving interval shortens, and a rule assuming a fixed 18,000 ft over-estimates the pressure that remains.

The behaviour above the tropopause

Nothing in the derivation above referred to any particular layer, so nothing changes at the tropopause. **Pressure continues to decrease exponentially through the tropopause and all the way up**, with no discontinuity, no flattening and certainly no reversal. The only thing that changes is the scale height, which stops shrinking once the temperature stops falling.

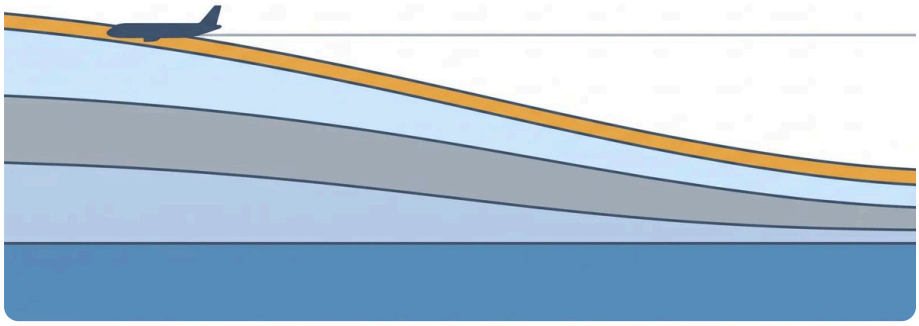
One consequence of that is worth noting because it simplifies calculations in the stratosphere. In an isothermal layer, the equation of state makes density directly proportional to pressure, since the temperature in the denominator is a constant. **Above the tropopause, therefore, pressure and density fall in exactly the same proportion**, and the density ratio and pressure ratio at any two levels in that band are identical. Below the tropopause they are not, because the falling temperature partly offsets the falling pressure — which is the quantitative point made earlier about engine thrust falling away faster once the tropopause is passed.

Horizontal pressure variation and why QNH is not a constant

Everything above concerns the vertical. Pressure also varies horizontally, and although the horizontal gradients are minute compared with the vertical ones — a few hectopascals over hundreds of miles, against 1 hPa per 27 ft vertically — they are what drives the wind and what makes altimetry an operational problem rather than a fixed calculation.

- A deep depression may have a central sea-level pressure below 960 hPa; a strong anticyclone may exceed 1040 hPa. The difference of 80 hPa is worth well over 2,000 ft of indicated altitude to an aircraft whose subscale setting is wrong by that amount.

- Surfaces of constant pressure — **isobaric surfaces** — are therefore not level. They slope downward towards low pressure, and an aircraft maintaining a constant indicated pressure altitude is following that slope rather than flying level. Flying towards a region of lower pressure while holding a constant flight level means descending, gradually, in true altitude.
- This is why the transition level moves with the pressure of the day, why QNH is re-issued regularly rather than published once, and why an altimeter setting has an area of validity as well as a value.



Isobaric surfaces slope down towards low pressure, so an aircraft at constant flight level gradually descends.

Deriving the pressure profile from the temperature profile

Everything needed to compute the standard atmosphere's pressure column has now been stated: the hydrostatic equation relates pressure change to density, and the equation of state relates density to pressure and temperature. Combining them eliminates the density and leaves a relationship between pressure and temperature alone, which is why the model needs only a temperature profile as its input. The result takes two forms, one for each kind of band in the model.

In a band with a constant lapse rate (such as the troposphere):

$$\frac{p}{p_b} = \left(\frac{T}{T_b}\right)^{\frac{-g}{RL}}$$

Where the subscript b denotes the base of the band and L is the lapse rate expressed as a rate of change of temperature with height (so $L = -0.0065$ K/m in the troposphere). With the standard values, the exponent evaluates to **5.256**.

In an isothermal band (such as the lower stratosphere), the temperature ratio is 1 and the expression above breaks down, so the exponential form is used instead:

$$\frac{p}{p_b} = e^{\frac{-g\,(h-h_b)}{R\,T_b}}$$

which is the scale-height relationship already introduced, written out.

Check the first form against the tropopause. At 11,000 m the temperature has fallen to 216.65 K from 288.15 K, so

$$\frac{T}{T_b} = \frac{216.65}{288.15} = 0.7519$$

and $0.7519^{\,5.256} = 0.2233$, so

$$p = 1013.25 \times 0.2233 = 226.3 \text{ hPa}$$

which is the tropopause pressure quoted throughout this note and printed in the table above. The whole of the standard pressure column is this one calculation repeated at each altitude.

The corresponding density relationship follows by applying the equation of state to both ends. Since $\rho \propto p/T$, the density ratio is the pressure ratio divided by the temperature ratio, so within a constant-lapse-rate band the exponent is simply reduced by one:

$\frac{\rho}{\rho_b} = \left(\frac{T}{T_b}\right)^{\frac{-g}{RL}-1}$, with the exponent evaluating to **4.256** in the troposphere.

Check it at the same point: $0.7519^{\,4.256} = 0.2971$, which is the density ratio at the tropopause given in the tables above — and it agrees exactly with the 0.223 / 0.752 division performed earlier by a completely different route. Two independent derivations arriving at the same number is the check worth making on any calculation of this kind.

What the exponents tell you at a glance. The pressure exponent is larger than the density exponent by exactly 1, which is another way of stating that pressure falls faster with altitude than density does inside the troposphere — the falling temperature holding the density up, as established earlier. Above the tropopause, where the temperature is constant, both fall at exactly the same exponential rate and the distinction disappears.

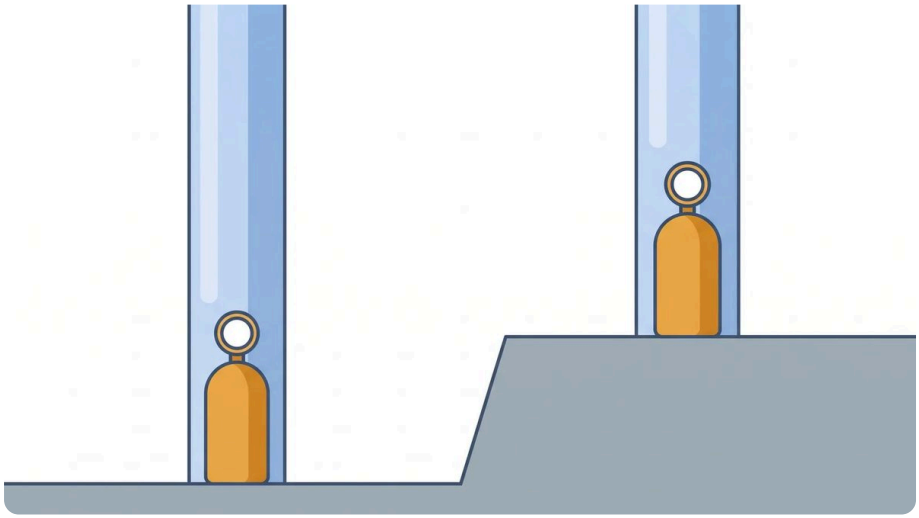
Measuring pressure in practice

The three datums described earlier — absolute, gauge and differential — correspond to three different instrument constructions, and a sensor built for one datum cannot simply be read as though it were another. The distinction is a physical one about what is on the reference side of the sensing element:

Instrument	What is on the reference side	Datum it reads	Where it is used
Liquid manometer (U-tube)	Open to atmosphere, or evacuated	Gauge, or absolute if the reference limb is evacuated	Low pressures and calibration work; the mercury barometer is the evacuated case
Bourdon tube	The atmosphere, acting on the outside of a curved flattened tube	**Gauge**	Hydraulic, pneumatic and oxygen system gauges — the classic workshop pressure gauge

Instrument	What is on the reference side	Datum it reads	Where it is used
Evacuated aneroid capsule	A sealed vacuum inside the capsule	**Absolute**	Altimeters, air data computers, manifold pressure gauges
Diaphragm or capsule with both sides connected	The second pressure to be compared	**Differential**	Airspeed indicators, cabin differential pressure indicators, filter and heat-exchanger pressure-drop warnings
Electrical transducer (piezoresistive, capacitive)	Whatever the construction provides — vacuum, vent or second port	Any of the three, by design	Modern air data and system sensing, where an electrical output is required

Two practical points follow. **An altimeter is an absolute-pressure instrument and an airspeed indicator is a differential one**, which is why they behave so differently when the static line fails: the altimeter loses its only input, while the airspeed indicator loses one of two and continues to display a difference that is now meaningless. And a **gauge instrument's reading depends on where it is**. The same accumulator charged to the same absolute pressure reads a few psi higher on a gauge at a high-elevation aerodrome than at sea level, purely because there is less atmosphere pressing on the reference side, which is the practical reason charging pressures are sometimes qualified by field elevation.



The same accumulator charge reads higher on a gauge at a high aerodrome, where less atmosphere presses on it.

Reading a specification correctly. "50 psi" on a servicing placard almost always means gauge, because that is what a workshop gauge reads. "50 psia" means absolute and is roughly 35 psi on the same gauge at sea level. "50 psid" means a difference between two named points and can be read from neither of the others without knowing both. Getting this wrong is a whole atmosphere out — 14.7 psi — which on a low-pressure system is a very large error indeed.

Density Variation with Altitude — Detailed

Air density is the mass of air per unit volume, measured in kg/m³. It is the **single most important atmospheric property for aerodynamics**, because both the lift equation and the drag equation contain density as a direct multiplier. If density halves, lift halves (for the same speed and angle of attack).

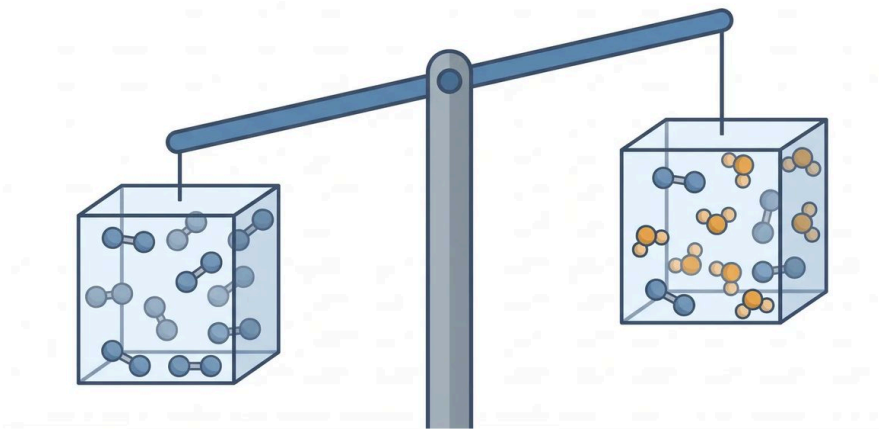
Density depends on both pressure and temperature, related by the **equation of state (ideal gas law)**:

$$\rho = \frac{P}{R \cdot T}$$

Where: ρ = air density (kg/m³), P = absolute pressure (Pa), R = specific gas constant for dry air (287.05 J/kg·K), T = absolute temperature (Kelvin)

From this equation, we can see three fundamental relationships:

Factor	Change	Effect on Density	Reason
Pressure	Decreases (higher altitude)	Density **decreases**	Fewer molecules per unit volume (P is in the numerator)
Temperature	Increases (hotter day)	Density **decreases**	Molecules move faster and spread apart (T is in the denominator)
Humidity	Increases (more moisture)	Density **decreases**	Water vapour molecules (mass 18) replace heavier N_2 (28) and O_2 (32) molecules



Water vapour (mass 18) displaces heavier nitrogen and oxygen, so humid air is less dense than dry air.

All three factors that reduce density — high altitude, high temperature, and high humidity — have the same effect on aircraft performance: **degraded performance**. Specifically:

- **Less lift** for a given airspeed and angle of attack → the aircraft must fly faster or at a higher angle of attack to maintain level flight
- **Less engine thrust/power** → less mass of air flows through the engine per second

- **Longer take-off and landing distances** → less acceleration from engines, less aerodynamic braking
- **Reduced rate of climb** → less excess thrust available for climbing
- **Lower service ceiling** → the altitude at which the maximum rate of climb falls to a specified small value (commonly 100 ft/min); the absolute ceiling, at which the aircraft can barely maintain level flight, is lowered with it

Which "speed" the lift statement refers to

The statement that halving the density halves the lift is exact, and the speed held constant in it is the **true** airspeed. That distinction is worth drawing out, because it is the reason an aircraft can climb into air of one third the sea-level density without the wing losing two thirds of its lift:

$$L = C_L \cdot \tfrac{1}{2}\rho V^2 \cdot S$$

Where C_L = lift coefficient (set by the angle of attack and the aerofoil), ρ = air density, V = **true** airspeed, S = wing area.

- At constant **true** airspeed and angle of attack, halving the density halves $\tfrac{1}{2}\rho V^2$ and therefore halves the lift. This is the statement above.
- At constant **indicated** airspeed and angle of attack, the lift does **not** change at all — because indicated airspeed *is* a measure of $\tfrac{1}{2}\rho V^2$, and holding it constant is precisely holding the dynamic pressure constant. The aircraft achieves this by flying faster in true terms as the density falls.

This is the whole reason aircraft are flown on indicated airspeed, as set out earlier in this note. It also means that in level flight at a fixed weight, an aircraft climbing at a constant indicated airspeed keeps very nearly the same angle of attack and the same lift the whole way up; what it is doing is converting the loss of density into an increase of true airspeed.

Ranking conditions for lift and for performance

Because all three density-reducing factors act in the same direction, the ranking of atmospheric conditions is unambiguous. The **best** conditions for lift, thrust and take-off performance are **cold, dry and low**; the **worst** are **hot, humid and high**. A question offering a set of conditions to rank is answered by working out which set gives the greatest density and which the least, and then reading the stem again to see which of the two it wants: best lift, best thrust and shortest take-off go with the greatest density, while the greatest true landing speed and the highest density altitude go with the least. Either way the ranking is done by taking the three factors in that order of importance:

Factor	Best case	Worst case	Relative strength of the effect
Temperature	Cold	Hot	Dominant — a 30 °C excess over standard is worth thousands of feet of density altitude
Pressure / elevation	High pressure, low elevation	Low pressure, high elevation	Strong — elevation is usually the largest single contributor at a mountain airfield

Factor	Best case	Worst case	Relative strength of the effect
Humidity	Dry	Humid (saturated)	Modest — roughly 700 ft of density altitude at 35 °C, and only when both hot and saturated

Applying the same reasoning to the landing case gives a result students often find counter-intuitive. The approach and touchdown are flown at a fixed *indicated* airspeed, so the *true* speed at touchdown is greatest when the density is lowest — that is, on a hot, humid day at a high elevation. A low-density day therefore raises the true touchdown speed, lengthens the landing roll, and increases the kinetic energy the brakes must absorb, which goes with the square of that speed.

Stalling: an angle in the air, a speed on the dial

One consequence of the lift equation is worth stating precisely because it is a standing source of confusion. The wing stalls at a particular **angle of attack**, and that angle does not change with altitude, temperature or density. What changes is the *speed* at which that angle is reached in level flight:

- The **indicated** stalling speed for a given weight and configuration is essentially the same at every altitude, because it corresponds to a fixed dynamic pressure.
- The **true** stalling speed rises steadily with altitude, in proportion to $1/\sqrt{\sigma}$. At 20,000 ft it is about 1.37 times its sea-level value; at 35,000 ft, about 1.80 times.

So an aircraft stalling at an indicated 120 kt at sea level still stalls at an indicated 120 kt at 35,000 ft — but it is then travelling through the air at about 216 kt true when it does so. This is one of the reasons high-altitude flight leaves so little margin: the true speeds at the low-speed and high-speed limits are converging even though the indicated speeds tell a much calmer story.

Ceilings: two definitions that are frequently swapped

An aircraft's ability to climb runs out gradually rather than abruptly, so two altitudes are defined and the distinction between them is examined directly:

Term	Definition	Rate of climb there	Why it is defined that way
Absolute ceiling	The altitude at which the aircraft can no longer climb at all — the excess thrust or power has fallen to zero	**Zero**	It is the true physical limit, but it is useless operationally: approaching it, the rate of climb becomes so small that the aircraft would take an unbounded time to reach it
Service ceiling	The altitude at which the maximum rate of climb has fallen to a specified small value, commonly **100 ft/min** (50 ft/min for some types)	**The specified small residual**	It is reachable in a sensible time and still leaves a usable climb capability, so it is the certificated and published figure

The service ceiling is therefore always the **lower** of the two, and it is the one quoted in the aircraft's documentation. The error to guard against is the statement that the service ceiling is where the aircraft can only just maintain level flight, or that the rate of climb there has fallen to zero — both of those describe the

absolute ceiling. If a definition mentions a rate of climb of zero, it is the absolute ceiling; if it names a small non-zero rate such as 100 ft/min, it is the service ceiling.

What moves a ceiling up or down

Both ceilings are density-limited, and anything that changes the density at a given pressure altitude moves them:

- **Colder air raises both ceilings.** Working the direction through: colder air at a given pressure altitude is *denser*; denser air means the engine ingests more mass and produces more thrust, while the drag at a given indicated airspeed is unchanged; so the excess thrust available for climbing is greater, and it does not fall to the limiting value until a higher altitude has been reached. A cold day is a high-ceiling day.
- **Hot air lowers both ceilings**, by the exact reverse of that argument.
- **Increased weight lowers both ceilings.** A heavier aircraft must fly at a higher lift coefficient or a higher speed, either of which increases drag, so the excess thrust runs out lower down.
- **High humidity lowers them slightly**, for the same density reason as everything else in this section.

The rate of climb falls off with altitude faster than a simple density argument suggests, because thrust falls with density while drag at the climb speed does not fall correspondingly. That widening gap is what closes the ceiling, and it is why the last few thousand feet of a climb take a disproportionate share of the time and fuel.

Density Altitude

Density altitude is a concept that combines the effects of altitude, temperature, and (to a lesser extent) humidity into a single, easy-to-use number. It is defined as:

The altitude in the ISA at which the actual air density would be found.

In other words, density altitude answers the question: "Given the current air density, what altitude in the ISA would have this same density?"

Example 5: An airport is at 2,000 ft elevation, and the QNH is standard, so its pressure altitude is also 2,000 ft. The day is hot at 35 °C. The ISA temperature at 2,000 ft is $15 - (2 \times 2) = 11\ ^{\circ}C$. The actual temperature is 24 °C warmer than ISA (ISA + 24). Each degree above ISA adds roughly 120 ft to the density altitude. So:

Density altitude ≈ 2,000 + (24 × 120) = 2,000 + 2,880 = **approximately 4,880 ft**

The aircraft sitting on a runway at 2,000 ft on this hot day will perform as if it were at nearly 5,000 ft in ISA conditions. Take-off distance will be significantly longer, climb rate will be reduced, and engine power will be less than sea-level ratings.

Maintenance relevance: When conducting engine ground runs, performance checks, or comparing measured engine parameters to manufacturer's reference values, **always correct for the current density altitude**. An engine producing apparently low thrust on a hot, high-altitude day may be performing perfectly normally once the reduced air density is accounted for. Conversely, accepting uncorrected values as satisfactory could mask a genuine engine deficiency.

Density altitude at sea level, and the myth of the mountain airfield

High density altitude is often assumed to be a problem confined to high-elevation aerodromes. It is not. Because temperature is the dominant term, a sea-level runway on a hot day can present a density altitude of several thousand feet, and an aircraft departing from it performs accordingly:

Worked example — a hot day at a sea-level airfield. Elevation 0 ft, QNH 1013 hPa so pressure altitude is also 0 ft, and the outside air temperature is +35 °C.

Standard temperature at sea level is +15 °C, so the deviation is ISA + 20.

Density altitude $= 0 + (20 \times 120) = 2{,}400$ ft.

The aircraft is at sea level and performing as though it were at 2,400 ft. A full calculation from the equation of state gives about 2,275 ft, so the rule again errs slightly on the pessimistic side. If the air is also saturated, add roughly a further 700 ft.

Worked example — the favourable case. A winter departure: pressure altitude 3,000 ft, outside air temperature –10 °C.

Standard temperature at 3,000 ft: $15 - (2 \times 3) = +9\ ^{\circ}\mathrm{C}$.

Deviation: $-10 - 9 = -19\ ^{\circ}\mathrm{C}$, that is ISA – 19.

Density altitude $= 3000 - (19 \times 120) = 3000 - 2280 = 720$ ft.

Direction check: the air is much colder than standard, so it is denser than standard, so the aircraft should perform *better* than its physical elevation suggests — and a density altitude of 720 ft against an elevation of 3,000 ft says exactly that. Density altitude can fall below zero on a very cold day at a low-elevation aerodrome, which is a perfectly meaningful result: the air is denser than standard sea-level air.

How accurate is the 120 ft per degree rule?

The rule replaces a curve with a straight line, so it is exact nowhere and adequate almost everywhere. Compared with a full calculation from the equation of state it over-estimates the density altitude by a few per cent in the cases worked in this note — roughly 130 ft in the sea-level example and about 350 ft in a 6,000 ft, ISA+35 case. That bias is in the conservative direction: the rule tells the operator the air is thinner than it really is, so a take-off calculation based on it has a small margin built in.

Two limitations should nevertheless be understood. The rule assumes the constant is 120 ft per degree in every condition, whereas the true sensitivity is about 118.6 ft per degree at standard temperature — essentially the same figure at every altitude in the troposphere — and falls further as the deviation grows, which is why the over-estimate above is driven by the size of the deviation rather than by height; and it takes no account of humidity, which always adds to the density altitude and never subtracts from it. Where the margins are thin — a short runway, a high elevation, high temperature and an obstacle in the departure path — the aircraft's own performance charts, which are computed rather than approximated, are the only acceptable source.

Where density altitude appears in maintenance work

Beyond the engine ground-run case above, the same quantity governs several checks an engineer carries out or witnesses:

- **Auxiliary power unit starting and load limits** are altitude- and temperature-dependent for exactly the same mass-flow reason as the main engines, and an APU that is slow to start or unable to carry full pneumatic load on a hot, high day may be entirely serviceable.
- **Cabin air-conditioning performance** on the ground is limited by the mass of air the packs can move and by the temperature of the air being conditioned. Complaints of poor cooling concentrate on exactly the days when the density altitude is highest.
- **Propeller and rotor performance checks** are density-sensitive, since both are wings and both produce thrust in proportion to density. Static rpm and torque checks are quoted against ambient conditions for that reason.
- **Brake and tyre condition after a hot-and-high landing** reflects the higher true touchdown speed and therefore the greater kinetic energy absorbed, which is why brake cooling schedules are expressed against energy rather than against a simple speed.

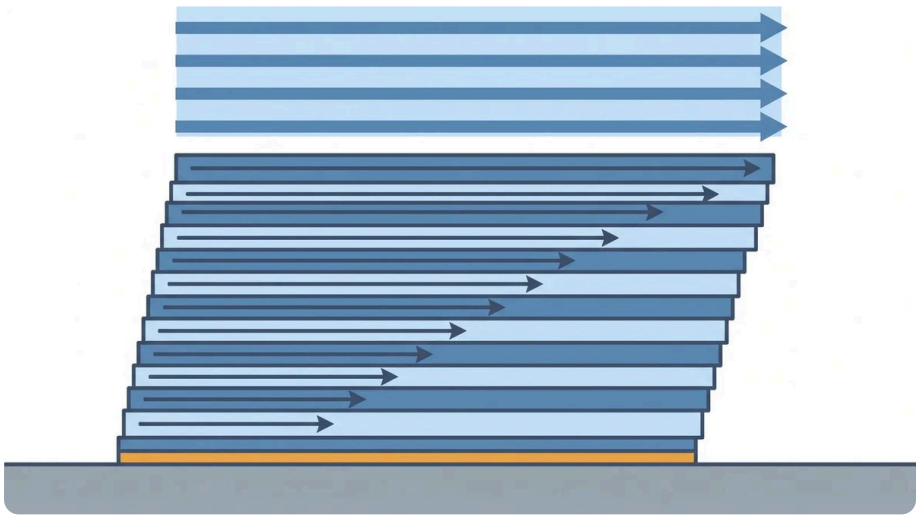
Other Physical Properties of Air

Pressure, temperature and density are what the standard atmosphere tabulates, together with the three ratios and the local speed of sound; humidity is not among them, because the model takes the air as dry and treats moisture as a separate correction. Those quantities are not the only properties of air that an aircraft responds to. Three more are worth setting out, because each has a direct aerodynamic or operational consequence and each varies through the atmosphere in a way that is predictable from what has already been established.

Viscosity: the property that makes drag possible

Viscosity is a fluid's internal resistance to shearing — its reluctance to have one layer slide over the next. Air is far less viscous than oil or water, but it is emphatically not zero, and the consequences of its small viscosity are large:

- Air in contact with a solid surface does not slip along it; it is brought to rest relative to the surface. That single fact is the origin of skin-friction drag and of the thin sheared layer next to every surface of the aircraft.
- An entirely inviscid fluid would produce *no drag at all* on a body moving steadily through it. The whole of profile drag exists because air has viscosity.



Air touching the skin is brought to rest, so a thin sheared layer forms above every surface.

The behaviour of air's viscosity with temperature is opposite to the everyday intuition formed from liquids. Warming a liquid makes it thinner and less viscous, because the molecules are pulled apart and the intermolecular attraction that resists shear is weakened. In a *gas* there is very little intermolecular attraction to begin with, and the resistance to shear comes instead from molecules migrating between layers and carrying their momentum with them. Heating the gas makes the molecules move faster, so more momentum is exchanged, and the viscosity **rises**.

The direction, stated explicitly: the viscosity of air *increases* as its temperature increases, and it is essentially independent of pressure over the range an aircraft meets. Since temperature falls with altitude in the troposphere, air aloft is *less* viscous than air at sea level — but the density has fallen far more than the viscosity has, and it is the ratio of the two, not the viscosity alone, that governs how the flow behaves around the aircraft.

Compressibility: where the air stops behaving like a liquid

Compressibility is a fluid's willingness to change its density when the pressure acting on it changes. Liquids are, for practical purposes, incompressible — which is why hydraulic systems work at all — whereas gases are highly compressible, as every gas law in this note has assumed.

At low speeds the pressure changes an aircraft produces in the air around it are small compared with the ambient pressure, so the resulting density changes are small enough to ignore. The flow can then be treated as though the air were incompressible, which simplifies the aerodynamics enormously. As speed rises the pressure disturbances grow with the square of speed, and at some point the density changes stop being negligible. The conventional threshold is **Mach 0.3**, above which compressibility effects begin to require explicit treatment.

Two connections to earlier material are worth drawing. First, that threshold is a *Mach number*, not an airspeed, because what matters is the speed of the aircraft relative to the speed at which pressure disturbances propagate — and that propagation speed is the local speed of sound, set by the temperature alone. Second, this is the reason the airspeed chain contains a compressibility correction between calibrated and equivalent airspeed: the pitot tube brings air to rest, compressing it slightly, so it senses marginally more than $\frac{1}{2}\rho V^2$. The

correction is always subtractive and it grows with both speed and altitude, since a given true airspeed is a higher Mach number in the colder air aloft.

Thermal properties: specific heat and thermal conductivity

Two thermal properties of air underpin results already used above:

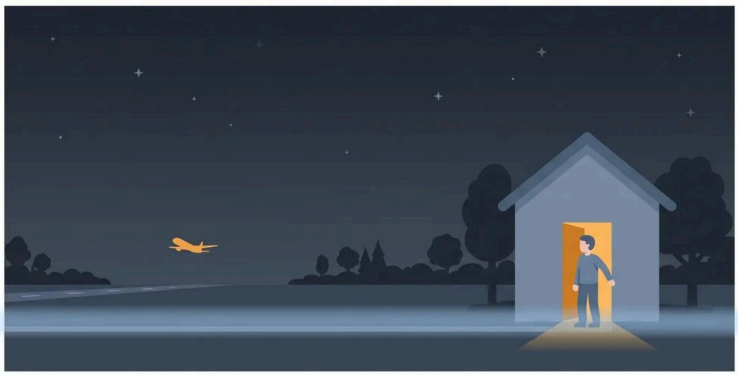
- **Specific heat capacity** is the energy required to raise the temperature of unit mass by one degree. For air it is small — roughly a quarter of that of water — which is why air heats and cools quickly and why a modest amount of energy produces a large temperature change in it. It is also the quantity that fixes the dry adiabatic lapse rate: that rate is gravitational acceleration divided by the specific heat of air at constant pressure, which is why it is a physical constant rather than a measurement. Note that air has two specific heats, one at constant pressure and one at constant volume, and their ratio is the $\gamma = 1.4$ that appears in the speed-of-sound expression.
- **Thermal conductivity** is very low. Still air is one of the better thermal insulators available, and this is why conduction plays almost no part in distributing heat through the atmosphere and why convection has to do the work instead. The same property is exploited deliberately in aircraft construction, where trapped air in cabin insulation blankets is doing most of the insulating.

The specific heat of air is very nearly constant over the range an aircraft meets, and the standard atmosphere treats it as a fixed constant rather than tabulating it. The thermal conductivity is a different case: it falls about 23 % between sea level and the tropopause, much as the viscosity falls about 21 % — both small movements beside the 70 % fall in density over the same interval, but large enough that the full standard-atmosphere documents do tabulate them. The US Standard Atmosphere gives dynamic viscosity, kinematic viscosity and thermal conductivity against altitude in a table of their own; the abridged tables used in flight and maintenance work carry none of the three: the standard-atmosphere table earlier in this note stops at temperature, pressure, density, the density ratio and the speed of sound. For this syllabus both thermal properties can be treated as constants, and the error in doing so is far smaller than the day-to-day variation of the atmosphere itself.

How the atmosphere carries sound and radio waves

Two forms of wave propagate through the atmosphere, and each is affected by properties that vary with height: temperature in the case of sound, pressure and humidity in the case of radio.

Sound travels at a speed set by the absolute temperature alone, as established above, and it therefore *refracts* wherever the temperature varies with height — bending towards the region where it travels more slowly, which is the colder air. On a normal day, with temperature falling with height, sound leaving the ground bends *upward* away from the surface, and an observer at a distance ends up in an acoustic shadow. Under a night-time radiation inversion the profile is reversed: the coldest air is at the surface, so sound bends *downward* and is returned to the ground instead of escaping. This is the physical reason aircraft and other distant noise carries so much further on a still, clear night, and it is a nuisance-complaint mechanism rather than an aircraft fault.



On a still, clear night the coldest air lies at the ground, so distant aircraft noise carries much further.

Radio waves are affected by the atmosphere in three distinct regions, and the effect is different in each:

- **In the troposphere**, the refractive index falls with height because pressure and humidity fall, so radio waves are bent gently downward. This extends the effective line-of-sight range of VHF communications and of radar somewhat beyond the geometric horizon. The convention for handling it is to leave the rays straight and enlarge the earth instead, working with an **effective earth radius** four-thirds of the true one; the four-thirds belongs to the radius, not to the range. Horizon distance goes as the square root of the radius, so the range grows by a factor of $\sqrt{4/3} = 1.155$, about 15 % — the same proportion at any height, because that factor carries no height term. Where a strong inversion or a sharp humidity gradient exists, the bending can become strong enough to trap signals in a **duct**, carrying them far beyond normal range while leaving aircraft just outside the duct unreachable. Anomalous VHF or radar performance in settled anticyclonic weather is very often ducting rather than an equipment fault.
- **Water in the atmosphere attenuates and reflects.** Rain absorbs and scatters at the higher frequencies, which is why heavy precipitation degrades satellite and radar links, and why the effect worsens as frequency rises. It is also the basis of weather radar, which deliberately uses a wavelength that precipitation-sized droplets return strongly — and that choice is exactly why it does not return anything useful from dry phenomena such as clear air turbulence, volcanic ash or dry hail.
- **In the ionosphere**, as described earlier, free electrons refract high-frequency signals back towards the earth and delay satellite navigation signals.

The pattern behind all three cases. A wave — sound or radio — bends towards the region in which it travels more slowly. For sound that region is colder air; for radio it is denser, moister air. Once that single rule is in hand, the direction of every one of these effects can be reconstructed rather than memorised, and the atmosphere's optical, acoustic and radio behaviour all become consequences of the same vertical profiles the rest of this note has established.

The Atmosphere as an Operating Environment

Everything above has treated the atmosphere as something the aircraft flies *through*. It is also something the aircraft lives *in*, and its properties act on the structure, the systems and the fluids continuously. This section collects the consequences that fall to the maintenance engineer rather than to the crew, and every one of them follows from a property already established.

Cold soak and the thermal cycle

An airframe cruising for several hours at –55 °C reaches thermal equilibrium with the air around it. Structure, fuel, hydraulic fluid, water and every component not deliberately heated arrives at approximately that temperature, and this **cold soak** persists for a considerable time after landing because the thermal mass is large and air is a poor conductor. Several familiar effects follow:

- **Frost and condensation on arrival.** A cold-soaked wing skin parked in warm humid air is frequently below the dew point of that air, so moisture condenses on it, and if the skin is below freezing it deposits as frost. Cold-soaked fuel in a wing tank can hold the upper skin below freezing for hours in ambient temperatures well above it, which is why frost can appear on an upper wing surface on a mild morning.
- **Water in fuel.** Air drawn into tanks through the vent system as fuel is consumed carries water vapour with it. Cooling that air below its dew point condenses the water, which then settles beneath the fuel; at cold-soak temperatures it can freeze. This is the physical reason for water drain checks and for the use of anti-icing additives, and it is a direct application of the saturation behaviour described earlier.
- **Dimensional change and seal behaviour.** Materials contract on cooling and do so by different amounts, so clearances change and elastomeric seals stiffen. Leaks that appear only after a long high-altitude sector, and disappear on the ground, are usually thermal.
- **Pressure change in sealed volumes.** As established in the section on the gas laws, a fixed mass of gas in a fixed volume loses pressure in proportion to its absolute temperature. Tyres, accumulators, oxygen bottles and fire-extinguisher bottles all read lower after a cold soak, and the correct comparison is against a pressure/temperature chart rather than against a single figure.

Pressurisation as a structural load

The cabin differential pressure introduced earlier is not a static curiosity. It is applied on every climb, held for the cruise and released on every descent, so the fuselage is a pressure vessel loaded and unloaded once per flight. That makes the number of flights, not the number of hours, the governing count for pressurised-structure fatigue, and it is why inspection programmes for fuselage skins, lap joints, window and door surrounds and frames are driven by cycles.

Two related points follow from the pressure profile. The differential is largest when the aircraft is highest, because the ambient pressure outside is lowest — so the structural load case is the high-altitude cruise, not the take-off. And a short-sector aircraft climbing to a modest level several times a day accumulates cycles far faster than a long-haul aircraft flying one sector in the same period, even though the long-haul aircraft spends more hours at a higher differential.

Humidity, condensation and corrosion

The relationship between temperature, dew point and relative humidity set out earlier explains most of the moisture an engineer finds inside an aircraft. Warm humid air enters the cabin and the unpressurised bays on the ground; the aircraft climbs; the structure cools far below the dew point of that air; and the water condenses on the coldest surfaces available, which are the skin, the frames and the stringers behind the cabin lining. On descent the process reverses only partially, and the residue accumulates.

The consequences are the familiar ones — corrosion in the lower fuselage, contamination of insulation blankets, and moisture in electrical connectors — and the countermeasures all address the same physical chain: drainage paths and drain valves at the lowest points, moisture barriers, and protective treatments where condensation is unavoidable. It is worth recognising that this is not a leak and not a fault: it is the inevitable result of taking humid air to an altitude where the airframe is far colder than that air's dew point.

Ozone, ultraviolet and the atmosphere's chemistry

Two atmospheric constituents attack materials directly, and both become more aggressive with altitude:

- **Ozone**, concentrated in the stratosphere as described earlier, attacks natural and some synthetic rubbers, causing surface cracking in elastomers under tension. This is one reason ozone-resistant compounds are specified for external seals and why bleed air on aircraft operating at high stratospheric levels is passed through a converter.
- **Ultraviolet radiation** is more intense at altitude because there is less atmosphere above to absorb it. It degrades paints, composite matrix resins, transparencies and any exposed polymer, which is why protective coatings on composite structures are functional rather than cosmetic and why their condition is an airworthiness matter.

Air as a working fluid in the aircraft's own systems

Finally, the atmosphere is not only outside the aircraft; it is drawn in and used. Every property described in this note follows it inside:

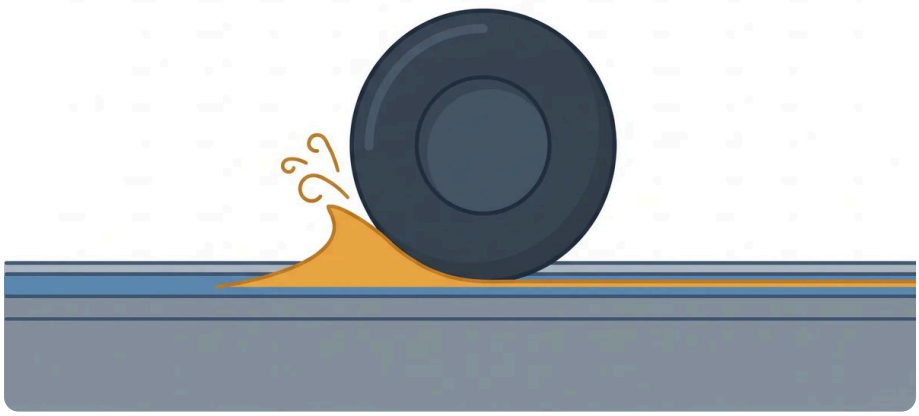
- **The mass of bleed air available falls with altitude**, for the same reason engine thrust does, so pressurisation and air-conditioning performance is at its most marginal exactly where it is most needed.
- **Compressing air heats it and expanding it cools it**, adiabatically and by exactly the mechanism described in the section on lapse rates. This is what an air-cycle machine exploits: bleed air is cooled in a heat exchanger, then expanded through a turbine, which extracts work and drops its temperature far below ambient with no refrigerant involved.
- **The water in that air condenses when it is cooled**, which is why an air-cycle system needs a water separator, and why the moisture it removes is a function of the humidity of the air the aircraft is standing in rather than of anything the system is doing.
- **Ram air cooling depends on density**, so a ram-air heat exchanger sized for a hot day at a high-elevation airfield is being sized for the case where both the mass flow through it and the temperature difference driving it are least favourable.

Atmospheric water on the runway

The last place the atmosphere acts on the aircraft is the one it is standing on. Precipitation that has fallen out of the atmosphere changes the runway from a dry friction surface into a contaminated one, and the physics of what then happens is worth understanding because the contaminants behave quite differently from one another.

Contaminant	Principal effect	Why
Standing water	Reduced braking friction and increased rolling resistance; risk of hydroplaning	The tyre must displace the water before it can contact the surface, and above a certain speed it cannot
Slush and wet snow	Substantial drag on the take-off run, plus impingement damage risk	Dense and heavy; it is thrown up into wheel wells, flaps and engine intakes, where it can freeze
Dry snow	Reduced friction, less displacement drag than slush	Low density, so less mass to push aside
Ice	Severely reduced friction, particularly with any water film on top of it	Very low coefficient of friction, and a melt film reduces it further

Hydroplaning (aquaplaning) occurs when the water cannot be displaced from beneath the tyre quickly enough and a wedge of water lifts the tyre clear of the surface. The tyre then rides on a water film with almost no friction available, so braking and directional control are lost together. Three forms are recognised: *dynamic*, on standing water, which is the case the classical speed relationship addresses; *viscous*, on a very thin film over a smooth or rubber-contaminated surface, which can occur at much lower speeds; and *reverted rubber*, where a locked wheel skidding on a wet surface generates enough heat to boil the water and revert the tyre's rubber, sustaining the skid.



Above a certain speed a wedge of water lifts the tyre clear of the runway, leaving it riding on a water film.

Classical dynamic hydroplaning speed (a widely used approximation):

$V_p = 9\sqrt{p}$ for a rotating tyre, and about $7.7\sqrt{p}$ for a non-rotating one

Where V_p is in knots and p is the tyre inflation pressure in psi.

Worked example. A tyre inflated to 150 psi:

$V_p = 9 \times \sqrt{150} = 9 \times 12.25 = 110$ kt for a rotating tyre.

What the relationship says. A higher tyre pressure raises the hydroplaning speed, because a higher pressure means a smaller contact patch carrying the same load and therefore a greater pressure available to push the water aside. That is one reason an under-inflated tyre is an operational hazard on a wet runway and not merely a wear problem — and it connects directly to the gas-law material earlier in this note, because a tyre that reads low after a cold soak and is judged serviceable without a temperature correction may be genuinely under-inflated once warm.

The non-rotating figure is the lower of the two, at $7.7 \times 12.25 = 94$ kt, which matters on landing: a wheel that has not yet spun up on touchdown is more vulnerable than one already rotating.

Runway contamination also interacts with the density effects described throughout this note, and the interaction is unfavourable in both directions. A cold, wet day gives good air density but poor runway friction; a hot, humid day gives poor density and, if there has been a shower, poor friction as well. The take-off performance calculation must account for both, and the two are assessed separately because they act on different parts of the take-off: density on the acceleration and the lift-off speed, contamination on the displacement drag during the run and on the stopping distance in the event of a rejected take-off.

Oxygen, combustion and the atmosphere

The atmosphere is the oxidiser for every fire an aircraft can suffer as well as for every engine it carries, and the properties already established govern both. A fire requires three things simultaneously — a **fuel**, an **oxidiser** and an **ignition source** — and removing any one of them prevents or extinguishes it. Since the fuel is generally what the aircraft is carrying and the ignition source is generally what has already happened, the oxidiser is the term most often attacked, and it is an atmospheric quantity.

Two further points sharpen the picture, and both come straight from earlier sections:

- **It is the oxygen *concentration* that governs whether a mixture will burn, not the total pressure alone.** As established in the section on composition, the proportion of oxygen is constant with altitude while the partial pressure falls. So a given volume of air aloft contains fewer oxygen molecules, which slows combustion, but the mixture is still 21 % oxygen and still capable of supporting it.
- **A fuel burns only within a range of mixture strengths.** Too little vapour and the mixture is too lean to propagate a flame; too much and it is too rich. Between the two lie the *flammability limits*, and those limits shift with temperature and with pressure — which means the flammability of a fuel tank ullage depends on the altitude and the fuel temperature, not on the fuel alone.

Making the atmosphere non-flammable: inerting

The nitrogen described in the section on composition as an inert dilutant is exactly what makes the atmosphere as safe as it is: at 78 % it is the majority constituent, and it takes no part in combustion. Fuel tank inerting takes that principle further by deliberately raising the nitrogen proportion in the space above the fuel. Reducing the oxygen concentration in the ullage to roughly 12 % by volume or below renders the vapour incapable of supporting combustion whatever the mixture strength — the mixture is outside its flammability limits at every fuel-vapour concentration, so no ignition source can start a fire.

The nitrogen is generated on board rather than carried, and the method is an elegant use of the atmosphere's composition. Bleed air, which is ordinary atmospheric air compressed and therefore of the same 78 / 21 makeup, is passed through an air separation module containing bundles of hollow fibres. The fibre wall is permeable to oxygen, water vapour and carbon dioxide but much less so to nitrogen, so oxygen diffuses out through the walls while nitrogen passes along the bore. What emerges is **nitrogen-enriched air**, which is delivered to the tank ullage, and *oxygen-enriched* air, which is dumped overboard. Nothing is consumed; the system simply separates the atmosphere it is given.

> **Why this is an atmospheric physics topic and not just a systems one.** The separation depends on the partial pressures across the membrane, so the performance of an inerting system varies with the bleed pressure and temperature available — that is, with altitude and with engine power. The system is at its most effective in the cruise, where the bleed pressure ratio is high and the ullage is being emptied by fuel consumption, and least effective on the ground. That matches the risk profile reasonably well, but it is a consequence of the atmosphere rather than a design choice.

Flash point, autoignition and why the numbers are quoted with conditions

Two temperatures characterise a liquid fuel's fire behaviour, and both are properties of the fuel *and its atmosphere* together:

- **Flash point** is the lowest temperature at which the liquid gives off enough vapour to form an ignitable mixture with the air above it. Below the flash point there is simply not enough vapour present to reach the lean flammability limit. Aviation kerosene has a comparatively high flash point — typically at least about 38 °C — which is one of the principal safety advantages it holds over aviation gasoline, whose flash point is far below any ambient temperature it will encounter.
- **Autoignition temperature** is the temperature at which the mixture will ignite spontaneously, with no spark or flame required. It is the relevant figure for a fuel or hydraulic leak onto a hot surface, and it is why bleed ducts, engine cases and exhaust components are shrouded, drained and routed away from flammable fluid lines rather than merely kept clear of electrical sources.

Both figures are quoted at standard atmospheric pressure, and both change with it. Lowering the pressure reduces the partial pressure of the fuel vapour as well as of the oxygen, and the net effect is that a tank ullage which is safely lean on the ground can move into the flammable range as the aircraft climbs. This is the physical reason inerting is required at all: the hazard is not constant through the flight, and the altitude at which it is greatest is not the one where the aircraft is easiest to reach.

Combined Worked Problems

Each problem below deliberately requires several of the relationships in this note at once, in the way a real calculation does. Work each one through before reading the solution.

Problem 1 — a cruise condition. An aircraft is cruising at FL390 at Mach 0.82. The total air temperature has been corrected and the static air temperature is –60 °C. Find the ISA deviation, the local speed of sound, the true airspeed, the air density and the equivalent airspeed.

ISA deviation. FL390 is above the tropopause, so the standard temperature is –56.5 °C, not a lapse-rate calculation. $-60 - (-56.5) = -3.5\ ^\circ\mathrm{C}$, that is ISA – 3.5.

Speed of sound. It depends on the actual temperature only, so use the actual –60 °C, which is 213.15 K:

$$a = \sqrt{1.4 \times 287.05 \times 213.15} = \sqrt{85660} = 292.7\ \mathrm{m/s} = 569\ \mathrm{kt}$$

True airspeed. $\mathrm{TAS} = M \times a = 0.82 \times 569 = 466\ \mathrm{kt}$.

Density. A flight level is a pressure altitude, so the ambient pressure is the standard value at 39,000 ft, about 197 hPa or 19,700 Pa. Use the actual temperature, not the standard one:

$$\rho = \frac{p}{RT} = \frac{19700}{287.05 \times 213.15} = \frac{19700}{61180} = 0.322\ \mathrm{kg/m^3}$$

Equivalent airspeed. $\sigma = 0.322/1.225 = 0.263$, so $\sqrt{\sigma} = 0.512$ and

$$\mathrm{EAS} = \mathrm{TAS} \times \sqrt{\sigma} = 466 \times 0.512 = 239\ \mathrm{kt}$$

What the answer says. The crew see a little under 240 kt on the airspeed indicator while moving through the air at 466 kt — almost double. The pressure came from the flight level and the temperature came from the probe; mixing them up, by using the standard temperature to find the speed of sound or the actual temperature to find the pressure, is the classic error in a problem of this shape.

Problem 2 — a hot-and-high departure. An aerodrome lies at 6,000 ft elevation. The QNH is 1013 hPa and the outside air temperature is +38 °C. Estimate the density altitude, the density ratio, and the effect on thrust and on the take-off run compared with a sea-level standard day.

Pressure altitude. The QNH is standard, so the pressure altitude equals the elevation: 6,000 ft.

ISA deviation. Standard temperature at 6,000 ft is $15 - (2 \times 6) = +3\ ^\circ\mathrm{C}$, so the deviation is $38 - 3 = +35\ ^\circ\mathrm{C}$.

Density altitude. $6000 + (35 \times 120) = 6000 + 4200 = 10{,}200\ \mathrm{ft}$. A full calculation gives about 9,840 ft.

Density ratio. The standard pressure at 6,000 ft is about 812 hPa; with the actual temperature of 311.15 K,

$$\rho = \frac{81200}{287.05 \times 311.15} = 0.909\ \mathrm{kg/m^3}, \text{ so } \sigma = 0.909/1.225 = 0.742$$

Consequences. Thrust falls roughly in proportion to σ, so about **26 % of the sea-level thrust is lost**. The true lift-off speed rises by $1/\sqrt{\sigma} = 1.16$, so it is about **16 % higher** than at sea level. Combining the two as $1/\sigma^2$ gives a ground roll roughly **80 % longer** than the sea-level standard-day figure — and the aircraft's own charts, which include effects this estimate omits, will normally show worse.

Note what dominated. Of the 4,200 ft of density altitude added to the elevation, every foot came from temperature; the elevation contributed its own 6,000 ft. Neither term is negligible, and this is the compounding case in numbers.

Problem 3 — the pressurised cabin. The same aircraft from Problem 1 is at FL390 with a cabin altitude of 6,500 ft. Find the cabin differential pressure, and compare the oxygen partial pressure inside the cabin with that outside and at sea level.

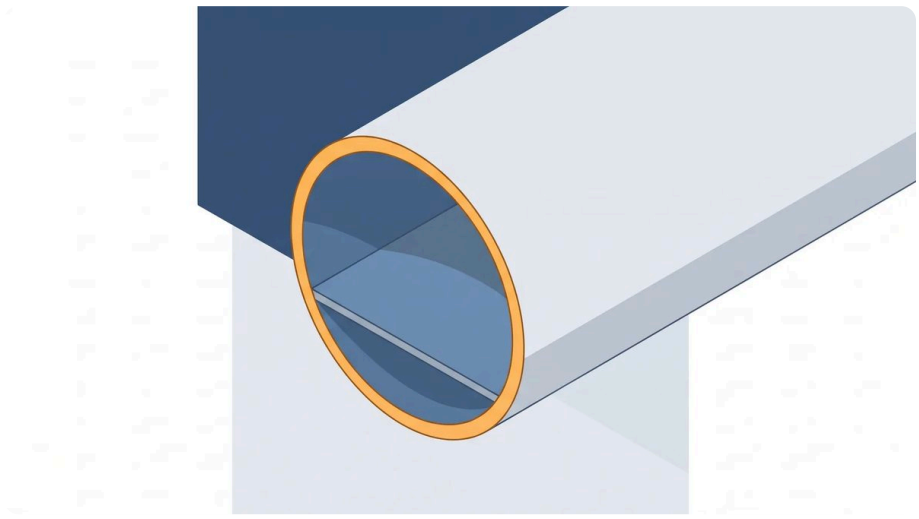
Pressures. Both altitudes are pressure altitudes, so both come from the standard tables: 197 hPa at 39,000 ft outside, and about 797 hPa at 6,500 ft inside.

Differential. $797 - 197 = 600$ hPa. Converting, $600 \div 68.95 = 8.7$ psi — the load the fuselage is carrying in tension over its entire surface, every flight.

Oxygen partial pressures. The composition is unchanged everywhere, so each is 20.95 % of the total:

- Sea level: $0.2095 \times 1013 = 212$ hPa — the reference.
- In the cabin: $0.2095 \times 797 = 167$ hPa, which is 79 % of the sea-level value.
- Outside: $0.2095 \times 197 = 41$ hPa, which is 19 % of the sea-level value.

What the answer says. The 600 hPa the pressurisation system is holding is the difference between a comfortable environment and one in which useful consciousness would last seconds. Note also that the oxygen *percentage* is 20.95 % in all three places — inside the cabin, outside the aircraft and at sea level. It is only the partial pressure that has changed, which is the point made at the start of this note and the reason pressurisation, rather than oxygen enrichment, is the solution adopted for the cabin.



At FL390 the cabin holds 797 hPa against 197 hPa outside, a 600 hPa load carried by the whole fuselage skin.

Problem 4 — a cold-weather approach. An aerodrome is at 1,500 ft elevation with a QNH of 1000 hPa and an aerodrome temperature of –25 °C. An approach procedure specifies a platform altitude of 5,000 ft. Find the pressure altitude of the aerodrome, the ISA deviation, and the approximate difference between the aircraft's indicated and true altitude at the platform.

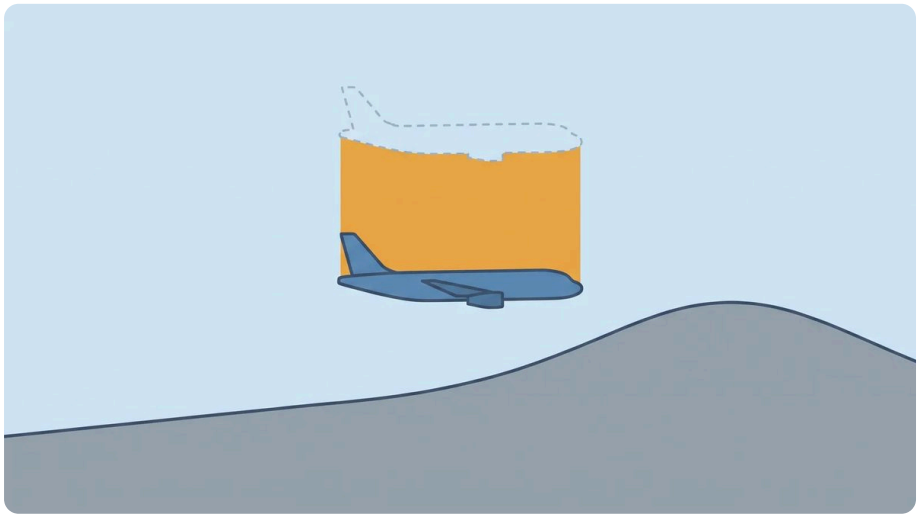
Pressure altitude of the aerodrome: $1500 + [(1013 - 1000) \times 27] = 1500 + 351 = 1{,}851$ ft.

ISA deviation: standard temperature at 1,851 ft is $15 - (2 \times 1.851) = +11.3\ ^\circ\mathrm{C}$, so the deviation is $-25 - 11.3 = -36.3\ ^\circ\mathrm{C}$ — about ISA – 36.

Altimeter error at the platform. The height above the QNH datum is 5,000 ft. At roughly 4 % per 10 °C of deviation, 36 °C gives $0.04 \times 3.6 = 14.4\,\%$, so

$5000 \times 0.144 = 720$ ft of over-reading.

What the answer says. The altimeter shows 5,000 ft while the aircraft is at roughly 4,280 ft true — it is more than 700 ft *lower* than the instrument claims, and the terrain does not care what the instrument claims. The direction is the dangerous one, the magnitude is comparable with an entire obstacle clearance margin, and this is the case for which mandatory cold-temperature corrections exist. Notice also how the deviation was found: against the standard temperature for the aerodrome's *pressure* altitude of 1,851 ft, not its elevation of 1,500 ft.



In very cold air the altimeter over-reads, so the aircraft is really over 700 ft lower than it shows.

Problem 5 — humidity, density and a piston aircraft. A light aircraft departs a sea-level airfield. Case A: QNH 1013 hPa, temperature +10 °C, dry. Case B: QNH 1003 hPa, temperature +32 °C, saturated. Compare the density altitudes and comment on what the naturally aspirated engine will do.

Case A. Pressure altitude $= 0 + [(1013 - 1013) \times 27] = 0$ ft. ISA temperature at sea level is +15 °C, so the deviation is $10 - 15 = -5\ ^\circ\mathrm{C}$. Density altitude $= 0 - (5 \times 120) = -600$ ft — below sea level, meaning the air is denser than standard.

Case B. Pressure altitude $= 0 + [(1013 - 1003) \times 27] = 270$ ft. ISA temperature at 270 ft is about +14.5 °C, so the deviation is about $32 - 14.5 = +17.5\ ^\circ\mathrm{C}$. Density altitude $= 270 + (17.5 \times 120) = 270 + 2100 = 2{,}370$ ft, and saturation at that temperature adds roughly a further 600 ft, giving about **3,000 ft**.

Comparison. The two departures are from the same runway on the same day of the year, and the density altitude differs by about 3,500 ft. The engine, whose power falls with density, will deliver noticeably less in Case B; the propeller, which is a wing and therefore also density-dependent, will produce less thrust for the same power; and the true lift-off speed will be higher. Note the ranking of the three contributions: the 22 °C of temperature difference is worth 2,700 ft, the 10 hPa of pressure difference 270 ft, and the humidity about 600 ft. Temperature dominates, as it does everywhere in this note.

Problem 6 — a gas-law check on a servicing figure. A tyre is inflated in a hangar at +20 °C to 180 psi gauge. The aircraft is towed out and left overnight, and the tyre temperature falls to –15 °C. Ignoring any change of volume, what should the gauge read in the morning, and is a reading of 155 psi evidence of a leak?

Convert to absolutes. Pressure: $180 + 14.7 = 194.7$ psia. Temperatures: $20 + 273.15 = 293.15$ K and $-15 + 273.15 = 258.15$ K.

Apply the pressure law at constant volume:

$$p_2 = 194.7 \times \frac{258.15}{293.15} = 194.7 \times 0.8806 = 171.4 \text{ psia}$$

which is $171.4 - 14.7 = 156.7$ psi on the gauge.

Conclusion. A reading of about 157 psi is exactly what the physics predicts, so 155 psi is within the noise of the measurement and is *not* evidence of a leak. Deflating the tyre back to some remembered number, or topping it up to 180 psi cold, would leave it substantially over-inflated once it warmed up again. This is why tyre pressures are checked against the temperature and why the check is specified at a stated condition rather than as a single figure.

Module 9 — Human Factors

9.1 — General

Human Factors is the study of how humans interact with their working environment — the equipment, procedures, other people, and the physical surroundings. In aviation maintenance, understanding human factors is not optional; it is a regulatory requirement and a matter of life and death. Research consistently shows that **approximately 80% of aviation accidents involve human factors**. This module equips maintenance engineers with the knowledge to recognise, understand, and mitigate the human contributions to error.

The Need to Take Human Factors into Account

Aircraft maintenance is performed by people — and people are fallible. No matter how skilled, experienced, or conscientious an engineer is, they will make errors. The goal of human factors training is not to eliminate human error (which is impossible) but to:

- **Understand** why errors occur — the underlying psychological, physiological, and organisational causes.
- **Recognise** the conditions that make errors more likely — fatigue, time pressure, poor communication, inadequate resources.
- **Design systems** that are error-tolerant — procedures, checklists, inspections, and organisational structures that catch errors before they cause harm.
- **Promote a safety culture** where reporting errors and hazards is encouraged, not punished.

What "Human Factors" Actually Means

The term is used loosely in the industry, and the loose use costs marks. Outside maintenance it is heard most often in connection with flight-deck design and Crew Resource Management, but that is one application of the subject rather than the subject itself. In the maintenance context the UK Civil Aviation Authority's CAP 715, the guidance document written to introduce this syllabus, defines human factors as the study of human capabilities and limitations in the workplace: the interaction between maintenance personnel, the equipment they use, the written and verbal procedures and rules they follow, and the environmental conditions of the system they work in, with the aim of optimising that relationship so as to improve safety, efficiency and well-being.

Three consequences follow from that definition, and each of them is a distinction the examination tests directly.

- **It concerns the person inside a working system, not the aircraft.** The strength of a lap joint, the fatigue life of a spar cap and the composition of an alloy belong to structures and materials. They become human factors questions only at the point where a person has to inspect, interpret, judge or repair them — which, in maintenance, is almost immediately.
- **It is not flight-deck human factors.** Analysing instrument layout and pilot response under abnormal conditions is a legitimate and closely related field, but it studies a different population doing a different job under different constraints. Maintenance human factors studies the engineer, the hangar, the manual, the shift and the handover.
- **It is a systems subject, not a character assessment.** The question it asks about an occurrence is not "who was careless" but "what about this task, this workplace and this organisation made that error likely, and what would have caught it". An investigation that arrives at the name of a person and stops has produced a disciplinary outcome, not a safety improvement.

Part-145 carries its own definition, and it is worth reading because it is deliberately broad. It describes human factors as the principles that apply to aeronautical design, certification, training, operations and maintenance, and that seek a safe interface between the human being and the other components of the system through proper consideration of human performance. Human performance, in the same sense, means the capabilities and limitations of the human being that have a bearing on the safety and efficiency of aviation. Notice what that pairing does: human factors is the body of principles, human performance is the raw material those principles have to work with. You cannot change the second, so the principles are entirely about arranging the five activities the definition names — design, certification, training, operations and maintenance — around it.

Because it is defined by the problem rather than by a technique, the subject borrows from several disciplines at once: human physiology; psychology, including perception, cognition, memory, social interaction and error; workplace design; environmental conditions; the human-machine interface; and anthropometrics, the measurement of the human body. A single maintenance task can call on all of them. Reading a torque value from a manual under a wing at three in the morning involves visual acuity and lighting (physiology and environment), working memory while you carry the figure to the tool (psychology), reach and posture in the space available (anthropometrics and workplace design), and the legibility of the page itself (human-machine interface). None of those is an aircraft systems question, and every one of them can produce a wrong torque.



Reading a torque value under a wing at three in the morning draws on several disciplines at once.

Why the Human Share of Occurrences Keeps Growing

Two things have been happening in parallel for the whole of the jet age, and they move in opposite directions.

The first is that the machinery has become dramatically more reliable. Damage-tolerant structural design, redundancy in every critical system, better materials and processes, built-in test equipment that reports a fault before anyone has to look for it, and engine reliability good enough to make long-range twin-engine operations routine — each of these removes a class of event whose cause was simply that a component broke. That improvement is real, it is enormous, and it is a large part of the reason the overall accident rate has fallen.

The second is that human reliability has not improved to anything like the same degree, and the reason is structural rather than cultural. As CAP 715 puts it, aircraft become more and more reliable through design and manufacture, but the human being cannot be re-designed; we have to accept that the human being is intrinsically unreliable, and can only work around that unreliability through good training, procedures, tools and duplicate inspections. Vigilance still decays with time on task. Memory is still displaced by interruption. Performance at four in the morning is still worse than performance at four in the afternoon. No engineering programme changes any of that.

Put those two trends together and the consequence is arithmetic rather than opinion: as the number of purely mechanical events falls and the number of events with a human contribution does not, the human contribution becomes a larger fraction of a smaller total. This is the single most misunderstood statistic in the module, so it is worth working through with numbers.

How a safer fleet produces a higher human percentage

Take a fleet that suffers 100 airworthiness events in a year. Suppose 40 of them are a component failing on its own with nobody at fault, and the other 60 have a human contribution somewhere in the chain. The human share is 60 out of 100, which is 60%.

Now improve the hardware. Better materials, redundancy and built-in test halve the pure component failures, from 40 down to 20. Nothing whatever has changed about the people, so their 60 events are still 60. The fleet now has 80 events in the year instead of 100 — a genuine improvement of a fifth — but the human share is 60 out of 80, which is 75%.

The human percentage rose by fifteen points in a year in which the people got neither better nor worse and the fleet got measurably safer. That is the mechanism behind every statement that human factors are a growing proportion of the problem. It also shows why the percentage on its own is a poor measure of an organisation's performance: it can be pushed up by good engineering and pushed down by bad.

The historical record supports the point. CAP 715 records that as early as 1940 it was being calculated that roughly 70% of aircraft accidents were attributable to human performance, and that when the International Air Transport Association reviewed the position some thirty-five years later it found no reduction in the human-error component of the accident statistics. Over that period the aircraft changed out of all recognition. The human share did not fall.

Within that human share, maintenance is not a footnote. A study of 93 aircraft accidents carried out in the United States in 1986 by Sears, reported in CAP 715, ranked the significant causes and major contributory factors. Deviation by the pilot from basic operational procedures appeared in 33% of them and inadequate cross-checking by the second crew member in 26%, but maintenance and inspection deficiencies appeared in 12% — essentially level with design faults at 13%, and ahead of several factors that receive far more attention. A comparable exercise by the UK CAA in 1998, examining the causes of 621 global fatal accidents between 1980 and 1996 and published as CAP 681, again found maintenance or repair oversight, error or inadequacy among the top ten primary causal factors.

The practical conclusion for an engineer sitting this examination is blunt: accidents and engineering faults traceable to maintenance and inspection are not a declining or a stable problem. Measured as a share of what still goes wrong, they are significant and increasing, and that trend is the reason human factors became a mandatory element of the licence syllabus and of maintenance organisation training rather than an optional extra for those who found it interesting.

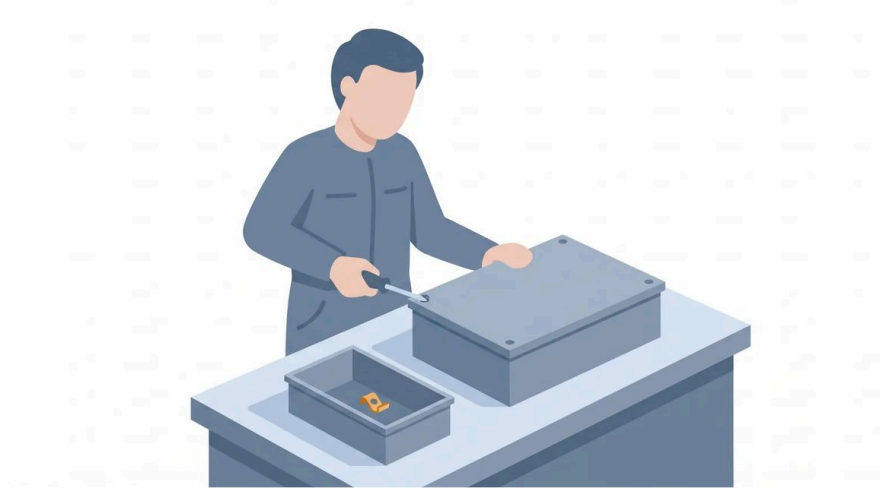
Reading Maintenance Error Statistics

Three quite different figures circulate in this subject and they are constantly confused with one another, in the hangar and in examinations. They are not in conflict; they count different things. Get the population right and each of them is useful.

Figure	The population it is drawn from	What it does not say
The figure quoted at the head of this note: around 80% of aviation accidents involve human factors	Every human contribution anywhere in the system taken together — flight crew, air traffic control, design, manufacture and maintenance	That maintenance engineers cause 80% of accidents. Maintenance is one contributor inside that total, alongside several larger ones.
Maintenance and inspection deficiencies were a factor in 12% of 93 accidents (Sears, 1986, reported in CAP 715)	Accidents in which a maintenance or inspection shortfall was one of the significant causes. An accident could carry several such factors at once, so the column does not total 100%.	That maintenance error is a minor issue. The same study put design faults at 13%, so the two are of comparable weight, and 12% of hull losses is a great many lives.
Omissions 56%, incorrect installation 30%, wrong parts 8%, other 6% (Reason's analysis of 122 maintenance incidents in one airline over three years, reported in CAP 715)	The internal breakdown of maintenance errors that are already known to have occurred. These four categories are exhaustive, so they total 100%.	That 56% of accidents are omissions. These are shares of maintenance errors, not of accidents, and the great majority of them never became an accident.

The first figure is also met in a narrower form, applied to maintenance rather than to aviation as a whole: that around 80% of maintenance-related incidents and accidents have human error as a contributing factor. In that form it is far less contentious, because a maintenance-related occurrence is very nearly by definition one in which something a person did, or did not do, is part of the story. Keep the two readings apart, because examination questions use both. The broad figure is about the whole system, in which maintenance is one slice; the narrow figure is about that slice on its own.

The third row deserves a moment on its own, because it tells an engineer where to concentrate. The largest single category of maintenance error is not doing the wrong thing; it is not doing something at all. Omission dominates because reassembly is intrinsically more error-prone than disassembly. There is normally one way to take an assembly apart and a very large number of ways to put it back together, of which only one is right and many differ from the right one only by something that has been left out — a seal, a lockwire, a clip, a cover, a torque. Several of the events summarised in the next section turned on something that was simply not done at all: screws not replaced, locking devices not fitted, covers not refitted.



A job that looks finished, with one part still in the tray: omission is the commonest maintenance error.

Two ways to misuse these numbers. The first is to quote the human-factors share as though it measured maintenance alone: "engineers cause eighty per cent of accidents" is not what any of these studies found, and it invites the defensive reaction that kills a reporting culture. The second is subtler and more damaging. It is to treat "human error" as the conclusion of an investigation rather than its starting point. Naming the error tells you what happened; it tells you nothing about why a competent person made it on that night, why nothing downstream caught it, and whether the same conditions are still sitting in the hangar tonight. Those are the questions this module exists to answer.

Who the Requirement Reaches

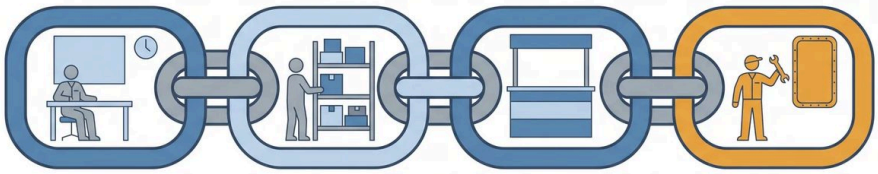
The obligation itself is short. Point 145.A.30(e) of Part-145 requires the maintenance organisation to establish and control the competence of its personnel to a standard agreed with the competent authority, and states that competence must include, in addition to the expertise needed for the job function, an understanding of the application of human factors and human performance issues appropriate to that person's function in the organisation. The rule text stops there. The structure of the training, the list of who must receive it and the length of the continuation cycle quoted in the next section all sit in the Acceptable Means of Compliance rather than in the rule.

That distinction matters in practice, and it is not a loophole. An Acceptable Means of Compliance is the Agency's published way of satisfying the rule; an organisation may propose an alternative means to its competent authority and have it accepted, but it cannot simply ignore the published one. In practice almost every organisation adopts the AMC, writes it into its maintenance organisation exposition, and is then bound by its own exposition — which is audited.

The AMC's list of who should receive initial and continuation human factors training is much wider than the people who touch the aircraft. It covers, as a minimum:

- post-holders, managers and supervisors;
- certifying staff, support staff and mechanics;
- technical support personnel such as planners, engineers and technical records staff;
- quality control and quality assurance staff;
- specialised services staff;
- human factors staff and human factors trainers;
- stores department staff and purchasing department staff;
- ground equipment operators.

The breadth of that list is the answer to a question the bank asks in several forms: whom does human factors matter to? It is equally important to technicians, engineers, planners and managers, and the reason is visible in the causal chains of every case in the next section. A planner who compresses a task into a window that does not fit it has created time pressure that a technician will absorb. A purchasing decision that leaves the correct part out of stock creates the moment at which somebody has to decide whether a near-equivalent will do. A stores absence removes an independent pair of eyes from the issue and check of equipment, without anyone deciding to remove a defence. None of those three people ever picks up a tool, and each of them can set an error in motion. The person holding the spanner is usually the last link in the chain and almost never the only one.



A planner, a purchasing decision and a stores absence all come before the technician, who is the last link.

Training is proportionate rather than identical. The AMC allows the syllabus to be adjusted to the nature of the organisation and to the nature of each function within it, and gives examples: planners may cover the scheduling and planning objectives in more depth and the shift-working material in less; a small organisation

that does not work shifts may treat teamwork and communication more briefly. Adjusted, however, is not the same as omitted — the topics remain the topics, and the syllabus for initial training is published as Guidance Material to the same point of Part-145.

On timing, the AMC expects all personnel, including people recruited from another organisation, to have received initial human factors training that meets the organisation's own training standard before they start their actual job function, unless a competence assessment justifies otherwise. There is one narrow relaxation: newly employed personnel who are working under direct supervision may receive the training within six months of joining. That relaxation is worth understanding rather than memorising: it is granted only because direct supervision is itself a defence, so it lapses at the moment the supervision does.

The SHELL Model

The **SHELL model** (developed by Edwards, 1972, modified by Hawkins, 1975) illustrates the relationships between humans and their working environment. The model places the individual (**Liveware**) at the centre, surrounded by four elements they interact with. The edges between the centre and each surrounding block are irregular — representing the need to match and adapt the interfaces.



SHEL, SHELL and the Second Liveware

Two spellings of the name are in circulation, SHEL and SHELL, and they refer to the same model. Edwards originated the concept in 1972, the name being taken from the initial letters of its components — Software, Hardware, Environment, Liveware — and Hawkins developed the modified diagram in 1975. The diagram above is that Hawkins version, and it is the one ICAO printed in Human Factors Digest No 1, published as Circular 216: the figure there is captioned "The SHEL model as modified by Hawkins", it places a person at the centre with a second Liveware block below for the other people that individual deals with, and the text alongside it gives the Liveware–Liveware interface a numbered paragraph of its own. So the spelling carries no information about the diagram — ICAO writes SHEL and its figure already has the second L, and the CAA's CAP 715 prints the same five-block figure under the same four-letter name. This note uses SHELL throughout, and the interfaces are named from the central person outwards: L–S, L–H, L–E and L–L.

The second L is not a fifth category of thing in the world. It is the recognition that other people are an interface in their own right and cannot be filed under Environment. Treating colleagues as part of the surroundings hides the failure modes that matter most in maintenance, because those failures are transactions rather than conditions: a handover that transfers a task but not its state, a briefing given verbally and remembered differently by two people, an instruction issued to somebody who is not qualified to question it, a junior engineer who sees something wrong and does not say so. None of those is a property of the hangar. Each is a property of an exchange between two people, and each has appeared as a causal factor in the accidents in the next section.

Why the Edges Are Ragged

The most important feature of the diagram is the one easiest to overlook: the blocks do not have straight edges, and they are drawn that way on purpose. The model's claim is that failures occur at the joints, not inside the blocks. Every block can be individually sound and the system can still fail, because two sound blocks do not fit each other.

Consider three cases where nothing is defective. A maintenance manual procedure that is entirely correct, but written for a different variant of the same type, is a good piece of software meeting a person who has no way of knowing it does not apply to the aircraft in front of him. A torque wrench that is in calibration and functioning perfectly, but whose scale cannot be resolved in the light available under a cowl, is good hardware meeting an environment that has disabled it. An accurate, complete verbal briefing given by one competent engineer to another, with no written record, is two competent people meeting an interface that has no memory. In each case an inspection of the parts finds nothing wrong. The mismatch is the fault.

There is a second-order effect that students almost always miss, and it is where real occurrences come from: the interfaces are not independent of one another. Each one silently sets the conditions under which the others can function. Degrade the lighting and you have not merely made the hangar unpleasant — you have disabled every defence that consists of somebody seeing something, without laying a finger on the hardware that provides those defences and without producing any record that they are now absent. Take a person's assistant away and you have done the same to every defence that consists of somebody else noticing. This is why an environmental or a staffing problem is never merely a comfort or an efficiency problem: it removes protection that the paperwork still says is in place.

One further asymmetry decides what to do about a mismatch. Four of the five blocks can be changed at will; the central Liveware cannot. Selection and training do adapt the person to the job, and they are genuine levers — a trained engineer performs a task quite differently from an untrained one. But the capacities that actually fail in these accidents are biological: the acuity of the eye in poor light, the persistence of working memory across an interruption, the level of alertness available at four in the morning. Instruction does not change them, and neither does willingness. So where a mismatch is found, the reliable fix is to change the software, the hardware, the environment or the way people are organised; asking the person to try harder is the weakest available response and the one that decays fastest once the briefing is over. CAP 715 states the aim as fitting the man to the job and the job to the man. Both halves are real. Only the second one still works on a bad night.

SHELL Interfaces Explained

Interface	Description	Example of Mismatch
L–H (Liveware–Hardware)	How the person interacts with tools, equipment, controls, and displays	Controls placed out of reach; displays too small to read; tools that don't fit the user's hand size
L–S (Liveware–Software)	How the person interacts with procedures, manuals, checklists, and regulations	Procedures that are ambiguous, too complex, or contradict each other; poorly written manuals
L–E (Liveware–Environment)	How the person is affected by the physical work environment	Excessive noise preventing communication; poor lighting during inspections; extreme temperatures
L–L (Liveware–Liveware)	How people interact with each other — communication, teamwork, supervision	Poor shift handover; language barriers; lack of assertiveness; personality conflicts

Working the Model Through One Job

The 1990 British Airways windscreen accident is summarised in the table at the start of the next section; it is used here only to show the model doing work. Reduce the whole event to a single sub-task — a shift maintenance manager selecting replacement bolts in a stores area during a night shift — and every interface

in the model can be seen failing quietly, none of them fatally on its own. The account below follows CAP 715's description of the accident.



At night in poorly lit stores, bolts are matched by sight and touch against a sample, without spectacles.

Interface	What was actually the case on that job	Why it mattered
L–S	The manual procedure for the job called for neither a pressure check nor a duplicate inspection afterwards.	One person's judgement was the only barrier between the selection and the flight. There was no stage at which a second person or a test could disagree.
L–H	The correct bolt and the bolt actually fitted differed by a small amount of diameter and by one character of part number. Both went in and both held on the ground.	Nothing in the hardware resisted the substitution. A design that cannot be assembled wrongly needs no vigilance; this one needed a great deal.
L–E	The selection was made at night, in a poorly lit stores area, by matching sight and touch against a sample bolt carried from the aircraft, and without spectacles.	Discriminating a small difference in diameter by eye and hand is at the edge of human capability in good light. In that light it was not a realistic task at all.
L–L	Short-handed, the manager elected to carry out the job himself and signed it off himself. The storeman advised that a different bolt was required, and the advice did not change the outcome.	The supervisory layer that would normally review the work was occupied doing it. A correct challenge was made by the one person present who was not under time pressure, and it was absorbed.

Not one of those four rows is a failure of technical knowledge, and the engineer concerned was not judged incompetent by anybody. Each is a mismatch at an interface and each, on its own, was survivable. The accident required all four to line up on the same task on the same night.

The clearest illustration of interfaces acting on each other is also in this case. Once the bolts were in position the countersink sat lower than it should have done. That is a physical cue: the hardware was reporting, in the only language it has, that the fit was wrong. It was not noticed. A cue is only a defence when the conditions allow somebody to see it and the working arrangement gives somebody the job of looking — which is the environment interface and the person-to-person interface jointly deciding whether the hardware interface gets to do its job. Read that way, the model stops being a diagram to memorise and becomes a checklist for planning work: for this task, on this night, which of the four is carrying the load, and what happens if it does not hold?

Examination points from this section: human factors is the study of how human capabilities, limitations, behaviour and performance interact with the work environment and the systems around it to affect safety — it is not the study of structural design and materials, and it is not the analysis of cockpit instruments and pilot response. The primary reason for studying it is to reduce human error, which is a contributing factor in the majority of aviation accidents and incidents. It is equally important to technicians, engineers, planners and managers, because a planning or management decision can set up an error just as readily as a hands-on slip. And the proportion of accidents attributable to maintenance and inspection causes is not declining: it is significant and rising.

Incidents Attributable to Human Factors/Human Error

The aviation industry has learned many of its human factors lessons through tragedy. Key maintenance-related accidents that shaped human factors awareness include:

Event	Human Factors Issues	Outcome / Lesson
Aloha Airlines 243 (1988)	Inadequate inspection of ageing structure; complacency; organisational failures	18 feet of fuselage roof separated. Led to ageing aircraft safety programmes.
British Airways BAC 1-11 (1990)	Wrong bolts used in windscreen replacement (84 of 90 too small); normalisation of deviance; time pressure; poor stores procedures	Windscreen blew out at 17,300 ft. Captain partially sucked out. Survived.
Continental Express 2574 (1991)	Shift handover failure — 47 screws left out of horizontal stabilizer leading edge after de-icing boot removal. Incomplete task not communicated.	In-flight breakup. 14 fatalities. Led to improved handover procedures.
Airbus A320 G-KMAM (1993)	Spoilers left in maintenance mode after a flap change; collars and flags not fitted; spoiler operation not fully understood; engineers more familiar with another type; inadequate shift handover briefing	Undemanded roll to the right after take-off from Gatwick. Returned and landed safely. A locked control surface passed the standard pilot functional checks.

Event	Human Factors Issues	Outcome / Lesson
Boeing 737-400 G-OBMM (1995)	HP rotor drive covers not refitted after overnight borescope inspections of both engines; verbal-only task handover; approved data not used; work by torchlight; required ground idle runs omitted	Almost all the oil lost from both engines in flight. Diverted to Luton and landed safely. One error repeated on both engines removed the redundancy.
China Airlines 611 (2002)	Inadequate repair of tail strike damage 22 years earlier; repair did not meet SRM requirements; fatigue crack grew undetected	In-flight breakup. 225 fatalities. Highlighted importance of repair quality.
Chalk's Ocean Airways 101 (2005)	Corrosion and fatigue in wing spar not detected during inspections	Wing separated in flight. 20 fatalities. Emphasised inspection quality.

Case Study: Both Engines, One Error

The double engine oil loss deserves a section of its own. It is one of the small number of maintenance cases that examinations use repeatedly, and it repays study in detail because nothing exotic happened in it: every single element of it was an ordinary night in an ordinary hangar. The account here follows the UK Air Accidents Investigation Branch report on the incident, published as Aircraft Incident Report 3/96.

In February 1995 a Boeing 737-400, registration G-OBMM, was due to receive routine 750-hour borescope inspections on both engines during the night before a charter flight from East Midlands Airport to Lanzarote. The task began normally. The line engineer in charge of the line night shift picked up the paperwork, went to the aircraft in a hangar some distance from the main line maintenance area, and started preparing the left engine. Preparation includes removing the high pressure rotor drive cover from the accessory gearbox, because that is where the core has to be turned from during the inspection. He removed it, and it hung on its lanyard.

He then needed the borescope equipment, which was in a tool store in another building, and a second person, because the HP spool has to be turned by somebody while the inspector looks through the probe. He walked across to base maintenance to ask for both. What happened next is the pivot of the whole event, and it was not a mistake by anybody at the time.

The base maintenance controller on that night shift had returned from a week's leave the previous evening. His shift should have had a foreman, four licensed inspectors and three licensed technicians; the foreman and three of the four inspectors were absent, leaving himself and one inspector to supervise the work. He was also aware that line maintenance had four people out of a nominal six and eight aircraft to deal with. And he had a personal interest in the task: company rules required him to perform at least two 750-hour borescope inspections in a twelve-month period to keep his company authorisation, opportunities were scarce, and he had recognised that he was in danger of letting it lapse. So he offered to take over the borescope inspections

personally if the line engineer would take over moving another aircraft that was in the way. The offer was accepted and the transfer of the task was logged at 2200 hrs.

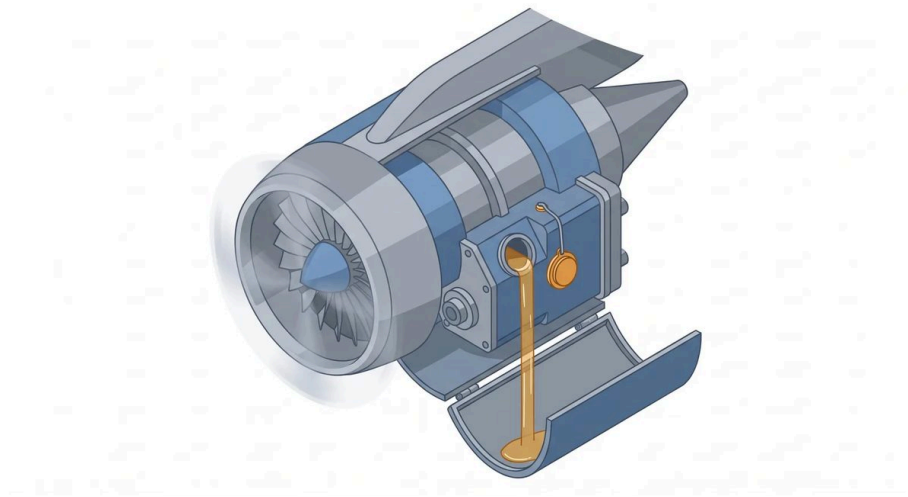
The handover that followed was verbal. The line engineer briefed the controller on where he had reached: the aircraft was dead, with no electrical or hydraulic power and nobody else working on it, the left engine was fully prepared, and the paperwork pack was with the aircraft. There was no written statement of the stage the job had reached and no annotated work stage sheet — and, as the report records, there was no suitable place on the line maintenance paperwork to note it. Procedurally, additional paperwork should have been raised. It was not, and the receiving engineer was content with what he had been told.

From there the conditions accumulated. The controller selected a fitter to assist him who was not licensed, though he had three years in the company and some previous experience of assisting on this inspection. The storeman was also absent, so the controller used his own key to draw the borescope equipment and he and the fitter checked it themselves — a check that would normally have been done by the storeman, and therefore by somebody who was not about to use the kit. The paperwork was line maintenance paperwork, unfamiliar to a base engineer and carrying no manufacturer's task cards; the controller considered it unnecessary to draw any additional reference material and worked from his own personal, annotated copy of the borescope training manual. The hangar was partly occupied by workshops and was not big enough to take the aircraft, so it was pulled in far enough for the doors to be partially shut against the fuselage sides just ahead of the tailplane. It had no sockets compatible with the engineers' lead lights, which reduced them to working by torchlight; the report notes that it was particularly dark under the engine cowls in the area of the accessory gearbox, and that against the lighting guidance it quoted, the hangar's level was graded as suitable only for infrequently used areas.

The controller pointed out the removed HP rotor drive cover on the left engine to the fitter and told him to prepare the right engine in the same way. Both inspections were carried out. Neither cover was refitted. The post-inspection ground idle engine runs required by the aircraft maintenance manual were not carried out either. The job was signed off in the aircraft technical log as complete.

The following morning the dispatching engineer reviewed the technical log before his pre-departure checks and saw two borescope inspections signed off as satisfactory. There was no outstanding item to query and nothing visible from outside a closed cowl. The one anomaly the crew did find was that the hydraulic power circuit breakers had been left open; the first officer asked the dispatching engineer about them, the engineer could see no reason for them to be open, and the first officer closed them. An unexplained departure from the expected state after an overnight maintenance visit is information about that visit. Here it was tidied away rather than accounted for, which is what usually happens to a small anomaly when the paperwork says everything is complete. The aircraft began its take-off roll at 1157 hrs. Climbing through about flight level 140 the commander noticed both oil quantity gauges reading around 15% and falling; the oil pressures then began to fall as well. The crew declared a Mayday, diverted, and landed at Luton, where both engines were shut down during the landing roll. Almost all the oil had been lost from both engines. Nobody was injured and the aircraft was undamaged.

Two details from the investigation are worth holding on to. The first is what the engine manufacturer's subsequent test showed: with the cover missing, motoring the engine over on the starter with the ignition off produced only a very fine oil mist from the open port in the gearbox casing, but starting the engine and running it at ground idle — exactly what the maintenance manual required after the inspection — produced a significant outflow of oil. The omitted check was not a formality. It was the action on that task that would have made the fault announce itself without anybody having to go looking for it, and it is specified after a borescope inspection for exactly that reason.



With the HP rotor drive cover left off its gearbox port, running the engine at ground idle pushes oil out.

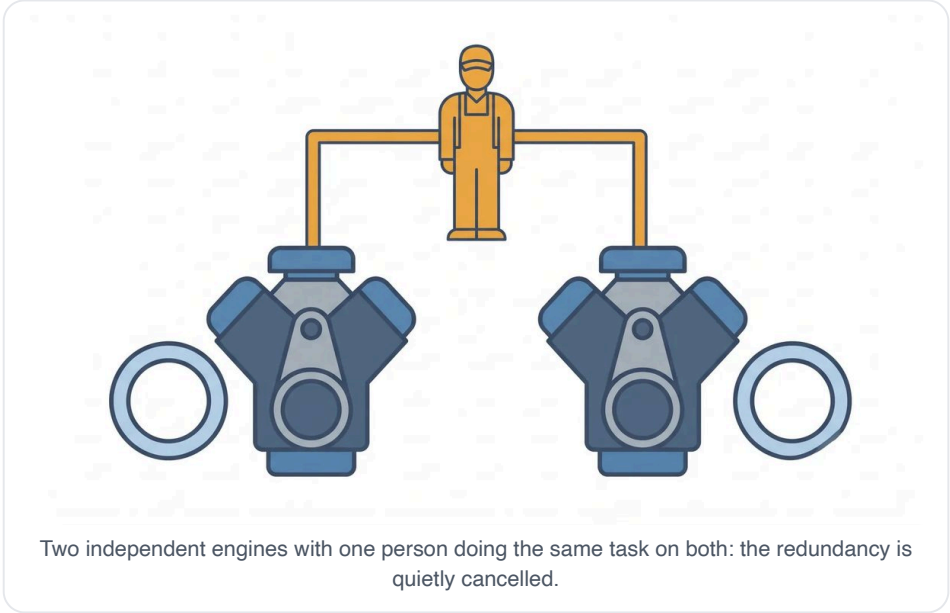
The second is how far up the report's causal factors reach. They record the aircraft being presented for service with the inspections signed off although the covers had not been refitted; failure to comply with the maintenance manual in a number of areas, most importantly the covers and the ground runs; that the operator's quality assurance department had not identified the non-procedural conduct of borescope inspections that had been prevalent among company engineers over a significant period; and that the Civil Aviation Authority, reviewing the company's procedures for its JAR-145 approval — the framework that preceded today's Part-145 — had detected limitations in aspects of the quality assurance system, including procedural monitoring, but had been satisfied they were being addressed and had not withheld approval. Fifteen safety recommendations were made. Two people got one job wrong on one night; the causal factors reach past both of them to the company's routine practice, to its quality department and to its regulator.

Why "coincidence" is not an available answer

Two engines on a twin are deliberately independent. They have their own oil systems, their own accessory gearboxes, their own indications, and the whole case for despatching a twin over water rests on the assumption that whatever fails on one of them has no bearing on the other.

Suppose an oil system failure of any kind has some small probability, call it p, on a given flight. If the two engines really are independent, the probability of both failing on the same flight is p multiplied by p. For an event as rare as total oil loss, p is already very small; p multiplied by itself is smaller again by that same factor. The result is a number so small that you would not expect the double event even once in the service life of a whole fleet.

So when both engines do lose their oil on the same flight, the observation is not evidence of very bad luck. It is evidence that the independence assumption was false — that something acted on both engines. The only thing that had acted on both engines was the maintenance. That is the reasoning behind the examination answer: the event is a direct result of human error, not a statistically expected coincidence, not something within accepted probability limits, and not loose bolts or a design weakness. The moment one person performs the same task on two supposedly independent systems, the redundancy that justified the design has quietly been cancelled.



Two independent engines with one person doing the same task on both: the redundancy is quietly cancelled.

The Rule That Came Out of Cases Like This

Point 145.A.48 of Part-145, on the performance of maintenance, is where the industry's answer to that reasoning now sits, and reading it against the case above shows how directly accident findings turn into rule text.

The procedures an organisation writes must be aimed at minimising multiple errors and preventing omissions. The Acceptable Means of Compliance spells out what that means: every maintenance task is signed off only after it has been completed; where tasks are grouped for the purpose of sign-off, the grouping must still allow critical steps to be identified; and work performed by personnel under supervision, such as temporary staff and trainees, is checked and signed off by an authorised person.

Then comes the provision that addresses the double oil loss directly. Procedures must minimise the possibility of an error being repeated in identical tasks and so compromising more than one system or function. In practice they should ensure that no one person is required to perform a maintenance task involving the removal and installation, or the assembly and disassembly, of several components of the same type fitted to more than one system on the same aircraft during a particular maintenance check, where a failure of those components could affect safety. Where unforeseen circumstances leave only one person available, the organisation may use a reinspection instead. Read that sentence with two engines and one borescope inspector in mind and its origin is obvious.

Alongside it sits the concept of the **critical maintenance task** and the **error-capturing method**. An error-capturing method is defined as an action the organisation has decided upon to detect maintenance errors made while performing maintenance, and the AMC is explicit that it must be adequate for the work and for how much the system was disturbed — a combination of actions such as a visual inspection, an operational check, a functional test and a rigging check may be needed for one task. The ground idle run in the case above was precisely such a method, and it was the last opportunity that job had to detect the omission. Dropping a check of that kind is a different and much larger decision than skipping an ordinary step, because it does not merely leave one thing undone: it removes the mechanism by which everything else would have been verified.

The AMC also lists where an organisation should look to decide which of its tasks are critical: information from the design approval holder, accident reports, the investigation and follow-up of incidents, occurrence reporting, flight data analysis, the results of audits and independent inspections, monitoring schemes, and feedback from training. That list is the formal mechanism by which the events in the table above stop being history and start being procedure.

What the Cases Have in Common

CAP 715 examines three of the UK events together — the 1990 windscreen accident, the 1993 spoiler incident and the 1995 oil loss — and draws out the features they share. Before the list, one finding that frames all of it.

> **These were not bad engineers.** In all three of those UK cases, the engineers involved were considered by their own companies to be well qualified, competent and reliable employees, and in every one of them the hands-on work was being done by a supervisor. In the 1988 Aloha inspection failure the two inspectors who found no cracks had 22 and 33 years of experience respectively, and post-accident analysis established that the aircraft carried over 240 cracks in its skin at the time they looked at it. If your working model of a maintenance accident is that somebody incompetent was involved, you have no model at all, and you will not see the next one coming.

The features common to all three UK cases were these:

- there were staff shortages;
- time pressures existed;
- all the errors occurred at night;
- shift or task handovers were involved;
- they all involved supervisors doing long hands-on tasks;
- there was an element of a "can-do" attitude;
- interruptions occurred;
- there was some failure to use approved data or company procedures;
- manuals were confusing;
- there was inadequate pre-planning, equipment or spares.

The most useful thing about that list is what kind of thing it is. None of these ten items is a cause in the sense of being the thing that made a particular event happen; each is a **precondition**, a state of affairs that raises the probability of error across every task carried out in it. That difference is what makes the list operationally valuable. A cause can only be identified after the event. A precondition is visible beforehand — on a roster, in a workload plan, on a walk round the hangar at two in the morning — and can be counted before anything has gone wrong. The ten group naturally into four families.

Resourcing. Staff shortages, time pressure, and inadequate pre-planning, equipment or spares are the same problem seen from three angles: the work to be done exceeds the capacity available to do it properly. The important point is where the shortfall is absorbed. It is not absorbed by the aircraft leaving late; it is absorbed by the individual, in the form of shortcuts, deferred checks and tasks accepted without the right data. Every one of the three UK cases has a resourcing shortfall standing immediately behind the error, and in the oil loss case the shortfall existed on both shifts at once.

Time of day. All the errors occurred at night, and that is not an artefact of when maintenance is scheduled. Alertness, vigilance and the ability to sustain concentration all fall through the small hours, and they fall furthest at the end of a night shift — which is exactly when a task started at the beginning of the shift reaches its final reassembly and inspection. The night shift concentrates the most error-prone part of the work into the least capable part of the human day. The physiology behind this is developed later in the module; here it is enough to note that it is present in every one of the three cases.

Information transfer. Handovers, interruptions, confusing manuals and failure to use approved data are all failures of the same commodity: knowing the true state of the job. A handover is the point at which the state of a task has to be moved from one person's head into another's, and it is the only routine moment in maintenance when that transfer is guaranteed to be necessary. An interruption does the same thing to one person across time — the state has to survive the gap, and human working memory is poor at that. A manual that is confusing, or approved data that is not in front of you, means the correct state of the job is not available

even in principle. In the oil loss case a handover, repeated interruptions and the absence of the manufacturer's task cards were all present on the same task.

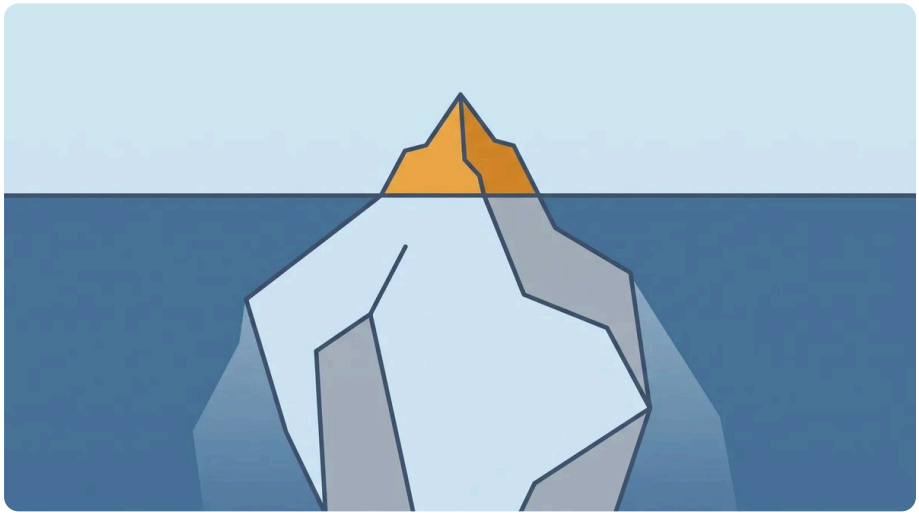
Role. Supervisors doing long hands-on tasks, and the "can-do" attitude, are the subtlest pair. A supervisor absorbed in a two-hour task is not supervising during those two hours, but the organisation's chart still shows a supervisor on shift, so the layer is believed to be there when it is not. Worse, when the senior person does the job, the check that a supervisor normally provides is being applied by the individual who most needs it. That is exactly what happened in the 1990 windscreen accident and in the 1995 oil loss, in both of which the senior engineer present carried out the work personally; in the windscreen case he also signed it off himself. The "can-do" attitude is not a vice — it is what keeps a fleet flying and it is why the industry works at all. It becomes dangerous at one specific moment: when it is the reason a task is accepted that the available resources, data or conditions do not actually support.

From Error to Accident: the Chain and the Iceberg

Two models explain why these events are both preventable and rare, and they answer different questions.

The first is the **error chain**, illustrated in CAP 715 with a Boeing diagram in which links contributed by management, by maintenance and by the crew run in sequence to an accident. The claim is simple and it is the most encouraging thing in the subject: break any one link and the accident does not happen. It does not matter which one. Applied to the oil loss, the chain had at least four links that could have been broken independently by different people on different days — raising the additional paperwork at the handover so the state of the job was recorded; drawing the manufacturer's task cards rather than working from a personal copy; carrying out the ground idle run the manual specified; or a quality audit at any point over the preceding period noticing that borescope inspections were routinely not being done to procedure. Any single one of those four, on its own, ends the story before the aircraft flies.

The second is the **iceberg**. CAP 715 draws accidents as the visible tip, with serious incidents, incidents, errors and minor events in descending layers below the waterline. Most errors made by maintenance engineers never produce an accident, and that is the source of the most dangerous inference in the industry: because nothing happened, nothing was wrong. The correct inference is different. On that occasion, either a defence caught the error or the circumstances happened to be forgiving — and neither of those is guaranteed to repeat. The same error in different circumstances is the accident. The submerged part of the iceberg is not a record of harmless events; it is a preview.



The iceberg: accidents are the visible tip; incidents, errors and minor events lie unseen below the waterline.

That is the argument for occurrence reporting, and it is a data argument rather than a moral one. Accidents are too rare to learn from quickly; incidents and errors are common enough to reveal a trend while there is still time to act on it, and they are the only large-volume evidence anyone has about what is happening below the waterline. CAP 715 records that all maintenance incidents in the UK had to be reported to the CAA's Mandatory Occurrence Reporting Scheme so that the data could be used to disclose trends and drive action. Across the EU the equivalent obligation now sits in Regulation (EU) 376/2014 on the reporting, analysis and follow-up of occurrences, which also requires that reports are handled in a way that does not penalise the reporter for honest error. That last provision is not sentiment. A reporting scheme that people are afraid of collects nothing, and an organisation with no reports is not an organisation with no errors — it is an organisation that has lost sight of its own iceberg.

✈ **Aviation Context:** EASA made Human Factors training mandatory under Part-145. Initial and recurrent training (every 2 years) is required for certifying and support staff, and for other personnel in proportion to their role. This reflects the understanding that technical knowledge alone is insufficient — engineers must also understand the human system they are part of.

What that training must contain is not left to taste. The syllabus for initial human factors training is published as Guidance Material to the same point of Part-145, and its opening topic is the need to address human factors — the subject of this note. Continuation training is meant to be evidence-driven rather than repetitive: the Acceptable Means of Compliance ties its duration and content to the organisation's own quality audit findings and to internal and external information about human error in maintenance, and expects feedback collected during the training to be passed formally to the quality department so that action can be initiated. That closes the loop opened by the accidents above. Occurrences produce reports, reports identify which tasks are critical and which conditions keep recurring, and both feed the training that the next shift receives. An organisation whose human factors training says the same thing every cycle regardless of what its own occurrence reports say is meeting the letter of the requirement and losing all of its value.

Murphy's Law

"Anything that can go wrong, will go wrong."

Attributed to Captain Edward A. Murphy Jr., a US Air Force engineer, in 1949 during rocket-sled deceleration tests at Muroc Army Air Field (now Edwards Air Force Base). A technician had wired all 16 strain gauges backwards. Murphy's observation was about designing systems to prevent human error — not about pessimism.

What the Law Claims, and What It Does Not

Taken as a statement about a single event, Murphy's Law is obviously false: most tasks go right. It is not a statement about a single event. It is a statement about *exposure* — about what happens to a small per-task probability when the task is repeated across a fleet, a network and a service life. The word doing the work in it is not "wrong" but "will", and the honest expansion is: given enough opportunities, every available failure path is eventually taken.

What "eventually" is worth in numbers

Suppose a particular reassembly step can be got wrong, and suppose a competent, rested engineer gets it wrong once in every thousand attempts — a rate most people would call excellent. Across a fleet the task comes round 5,000 times a year.

The expected number of wrong reassemblies is 5,000 multiplied by 0.001, which is 5 per year. The chance of a whole year passing with none at all is under one per cent. In other words, at this rate the error is not a risk. It is a schedule.

Now suppose the organisation responds with a briefing, and suppose the briefing is unusually effective and halves the error rate to one in two thousand. The expected number falls to 2.5 a year. That is a real improvement, and it is nowhere near enough, because the failure path is still open and the fleet still flies 5,000 of these tasks.

Now suppose instead that the component is modified so that it physically cannot be assembled the wrong way. The expected number is zero, and it stays zero on night shifts, in the rain, at the end of a twelve-hour duty, and for the engineer who joined after the briefing.

The numbers are an illustration rather than data from any particular fleet, but the shape of the result is completely general, and it is the whole argument for designing errors out rather than exhorting people not to make them.

Two misreadings are worth naming because both appear as distractors. The first is fatalism: if everything that can go wrong will, why bother? That inverts the law's meaning. Used fatalistically it is an excuse offered after an event; used as engineering it is an instruction issued before one, and the instruction is specific — find the failure paths and close them. The second is blame. Murphy's Law says nothing about whose fault anything is.

It is not the proposition that errors are always due to neglect, and it is not the proposition that blame rests with the last person to touch the aircraft. Both of those are statements about attribution; the law is a statement about probability and design.

Complacency: the Attitude That Keeps Proving It

CAP 715 introduces the law by way of an observation about people rather than about probability: there is a tendency among human beings towards complacency, and the belief that an accident will never happen to me or to my company is a major obstacle to getting anyone to take human factors seriously, recognise risks and act on them rather than paying lip service. That is why the examination answer to what perpetuates Murphy's Law is complacency, and it is worth understanding why complacency specifically, rather than the two answers that sound equally plausible.

Inattention is a momentary state. Attention wanders, it comes back, and the person often knows afterwards that it wandered. It is exactly the failure mode that ordinary defences are built for: a checklist that has to be signed line by line, an independent inspection, a functional check. Inattention produces errors, and the system is designed on the assumption that it will.

Carelessness is not a mechanism at all. It is a judgement about how much somebody cared, made after the fact by somebody who knows the outcome. As an explanation it is unfalsifiable and it predicts nothing: it cannot tell you which task is at risk tomorrow, and it points at no defence you could install. Investigators avoid it for that reason.

Complacency is different in kind from both. It is a settled expectation built by success. Every previous occasion on which the task went well strengthens the belief that this occasion will too, and that belief acts not on the hands but on the *defences*. A complacent engineer does not merely make a slip; he stops looking for the slip. The inspection becomes a formality, the check becomes a signature, the manual stays in the office because "I have done this a hundred times". Inattention defeats the person and leaves the barriers standing. Complacency dismantles the barriers first.

It also has a mechanism that makes it self-reinforcing, and the direction of that mechanism is worth stating carefully. As a task is repeated without incident, the *perceived* risk falls. The *actual* risk does not change at all, because nothing about the hazard has been touched. The gap between the two therefore widens with every uneventful repetition, and it widens fastest for exactly the people with the most experience of the task. That is not a paradox; it is what happens when the only evidence available is the absence of a bad outcome, and most errors do not produce bad outcomes.

The double oil loss is the worked case. The report found that the non-procedural conduct of borescope inspections had been prevalent among the company's engineers over a significant period. Every one of those

earlier inspections was completed and every aircraft came back. Read as evidence, that record said the shortcut was safe; in reality nothing about it had ever been safe, and the accumulated absence of consequences was simply a run of luck being mistaken for a result. That is also how a shortcut becomes a norm and eventually becomes the way the job is done — the process the accident table above calls normalisation of deviance in the 1990 windscreen case. Complacency is one of the twelve error precursors treated elsewhere in this module; what matters here is its specific relationship to Murphy's Law. The law says that an open failure path will eventually be taken. Complacency is the state of mind in which nobody closes it, because nothing has gone wrong yet.

One corollary follows immediately, and CAP 715 states it plainly: it is not true that accidents only happen to people who are irresponsible or sloppy, or to organisations that were already known to be poor. Errors are made by experienced, well-respected individuals, and accidents occur in organisations previously thought to be safe. An examination option claiming that errors only happen to untrained or uncertified personnel is not a near miss; it is the exact belief the law exists to attack.

Application in Aviation Maintenance

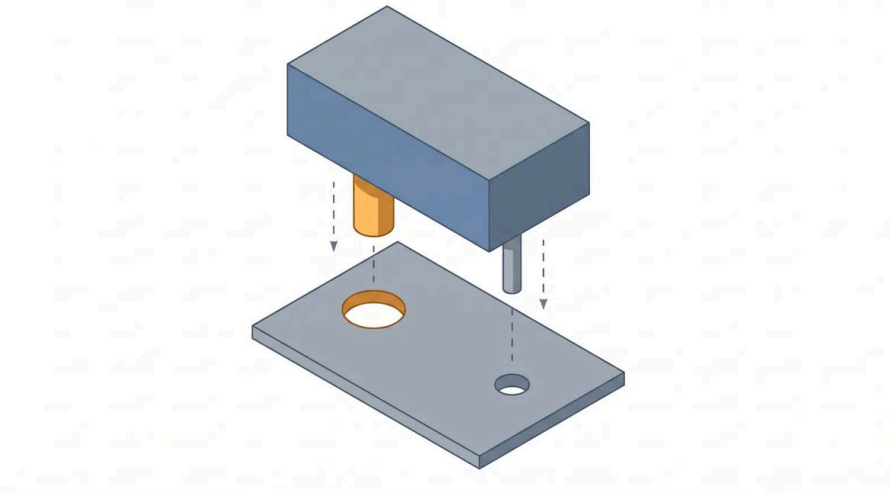
- **Design for error prevention** — if a connector can be plugged in backwards, someone eventually will plug it in backwards. Use keyed connectors, poka-yoke (mistake-proofing), and asymmetric designs.
- **Assume errors will happen** — design systems with multiple layers of defence (checklists, independent inspections, BITE).
- **Plan for the worst case** — if a failure mode exists, it will eventually occur. Identify failure modes and mitigate them proactively.
- **Don't rely on human perfection** — procedures, training, and good intentions alone do not prevent errors. System design must account for human fallibility.

How Error-Proofing Actually Works

"Design the error out" is easy to say and covers three quite different mechanisms, which are not equally reliable. The useful way to rank them is by how many human acts each one depends on, because every human act in a defence is a place where the defence can fail on a bad night. Fewer acts, stronger defence.

First, and strongest: make the wrong action impossible. Different thread forms and pitches so a line cannot be cross-connected. Connector keyways, differing shell sizes and different pin counts so a harness will not mate with the wrong socket. Asymmetric bolt patterns and mounting lugs of different sizes so a component only goes on one way round. The wrong assembly does not go together, so fatigue, time pressure, poor light, an interruption and inexperience have nothing to act on. This is the only category whose effectiveness does not depend on the state of any person, which is why it is worth a great deal of design effort and why it is the

first thing a maintainability review should look for. Its limitation is that it has to be decided at design time; an engineer meeting the aircraft twenty years later cannot add it.



Mounting pins of different sizes let a component go on only one way round, so the wrong fit cannot happen.

Second: make the wrong action announce itself under test. Functional checks, operational checks, rigging checks, leak checks and ground runs. These do not prevent the error; they cause the system to reveal it before the aircraft flies. They depend on two human acts — that the check is actually carried out, and that the check chosen is adequate for how much the system was disturbed, in the sense the AMC uses earlier in this section. With both right, the result no longer depends on anybody's vigilance, because the aircraft does the detecting. With the second wrong, a check can pass a fault it never exercised: the G-KMAM spoilers in the table above were left in maintenance mode and still passed the standard pilot functional checks. That is a much narrower dependency than it sounds, and it is why this category outranks the next one.

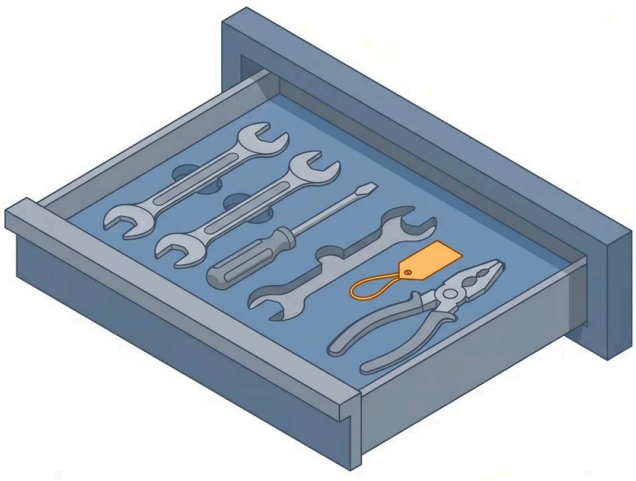
Third: make the wrong action visible. Lockwire, torque witness marks, streamered blanks and covers, warning tags, a removed part left hanging on a lanyard. These raise the probability that somebody notices, and they cost almost nothing, so they are everywhere and they are worth having. But they depend on two human acts rather than one: somebody must look, and the conditions must permit looking. Degrade the light, the access or the time available and a passive cue quietly stops being a defence while remaining fully in place. A duplicate or independent inspection sits at this level too, even though it adds a person rather than a cue. It is a second person looking, so it depends on the same two acts: the inspection has to be done, and that person has to see the defect in the light, the access and the time available. What it adds over a passive cue is real — the second person did not do the work, so they do not arrive carrying the first person's expectation of what it should look like — but what it does not add is any independence from human vigilance, which is why a check the aircraft itself answers is the stronger barrier wherever the task admits one.

Below all three: exhortation. Briefings, posters, "be careful", a stern word after the last event. These depend on the person remembering the message and acting on it while under exactly the load that caused the

problem, and their effect decays from the moment the briefing ends. They are not worthless — awareness is a precondition for everything else — but treating them as a control is the error the whole subject exists to correct.

In plain sight, and still missed. The third category fails more often than engineers expect, and the double oil loss described in the previous section is the proof. The removed cover was not hidden in a toolbox. It was hanging from the accessory gearbox on its lanyard, on both engines, and it was explicitly pointed out and discussed during the job. It was still not refitted — because these people were working under a cowl at night with a torch in one hand, in a hangar whose lighting was rated for infrequent use, and a cue in those conditions is a cue nobody is going to look at. The defence that would have caught it sat in the second category and needed nobody to look at anything at all. The practical rule: when the environment is working against you, do not add another cue. Move the defence from noticing to testing.

There is a converse worth stating, because it is where the third category earns its place. A cue can be strengthened by making it impossible to ignore rather than merely possible to see: a blank with a long red streamer that hangs into the walkway, a cover that fouls the cowl so the cowl will not latch, a tool tag that stays in the toolbox slot until the tool comes back. Each of these takes a passive cue and gives it an active consequence, so that missing it costs the engineer something before the aircraft is released rather than after. That is the cheapest available upgrade in maintenance human factors and it needs no design authority approval, only somebody who has thought about how the job actually goes.



A tool tag stays in the toolbox slot until the tool comes back, so a missing tool cannot be overlooked.

Murphy's Law in Engineering Terms:
If there are *n* ways to perform a task, and one of them leads to a catastrophic outcome, then someone, at some point, will perform it that way. The probability is not *if* but *when*. The correct response is to make the catastrophic way impossible — not to tell people to be more careful.

Murphy's Law on the Hangar Floor

The examination has a favourite scenario for this topic, and it is worth working through because the reasoning generalises. A spanner is placed on a wing surface while the engineer works. It is knocked or kicked off, falls into an open engine cowl, and breaks a sensor connector on the way down. What kind of event is that?

It is Murphy's Law. An available failure path was left open, the task was repeated enough times, and on this occasion the path was taken. It is not a team error, which would be a failure in the way two or more people coordinated their work — here nobody miscommunicated anything. It is not a lapse either: in the error taxonomies used elsewhere in this module a lapse is a failure of memory, and nobody forgot a step. Nothing was forgotten and nobody was misinformed. What happened is that the job was set up in such a way that a bad outcome was reachable, and reachable outcomes are eventually reached.

The apparent malice of it — the tool finding the one opening that mattered — has a mundane explanation, and understanding it converts the joke into a control. Where a dropped object ends up is not evenly distributed over the ramp. Aircraft surfaces are curved and sloped, so an object that starts moving keeps moving in a predictable direction; and the apertures that are open during maintenance are open precisely because people are working at them, so they are close to where objects get put down. The distribution of landing places is therefore heavily biased towards the openings, which is exactly the set of places where an object does damage. Murphy's Law, in this reading, is the working assumption that you will get the biased outcome rather than the average one — which, given the bias, is simply good estimation.

The controls follow directly, and they are all about the task set-up rather than about care:

- **Nothing rests on an aircraft surface.** Not a tool, not a torch, not a cup, not a part. A trolley or a mat costs nothing and removes the failure path entirely, which puts it in the strongest category described above.
- **Openings are blanked as soon as they are opened.** A cowl that is open for two hours is a target for two hours. Blanks and covers with streamers do two jobs at once: they close the path, and their absence is visible from a distance.
- **Tools are accounted for before anything closes.** Shadow boards, foam cut-outs and tool tallies exist so that a missing item is discovered while the panels are still open and while the search area is still small. Discovering it afterwards means either an unserviceable aircraft or an unrecorded object loose in a structure.
- **The account is against the task, not the shift.** Checking the box at the end of a twelve-hour shift tells you a tool is missing but not which of six jobs it is in. Checking at the end of each task tells you where to look.

CAP 715 offers a smaller example in the same spirit, and after two cases in which lighting was a factor it is worth taking seriously. A torch is very useful to an engineer, but Murphy's Law dictates that the batteries will

run down at the moment the engineer is on the far side of the airfield from the stores. The response is not to be more careful with torches; it is to carry a spare set of batteries rather than gamble on attempting a job without enough light. The pattern is identical to the one in the accidents: identify the failure path, decide in advance what happens when it is taken, and remove the need for anybody to be lucky.

✈ **Examination points on Murphy's Law:** state it as the notion that if something can go wrong, it will. It is perpetuated mainly by **complacency** — the settled belief that it will not happen to me or to my company — rather than by momentary inattention or by carelessness. Its constructive application in maintenance is to design systems, procedures and checks that assume errors will occur and to build in defences that catch them; it is not a licence to accept errors and merely document them without preventive action, and it is not answered by hiring only experienced staff and trusting their expertise. And where a question describes a chance chain such as a spanner falling from a wing into an open cowl and damaging a connector, the answer being tested is Murphy's Law.

Module 10 — Aviation Legislation

10.1 — Regulatory Framework

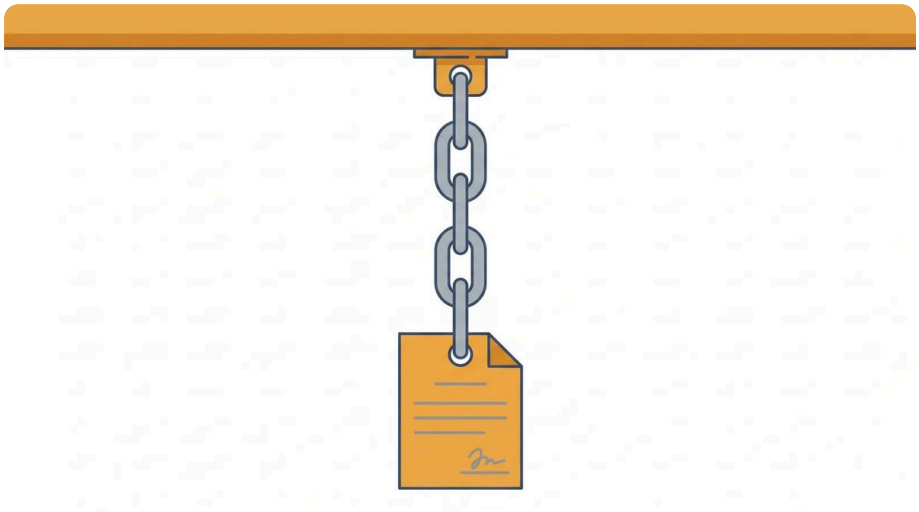
Introduction

Aviation is one of the most heavily regulated industries in the world. A single aircraft can cross multiple national boundaries in a single flight, carrying hundreds of passengers. This demands a **harmonised international regulatory framework** that ensures uniform safety standards everywhere. Module 10.1 establishes the foundation — the organisations, regulations, and relationships that govern everything an aircraft maintenance engineer does.

Why this matters: Every certificate you sign, every maintenance task you perform, and every part you fit traces back to the regulatory framework described here. Understanding *where* rules come from and *how* they interact is essential for legal compliance.

Reading the chain of authority

The framework set out below is a chain, and every link takes its force from the link above it. A convention between States fixes an international baseline. A regulation of the European Union turns the part of that baseline for which the Union has taken responsibility into law that binds directly in every Member State. The European Commission adopts the technical detail as further regulations. The Agency publishes the means of complying with that detail. Your own organisation then writes an exposition or manual describing how it will comply, and the competent authority approves it. When you sign a certificate of release to service, you exercise a privilege that traces, link by link, all the way back up that chain.



The framework is a chain: each link takes its force from the one above, down to your release certificate.

Read downwards, the chain tells you what the rules say. Read upwards, it tells you what has gone wrong when something does not fit — a company instruction that no approved exposition covers, a maintenance

data reference with no design approval behind it, a licence privilege the licence itself does not carry. Almost every question in this module is, underneath, a question about which link you are standing on and which link you are relying on.

How to read a rule reference

Nearly every requirement you will meet is cited in the same shape: the Part, a letter, then a number. The letter is the piece most students skip, and it is the piece that tells you whom the requirement is aimed at. **Section A binds the person or the organisation; Section B binds the competent authority.** Point 66.B.1 says so in as many words: that Section "establishes the procedures including the administrative requirements to be followed by the competent authorities in charge of the implementation and the enforcement of Section A of this Annex (Part-66)". Point 145.A.10 and point 147.A.05 make the mirror statement about their own Section A, which "establishes the requirements to be met by an organisation".

Reference	Where it sits	Whom it binds
M.A.201	Annex I (Part-M), Section A	the owner, the lessee or the operator
145.A.50	Annex II (Part-145), Section A	the approved maintenance organisation
145.B.350	Annex II (Part-145), Section B	the competent authority
66.A.20	Annex III (Part-66), Section A	the licence holder
66.B.10	Annex III (Part-66), Section B	the competent authority
147.A.05	Annex IV (Part-147), Section A	the maintenance training organisation
21.A.20	Annex I (Part 21) to (EU) No 748/2012, Section A	the applicant or the approval holder
21.B.80	Annex I (Part 21), Section B	the Agency, acting as the certifying authority

A reference carrying no letter at all — 145.1, 66.1, 147.1, CAO.1, T.1 — sits ahead of both Sections and does the same job in each of these five Parts: it names the competent authority. Those five points are worth knowing individually, because the test they apply is not the same in all of them.

Knowing which text you are reading

An EU regulation is amended by later regulations rather than reprinted, so two versions of the same instrument are in circulation at any time. The version published in the Official Journal is the authentic one; the version most people actually read is a consolidated text, in which every amendment made up to a stated date has been folded into the original. A consolidated text carries a printed warning that it "is meant purely as a documentation tool and has no legal effect" and that the authentic versions are those published in the Official Journal. It also carries markers through the body showing which amending regulation introduced each passage, which is how you can see that a point you are relying on was rewritten and when.

The practical consequence is a habit worth forming early: **always note the consolidation date of the text you quoted.** A requirement recited correctly from a text that is three amendments out of date reads exactly like a correct citation, and nothing on the page will tell you otherwise. It is a common way for a finding to be raised against a procedure that was written carefully and then never revisited.

In the exam and on the job: the examination tests the current wording, and your organisation's exposition has to match it as well. When a regulation changes, three separate things have to be brought up to date — the rule text you consult, the company procedure that implements it, and the training that tells staff what to do. They are usually updated by three different people at three different times, and the gap between them is where findings live.

ICAO — The Global Foundation

The **International Civil Aviation Organization (ICAO)** was established by the **Chicago Convention of 1944**. ICAO is a specialised agency of the United Nations, headquartered in Montreal, Canada. Its purpose is to develop **Standards and Recommended Practices (SARPs)** that provide the basis for harmonised global aviation safety.

Concept	Detail
Standards	Must be complied with by contracting states — departures must be formally notified to ICAO
Recommended Practices	Contracting states are encouraged to follow them but have more discretion

ICAO publishes **19 Annexes** to the Chicago Convention. Key annexes for maintenance engineers include:

Annex	Subject	Relevance
Annex 1	Personnel Licensing	Basis for Part-66 licence requirements
Annex 6	Operation of Aircraft	Operator continuing airworthiness obligations
Annex 8	Airworthiness of Aircraft	Minimum airworthiness certification standards

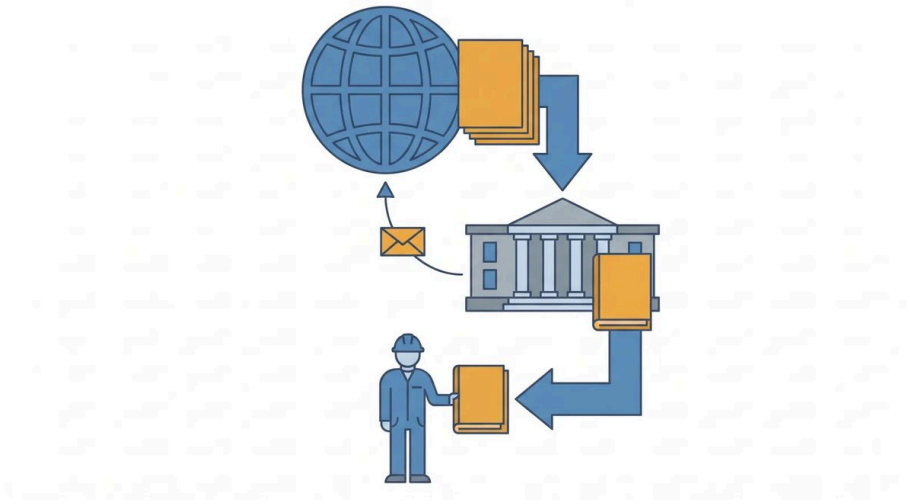
Exam tip: ICAO Annex 8 covers *Airworthiness of Aircraft* — this is frequently tested. Don't confuse it with Annex 1 (Personnel Licensing).

What a contracting State undertakes, and what it does not

ICAO adopts Standards and Recommended Practices; it does not licence engineers, approve maintenance organisations or issue certificates of airworthiness. Those acts belong to States. The word "contracting" in "contracting State" is therefore doing real work: what a State signs up to is the obligation to bring its own law into line, and it is that national law — in Europe, EU regulation — that reaches you. Regulation (EU) 2018/1139 records the mechanism precisely when it defines "international standards and recommended practices" as those "adopted by ICAO in accordance with Article 37 of the Chicago Convention" (Article 3, point (22)).

Article 38 of the Convention is the other half of the mechanism, and the same EU regulation points at it: Member States must use the information the Agency compiles "when implementing their obligations under

Article 38 of the Chicago Convention and when providing to ICAO information under the ICAO Universal Safety Oversight Audit Programme" (Article 90(4)). Article 38 is the duty to notify a difference — to tell ICAO, publicly, where national practice departs from a Standard.



ICAO standards reach you only through the State's own law; a State notifies ICAO back where it departs.

Two things follow, and both matter. First, a filed difference does not make a State non-compliant; it makes the departure visible. Second, and more importantly, a notified difference binds nobody else: another State is not obliged to accept an aircraft or a licence that relies on it. It is also the reason that recognising another State's certificates needs an instrument of its own, which is the subject of the last section of this note.

How ICAO material reaches you through EU law

The link is not merely cultural, it is written into the Basic Regulation. Article 17 of Regulation (EU) 2018/1139 empowers the Commission to adopt the airworthiness implementing acts, and Article 17(2) directs that when it does so it "shall ensure compliance with the essential requirements referred to in Article 9 and take due account of the international standards and recommended practices, in particular those set out in **Annexes 1, 6 and 8** to the Chicago Convention". Those are the three Annexes named in the table above, and this is the reason they are the three that matter to a maintenance engineer: the regulation that produces your rulebook names them.

In two places the incorporation is harder still. Article 9(2) makes noise and emissions compliance a matter of "Amendment 14 to Volume I, Amendment 11 to Volume II, and Amendment 2 to Volume III, all as applicable on 1 January 2024, of Annex 16 to the Chicago Convention" — an ICAO document made binding in EU law by reference to a named amendment. And in Part-T, the Part covering third-country registered aircraft dry leased in by an EU air carrier, point T.A.201 requires that the aircraft "holds a valid certificate of airworthiness issued in accordance with ICAO Annex 8" and is maintained under a programme complying with the requirements of the State of Registry and "the applicable requirements of ICAO Annex 6". There the ICAO Annexes are quoted as the operative standard inside a European rule.

The traffic also runs the other way. Under Article 90 of Regulation (EU) 2018/1139 the Agency assists the Commission in relations with third countries and international organisations, may cooperate with ICAO, and

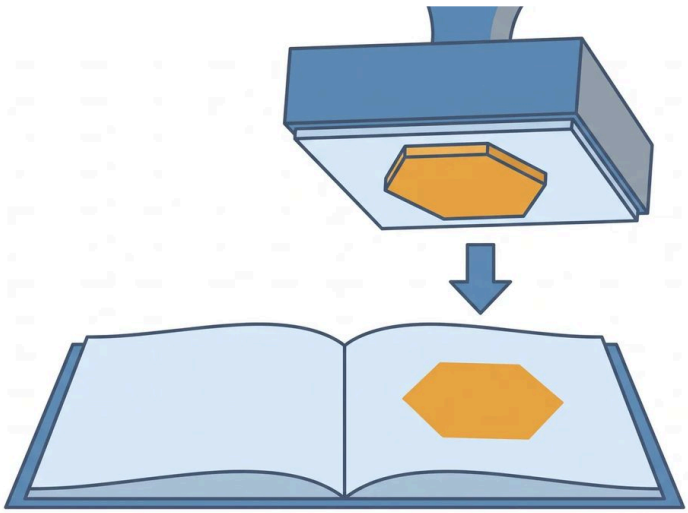
"may act as a Regional Safety Oversight Organisation in the ICAO framework". Where a Member State transfers responsibility for a task attributed to it by the Convention, Article 64 requires it to notify ICAO that the Agency or another Member State is carrying out those functions on its behalf, and states expressly that such reallocations are "without prejudice to the rights and obligations of the Member States under the Chicago Convention". Europe reorganises who does the work; it does not reassign the Convention obligation.

Three further Annexes a maintenance engineer meets

Beyond the three named in Article 17(2), three more Annexes shape work you will actually do — one when something goes wrong, one when an engine or airframe is certificated, and one that explains why your organisation has a safety management system at all.

Annex	Subject	Where it reaches you
Annex 13	Aircraft Accident and Incident Investigation	Independent safety investigation, kept separate from blame and from regulatory enforcement. In the Union the subject is governed by Regulation (EU) No 996/2010, from which Regulation (EU) No 376/2014 takes its definitions of accident, serious incident and incident and its definition of the safety investigation authority.
Annex 16	Environmental Protection	Aircraft noise and engine emissions. Incorporated into EU law by Article 9(2) of Regulation (EU) 2018/1139, and the reason a noise certificate is one of the documents Regulation (EU) No 748/2012 provides for.
Annex 19	Safety Management	Safety management systems and the State-level safety programme. Regulation (EU) 2018/1139 requires each Member State to establish a State Safety Programme (Article 7) and a State Plan for Aviation Safety (Article 8); Regulation (EU) No 376/2014 defines both the State Safety Programme and the safety management system, and the occurrence data you file feeds them.

The pattern is worth naming, because it is the whole point of this section: an ICAO Annex is not something you consult at the bench. You meet it as the shape of a European rule — the reason your licence looks the way it does, the reason a noise certificate exists, the reason your organisation runs a reporting system. If you can say which Annex a European requirement is answering, you understand the requirement rather than merely remembering it.

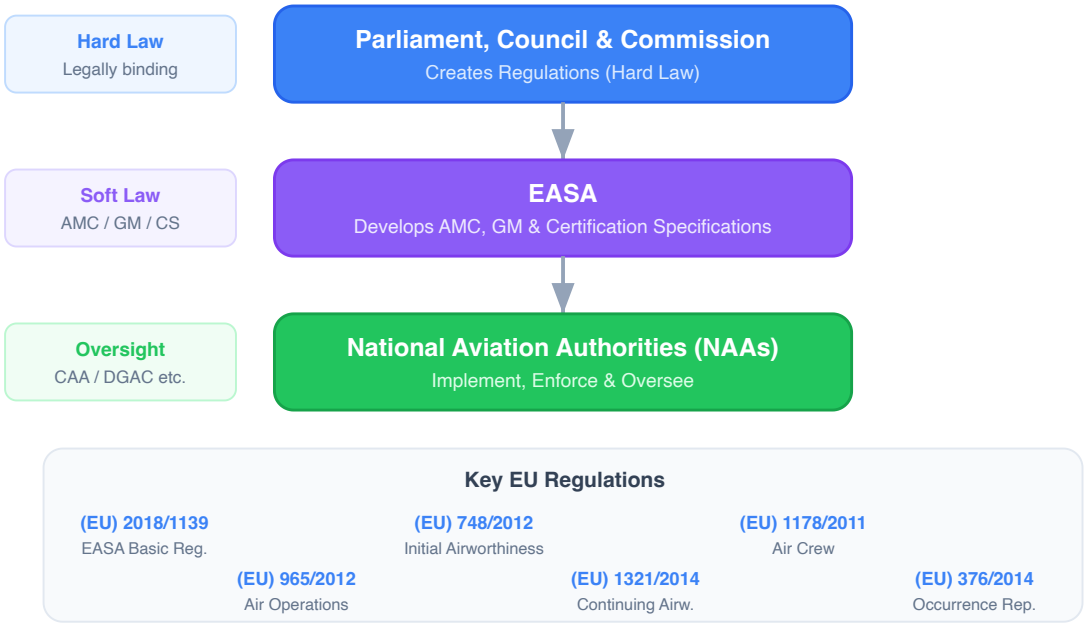


An ICAO Annex reaches you as the shape it gives a European rule, not as a document at the bench.

Direction check: a Standard that a State has not implemented does not reach an operator or an engineer directly, which is why a question offering "legally binding worldwide without any national implementation" is always wrong. And the arrow does not reverse: no European rule can excuse a State from a Convention obligation, it can only change who discharges it.

The European Regulatory Structure

Within Europe, the aviation regulatory framework operates on three levels:



The European Commission

The European Parliament and the Council adopt the basic **regulations** on a proposal from the European Commission, and the Commission adopts the implementing regulations beneath them — these are "hard law", directly applicable and legally binding in all EU Member States without needing national transposition. The key aviation regulations are listed in the diagram above.

EASA — European Union Aviation Safety Agency

EASA is an agency of the European Union, headquartered in **Cologne, Germany**. Established under what is now **Regulation (EU) 2018/1139** (the "EASA Basic Regulation"), EASA's core responsibilities include:

- Drafting implementing rules for the European Commission to adopt
- Publishing **Acceptable Means of Compliance (AMC)** — methods accepted by the Agency as a means of showing compliance with regulations
- Publishing **Guidance Material (GM)** — non-binding explanatory information
- Publishing **Certification Specifications (CS)** — technical standards for type certification (e.g., CS-25 for large aeroplanes)
- Issuing Type Certificates and approvals for aircraft types, parts, and organisations
- Conducting standardisation inspections of NAAs

⚖ Hard Law vs Soft Law

Category	Examples	Binding?
Hard Law	EU Regulations (e.g., 2018/1139, 1321/2014)	Yes — legally binding, directly applicable
Soft Law	AMC, GM, Certification Specifications	No — but AMC provides an accepted means of compliance. Deviation requires an **Alternative Means of Compliance (AltMoC)**

National Aviation Authorities (NAAs)

Each EU Member State has a **National Aviation Authority** (also called "competent authority") responsible for:

- Implementing and enforcing EU aviation regulations within their state
- Issuing, renewing, and revoking Part-66 licences
- Approving Part-145, Part-CAMO, and other organisations
- Conducting oversight and audits
- Investigating violations and taking enforcement action

Examples: DGAC (France), LBA (Germany), AESA (Spain), ENAC (Italy).

Two families of act, and how to tell them apart

The regulations that govern your work were not all made by the same institution or by the same procedure, and knowing which is which explains a great deal about how quickly each one can change. The distinction is

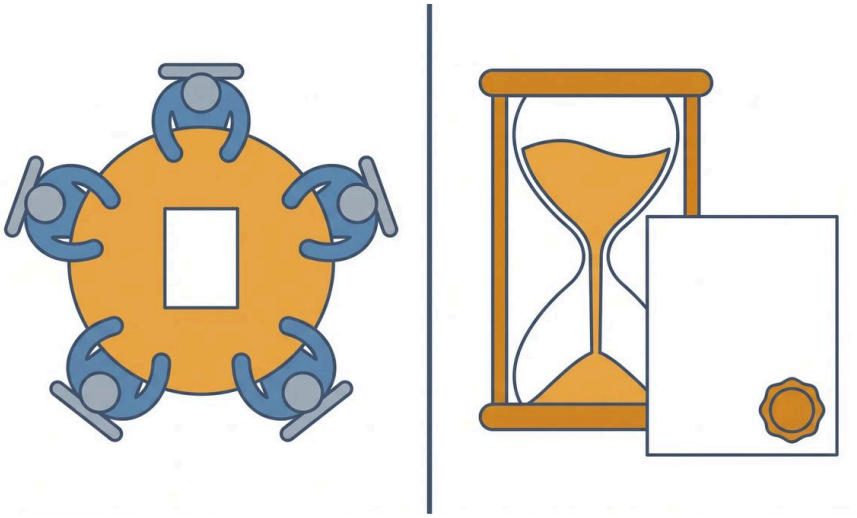
printed on the first line of every act.

Regulation	Adopted by	Legal route
(EU) 2018/1139, of 4 July 2018	the European Parliament and the Council	the basic act; adopted as legislation, and the source of the Commission's powers in the four implementing regulations below
(EU) No 376/2014, of 3 April 2014	the European Parliament and the Council	a separate legislative act on occurrence reporting
(EU) No 748/2012, of 3 August 2012	the European Commission	implementing act adopted on the basis of the basic regulation
(EU) No 1321/2014, of 26 November 2014	the European Commission	implementing act adopted on the basis of the basic regulation
(EU) No 965/2012, of 5 October 2012	the European Commission	implementing act adopted on the basis of the basic regulation
(EU) No 1178/2011, of 3 November 2011	the European Commission	implementing act adopted on the basis of the basic regulation

So the two instruments that set the framework and the reporting duty were adopted by the co-legislators, and the four that carry the working detail were adopted by the Commission alone. The Parliament and the Council are not absent from the second family, but their role there is structural rather than item by item: they wrote the empowerments into the basic regulation, and for one category of Commission act — delegated acts — either of them can revoke the delegation or block an individual act after it has been adopted. That is the mechanism the next two headings describe.

Implementing acts and delegated acts

Regulation (EU) 2018/1139 gives the Commission two different powers, and they are scrutinised at two different moments.



Implementing acts are scrutinised in committee before adoption; delegated acts in an objection window after.

Implementing acts carry the detailed rules and are made under Article 127. Article 127(1) provides that "the Commission shall be assisted by a committee" and that "that committee shall be a committee within the meaning of Regulation (EU) No 182/2011"; Article 127(3) provides that where reference is made to that paragraph, Article 5 of Regulation (EU) No 182/2011 — the examination procedure — applies. The substantive empowerments to adopt implementing acts close with a reference to that procedure. Article 17(1), the empowerment for airworthiness, closes: "Those implementing acts shall be adopted in accordance with the examination procedure referred to in Article 127(3)." Regulation (EU) No 1321/2014 and Regulation (EU) No 748/2012, and every amendment to them, come through that door. **The scrutiny happens before adoption, in the committee.**

Delegated acts are made under Article 128, and the arrangement is deliberately more guarded. Article 128(2) confers the power for a period of five years from 11 September 2018, tacitly extended for identical periods unless the Parliament or the Council opposes not later than three months before the end of each period, and it lists exhaustively the articles that confer it: Articles 19, 28, 32, 39, 47, 54, 58, 61, 62(13), 68(3), 84(4), 105 and 106. Article 128(3) allows either the Parliament or the Council to revoke the delegation at any time, without affecting delegated acts already in force. Article 128(6) provides that a delegated act enters into force only if neither institution objects within two months of notification, a period extendable by a further two months at the initiative of either. **Here the scrutiny happens after adoption, in the objection window.**

Article 128(4) adds one further step on the delegated route only: "Before adopting a delegated act, the Commission shall consult experts designated by each Member State." The Member States are therefore heard on both routes — through the committee on the implementing route, and through designated experts on the delegated one. What differs is who holds the power to stop the act, and when.

A rule from opinion to shop floor: Regulation (EU) 2023/989

The abstract description above is easier to hold on to as a single worked case. Regulation (EU) 2023/989 is the act that rewrote the Part-66 basic training syllabus — including this module — and its own text records every step of its making.

Step	What happened	Where it is recorded
1. Agency proposal	EASA developed the change and issued an opinion to the Commission. The regulation's recitals state that "the amendments are based on Opinion No 07/2022" of the Agency, "in accordance with Article 75(2), point (b), and Article 76(1) of Regulation (EU) 2018/1139".	Recital 6 of (EU) 2023/989
2. Commission drafting	Because the measure comprises technical rules, Article 75(2)(b) of the Basic Regulation applies: the Commission "may not change their content without prior coordination with the Agency".	Article 75(2)(b), (EU) 2018/1139
3. Committee opinion	The draft went to the committee. The regulation records that "the measures provided for in this Regulation are in accordance with the opinion of the committee established by Article 127 of Regulation (EU) 2018/1139".	Recital 11 of (EU) 2023/989
4. Adoption	The Commission adopted the act. Done at Brussels, 22 May 2023.	Signature block of (EU) 2023/989

5. Publication	Official Journal L 135, 23 May 2023, page 53.	Header of (EU) 2023/989
6. Entry into force	"On the twentieth day following that of its publication in the Official Journal of the European Union."	Article 3, (EU) 2023/989
7. Application	"It shall apply from 12 June 2024. However, Article 1(1) and Article 2 shall apply from 12 June 2023."	Article 3, (EU) 2023/989
8. Transition	A basic training course already under way on 12 June 2024 had to be completed, examinations included, before 12 June 2026.	Article 6(4) of (EU) No 1321/2014, as amended

Three points in that sequence repay attention. **Entry into force and application are different dates**, and it is the later one that changes what you must do: this act was in force from mid-June 2023 but did not apply to training until 12 June 2024. **A single act can carry more than one application date**, as this one does for its Article 1(1) and Article 2. And **the transitional provision is written into the amended regulation, not into the amending one** — Article 6(4) to (6) of Regulation (EU) No 1321/2014 — which is where you have to look for it a year later, because the amending act has by then disappeared into the consolidated text.

What the Agency may do, and what it may not

Article 75(2) of Regulation (EU) 2018/1139 lists the Agency's functions. It is to "undertake any task and formulate opinions on all matters covered by this Regulation"; "assist the Commission by preparing measures"; provide the Commission with technical, scientific and administrative support; take the measures within the powers conferred on it; "conduct inspections, other monitoring activities and investigations as necessary to fulfil its tasks under this Regulation, or as requested by the Commission"; carry out, on behalf of Member States, functions ascribed to them by the Chicago Convention; and "assist the national competent authorities in carrying out their tasks, in particular by providing a forum for exchanges of information and expertise".

Article 76 then converts those functions into the outputs the Agency issues, and the distinction between them is exactly the hard-law and soft-law distinction this section is about. The Article runs to seven paragraphs, and each of them names an output:

- **Opinions** — Article 76(1). The Agency assists the Commission in preparing amendments to the Basic Regulation and the delegated and implementing acts made under it, and "the documents that the Agency submits to the Commission for those purposes shall take the form of opinions". An opinion is a proposal, not a rule; it becomes binding only if and when the Commission adopts it.
- **Recommendations** — Article 76(2). The Agency issues recommendations addressed to the Commission for the application of Articles 70 and 71, the safeguard and flexibility provisions set out later in this section. A recommendation carries no obligation of its own: it is the Agency's assessment of a case a Member State has opened, and it is the Commission that decides what follows from it.
- **Certification specifications, acceptable means of compliance and guidance material** — Article 76(3), issued in accordance with Article 115 and with the applicable delegated and implementing acts.
- **Decisions** — Article 76(4), taken for the application of paragraph 6 of this Article, Articles 77 to 83, 85 and 126, and where tasks have been allocated to the Agency under Articles 64 and 65. A decision binds the

person it is addressed to.

- **Reports** — Article 76(5), on the inspections and other monitoring activities the Agency conducts under Article 85.
- **Reaction to an urgent safety problem** — Article 76(6), described below.
- **Opinions on individual flight time specification schemes** — Article 76(7), where a Member State proposes a scheme that departs from the certification specifications the Agency has adopted. That output belongs to the flight and duty time rules for aircrew rather than to maintenance, but it completes the Article.

Reacting to an urgent safety problem

Article 76(6) requires the Agency to react "without undue delay" to an urgent safety problem falling within the scope of the Regulation, and it does so along two different routes according to whether it is the competent authority for the persons concerned. Getting the direction right here is the whole of the point.

- **Where the Agency is the competent authority** — Article 76(6)(a) — it determines the corrective action to be taken by those natural and legal persons and disseminates the related information to them, "including directives or recommendations". It may also issue safety bulletins containing non-binding information or recommendations addressed to other persons involved in aviation activities.
- **Where it is not** — Article 76(6)(b) — it determines safety objectives to be achieved and recommends corrective action to the national competent authorities, disseminating the related information to them. Those authorities must then inform the Agency without undue delay of the measures they have taken, and where the problem affects more than one Member State they must cooperate with the Agency so that the measures are coordinated.

So the Agency does not issue instructions directly to an engineer in a Member State where it is not the competent authority. It sets the objective; the national authority chooses and imposes the measure; and the reporting obligation runs back up from the authority to the Agency, not down.

How the Agency's soft law is produced

Article 115 makes the Management Board responsible for establishing transparent procedures for issuing opinions, certification specifications and other detailed specifications, acceptable means of compliance and guidance material. Those procedures must draw on the expertise of the Member States' civil and, where appropriate, military aviation authorities; involve experts from interested parties or European standardisation bodies where necessary; and "ensure that the Agency publishes documents and widely consults interested parties, in accordance with a timetable and a procedure which includes an obligation on the Agency to give a written response to the consultation process". Article 115(2) requires a procedure for prior consultation of the Member States, for which the Agency may create a working group in which each Member State is entitled to designate an expert. Article 115(3) requires the results to be published in the official publication of the Agency.

The reason this matters to an engineer is that it explains the standing of the document in your hand. An acceptable means of compliance has been consulted on publicly and published formally; it is not an opinion somebody at the Agency holds. But it went through a procedure established by the Management Board, not through the committee under Article 127, and that is precisely why it does not carry the force of the regulation it explains.

The Management Board

Article 99(1) sets the composition: the Management Board "shall be composed of representatives from Member States and from the Commission, all with voting rights". Each Member State appoints one member and two alternates, an alternate representing the member only in his or her absence; the Commission appoints two representatives and their alternates; and the term of office is four years and is extendable. Under Article 99(2) members are appointed for their knowledge and experience in civil aviation and shall have overall responsibility at least for civil aviation safety policy in their own Member State. Article 99(5) adds four members of the Agency's advisory body, participating with observer status and representing as broadly as possible the different views held there. Article 100 provides that the Board elects a Chairperson and a Deputy Chairperson from among its voting members for four years, extendable once.

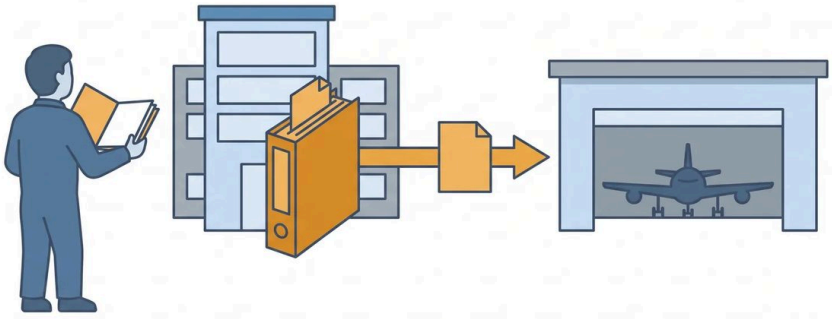
Two features of that composition explain the Agency's character. It is the Member States and the Commission together who hold the votes, so the Agency's procedures are agreed by the same States whose authorities will apply them; and the advisory body, which under Article 98(4) represents "the full range of interested parties affected by the work of the Agency", sits with observer status and no vote.

Standardisation: who inspects whom

Article 85 is titled "Monitoring of Member States", and the title states the direction. The Agency "shall assist the Commission in monitoring the application by the Member States of this Regulation and of the delegated and implementing acts adopted on the basis thereof by conducting inspections and other monitoring activities", and it reports to the Commission on them. Those activities also aim to help Member States apply the rules uniformly and to share best practice.

The powers are real ones. Article 85(2) allows the Agency to require any national competent authority and any natural or legal person subject to the Regulation to provide necessary information and oral explanations, to enter their premises, land and means of transport, and to examine, copy or extract any relevant document, record or data, irrespective of the medium. Article 85(3) requires the Agency to inform the Member State in good time of the subject matter, purpose, start date and the identity of the staff involved, and inspectors act on production of a written authorisation. Article 85(4) obliges the Member State to facilitate the activity. Article 85(7) requires the Agency to publish a summary of information about each Member State's application of the Regulation.

The subject of a standardisation inspection is therefore the authority, not you. But the finding lands on the authority's own procedures, and the authority closes it by changing what it requires of the organisations it oversees — which is how a standardisation finding reaches a hangar without ever naming it.



Standardisation inspects the authority; its changed procedures then reach the hangar it oversees.

Which authority is competent for you

Article 62(1) states the constitutional position: "the Commission, the Agency and the Member States shall cooperate within a single European aviation safety system". Article 62(2) then lists what the Agency and the national competent authorities each do within it — receive and assess applications and issue or renew certificates and receive declarations; perform oversight of certificate holders and of the products and organisations concerned; conduct the necessary investigations, inspections including ramp inspections, audits and other monitoring activities; "take all necessary enforcement measures, including amending, limiting, suspending or revoking certificates issued by them, grounding of aircraft and imposing penalties, in order to terminate identified infringements"; prohibit, limit or condition activities in the interest of safety; and ensure an appropriate level of qualification of their own staff, including by providing adequate training.

Article 62(4) then allocates the work. The Agency is responsible where tasks are attributed to it under Articles 77 to 82 or allocated to it under Articles 64 and 65; the aerodrome's own national competent authority is responsible for aerodrome certificates; and "in all other cases, the national competent authority of the Member State where the natural or legal person applying for the certificate or making the declaration has its principal place of business or, if that person has no principal place of business, where it has its place of residence or place of establishment, shall be responsible for those tasks". Regulation (EU) No 1321/2014 defines "principal place of business" as "the head office or the registered office of the undertaking within which the principal financial functions and operational control of the activities referred to in this Regulation are exercised" — so it is where the business is actually run, not where a letterhead says it is.

Each Part then applies that general rule in its own opening point, and **the tests are not identical.** The differences are easy to miss, because the Part-145 test is the one usually quoted.

Part	Point	The test	Who is competent

Part-145	145.1	the organisation's principal place of business	the authority designated by that Member State, or by another Member State under Article 64 of (EU) 2018/1139, or the Agency if responsibility has been reallocated under Article 64 or 65; and **the Agency** where the principal place of business is outside a territory for which a Member State is responsible under the Chicago Convention
Part-147	147.1	the organisation's principal place of business	the authority designated by that Member State; the Agency for organisations whose principal place of business is in a third country
Part-CAO	CAO.1(1)	the organisation's principal place of business	the authority designated by that Member State; the Agency where it is in a third country
Part-66	66.1(a)	**the Member State to which the person first applies** for the licence	the authority designated by that Member State; another Member State's authority only by agreement with the first, in which case the original licence is revoked, the records under 66.B.20 are transferred and a new licence is issued on the basis of them
Part-T	T.1	**the State that issued the Air Operator Certificate**	the authority designated by the Member State that issued the AOC to the operator

Read the table as three different questions rather than one. For an *organisation*, the question is where the business is run. For a *licence*, it is where you first applied — your licence does not follow you when you move, and changing authority is a deliberate transaction in which the old licence is surrendered, not a notification. For *Part-T*, it is neither: it follows the air operator certificate, so an aircraft on a third-country register is overseen by the State of the operator's AOC and not by its State of Registry.

⚠ **Exam trap:** "the Member State where the organisation has its principal place of business" is the right answer for a Part-145 organisation, but only because that organisation is inside the Union. Point 145.1 has a second limb, and it points the other way: where the principal place of business lies outside a territory for which a Member State is responsible under the Chicago Convention, the competent authority is **the Agency**. A maintenance organisation in a third country holding an EASA Part-145 approval is overseen by EASA directly, not by any national authority.

Independence, conflict of interest and the authority's own powers

Article 62(3) requires Member States to "ensure that their national competent authorities are independent when taking technical decisions on certification, oversight and enforcement and exercise their tasks impartially, and transparently and are organised, staffed and managed accordingly", and to ensure they have the necessary resources and capabilities to do the work efficiently and in a timely manner. Independence here means independence in the technical decision, which is a narrower and more useful idea than institutional separation: it is the reason an inspector's finding is not negotiable through commercial channels.

Article 7a of Regulation (EU) No 1321/2014 fills in the practical detail. Where a Member State designates more than one entity as competent authority, the areas of competence of each must be clearly defined, in

particular as to responsibilities and geographic limitations, and coordination must be established between them. Member States must ensure that competent authority personnel do not perform certification and oversight activities where there are indications that this could result, directly or indirectly, in a conflict of interest, "in particular when relating to family or financial interest". And the authority is empowered to examine records, data, procedures and any other relevant material, make copies or extracts, ask for an oral explanation on site from any of the organisation's personnel, enter relevant premises, operating sites or means of transport, "perform audits, investigations, assessments, inspections, including unannounced inspections", and take or initiate enforcement measures.

Article 62(6) sets the shape of that oversight: it "shall be continuous and based on priorities set in the light of the risks to civil aviation". Oversight is therefore not a fixed cycle of visits; a poor finding history is itself a reason for more attention. Article 62(11) provides the counterweight for organisations: any person subject to the Regulation may bring alleged differences in the application of the rules between Member States to the Agency's attention, and where those differences seriously hamper their activities the Agency and the authorities concerned must cooperate to eliminate them, with the Agency presenting the matter to the Commission if they cannot.

Moving oversight to the Agency

Article 65 lets an organisation ask the Agency to become its competent authority. The condition is a real one and worth stating precisely: the organisation holds, or is eligible to apply for, a certificate from one Member State's national competent authority, "but it has or it intends to have a substantial proportion of facilities and personnel covered by that certificate located in one or more other Member States". Two or more organisations in a single business grouping, each with its principal place of business in a different Member State and each holding or eligible for a certificate for the same type of activity, may make the request together.

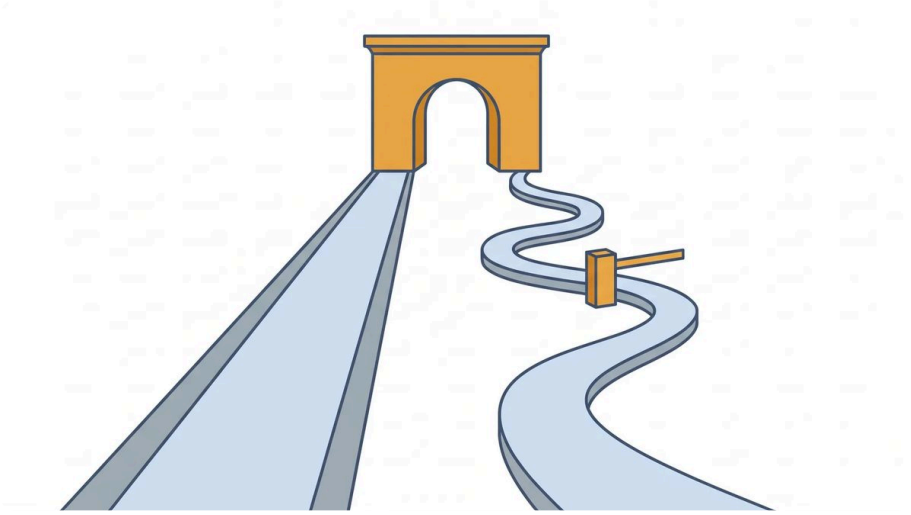
The sequence, and the direction of each step, is:

- The organisations send the request to the Agency *and* to the national competent authorities of the Member States where they have their principal places of business.
- The Agency and those authorities consult one another without undue delay, and may seek supplementary views from the organisations.
- If the Agency or any authority concerned considers the request would adversely affect its own ability to conduct certification, oversight and enforcement, it must inform the organisations, with justification, **within a maximum of 180 days** of receiving the request.
- Otherwise the Agency and the authorities conclude detailed arrangements identifying the tasks being reallocated and the legal, practical and administrative arrangements for an orderly transfer; the organisations are consulted on those arrangements before they are finalised.
- On conclusion of the arrangements the Agency becomes the competent authority for the tasks covered, and the Member States concerned are relieved of responsibility for them.

Article 65(6) makes the move reversible: an organisation for which the Agency acts as competent authority may ask the Member States where it has its principal places of business to resume responsibility. Note who initiates in each direction — the organisation asks in both cases; neither the Agency nor a national authority can move an organisation's oversight unilaterally under this Article.

Acceptable means of compliance in practice

An acceptable means of compliance is not a requirement; it is a route through one that the Agency has already accepted. Following it is the simplest way through an audit, because the argument that the route works has already been had. Departing from it is entirely lawful — but the burden of showing that your route achieves compliance moves onto you, and the rules set out how you discharge it.



Acceptable and alternative means of compliance are two routes to one requirement; neither bypasses it.

All three instruments are formally defined in EU law, and the definitions repay reading word for word. Article 2 of Regulation (EU) No 1178/2011 defines an acceptable means of compliance as "non-binding standards adopted by the Agency to illustrate means to establish compliance"; an alternative means of compliance as "those means that propose an alternative to an existing AMC or those that propose new means to establish compliance ... for which no associated AMC have been adopted by the Agency"; and certification specifications as "technical standards adopted by the Agency indicating means to be used by an organisation for the purpose of certification". The first two of those definitions still name Regulation (EC) No 216/2008 as the act being complied with; the definition of a certification specification names no act at all. Where the repealed regulation is named, Article 139(4) of Regulation (EU) 2018/1139 directs that the reference be read as a reference to the current act, by the mechanism described later in this note.

Two things fall out of that wording. An alternative means of compliance has a second limb that is easy to miss: it covers not only proposing an alternative to an existing acceptable means, but also proposing a means where the Agency has published none at all. And all three definitions describe standards that *illustrate* or *indicate* a means of compliance — never standards that impose a requirement of their own.

Point 145.A.120 states the mechanism for a maintenance organisation. "An organisation may use any alternative means of compliance to establish compliance with this Regulation." If it wishes to do so, then "prior to using it" it must "provide the competent authority with a full description", including any revisions to manuals or procedures that may be relevant and an explanation of how compliance is achieved; and "the organisation may use those alternative means of compliance subject to prior approval from the competent authority". Points CAMO.A.120 and CAO.A.017 carry the same requirement for a continuing airworthiness management

organisation and a combined airworthiness organisation, adding that use begins on receipt of the authority's notification.

The direction is fixed and is tested: description first, approval second, use third. An alternative means of compliance used before approval is not an alternative means of compliance; it is a non-compliance.

The approved alternative then becomes a permanent, visible feature of the organisation. Point 145.A.70(a)(17) requires the maintenance organisation exposition to include "a list of the currently approved alternative means of compliance used by the organisation", and point CAMO.A.300(a)(14) requires the same list in the continuing airworthiness management exposition. On the authority's side, point 145.B.120 — and points 66.B.2 and CAMO.B.120 in the same terms — provides that the Agency develops the acceptable means of compliance, that alternative means may be used, and that "competent authorities shall inform the Agency of any alternative means of compliance used by organisations under their oversight or by themselves". Point CAMO.B.120 adds that the authority must establish a system to evaluate them consistently, must evaluate each one proposed by analysing the documentation and, if it considers it necessary, inspecting the organisation, and on finding it compliant must without undue delay notify the applicant that it may be implemented. Point 145.B.220 requires the authority's record-keeping system to hold the documents supporting the use of alternative means of compliance.

⚠ **Exam trap:** an alternative means of compliance is an alternative *to the acceptable means of compliance*, never to the regulation. Point CAMO.A.120(a) says so explicitly: "Alternative means of compliance **to the AMC adopted by the Agency** may be used by an organisation to establish compliance with Regulation (EU) 2018/1139 and its delegated and implementing acts." No approval of an alternative means of compliance lets an organisation depart from a requirement in the regulation itself — only from the Agency's suggested way of meeting it.

When a certification specification becomes binding

Certification specifications sit in the same soft-law family, but they acquire binding force by a route of their own, and it is worth following once. Point 21.B.80 of Annex I (Part 21) to Regulation (EU) No 748/2012 provides that "the Agency shall establish the type-certification basis and notify it to the applicant for a type-certificate or a restricted type-certificate", and that the type-certification basis "shall consist of the certification specifications for airworthiness designated by the Agency for the product ... as applicable on the date of application for that certificate and any special condition prescribed by the Agency", subject to stated exceptions. Point 21.A.20(a) then requires the applicant to demonstrate compliance with that type-certification basis as established and notified.

So the specification itself is not a rule of general application; but once the Agency has designated it as part of a particular product's certification basis, and the applicant has been notified, compliance with it is mandatory *for that certification*. That is the sense in which certification specifications are, as the standard formulation goes, generally non-binding but binding when incorporated into a certification basis. Point 21.B.80(a)(2) and (3) also show where the flexibility lies: the Agency may accept an alternative to a designated specification that cannot be complied with where compensating factors provide an equivalent level of safety, or accept or prescribe other means of demonstrating compliance with the essential requirements.

Two escape valves, and they are not the same

Beyond the alternative means of compliance, the Basic Regulation provides two further ways of departing from what the rules say, and they are frequently confused because both end with the Commission and an implementing act. They start in different places and answer different problems.

	Flexibility provisions	**Safeguard provisions**
Article	Article 71 (and Article 76(4) for the Agency)	Article 70
The problem	urgent unforeseeable circumstances affecting a person, or that person's urgent operational needs	a problem involving a serious risk to aviation safety requiring immediate action by the Member State
Who acts	a Member State grants an exemption to a natural or legal person; the Agency may grant the same to a person it has certificated	the Member State itself takes a measure
What may be set aside	requirements applicable under Chapter III, **other than the essential requirements**, and the delegated or implementing acts made on the basis of that Chapter	nothing is set aside; the Regulation does not prevent an immediate reaction where the conditions are met
Conditions	the circumstances cannot be addressed in compliance with the requirements; safety, environmental protection and compliance with the essential requirements are ensured, where necessary through mitigation; distortion of market conditions is mitigated as far as possible; and the exemption is limited in scope and duration to what is strictly necessary and applied without discrimination	the problem involves a serious risk and immediate action is required; the Member State cannot adequately address it in compliance with the Regulation; and the action is proportionate to the severity of the problem
Notification	immediately to the Commission, the Agency and the other Member States through the repository established under Article 74	immediately to the Commission, the Agency and the other Member States through the same repository
Review trigger	a duration exceeding eight consecutive months, or the same exemption granted repetitively to a total exceeding eight months	immediate: the Agency assesses without undue delay on receiving the notification
Outcome if the conditions were not met	the Commission adopts an implementing act; on notification the Member State must immediately revoke the exemption	the Commission adopts an implementing act; on notification the Member State must immediately revoke the measure

There is a third route, and it is the constructive one. Article 71(3) provides that where a Member State considers that compliance with the essential requirements can be demonstrated by means other than those laid down in the delegated and implementing acts, and that those means "present significant advantages in terms of civil aviation safety or of efficiency", it may submit a reasoned request for **amendment of the delegated or implementing act itself**. The Agency then issues a recommendation to the Commission on whether the request meets the conditions, and the Commission considers amending the act. An exemption sets a rule aside for one person for a limited time; an Article 71(3) request changes the rule for everybody.

Point 145.B.220 requires the competent authority's records to include the use of the safeguard and flexibility provisions under Articles 70, 71(1) and 76(4), so none of these routes is informal.

Regulation (EU) 2018/1139 — The EASA Basic Regulation

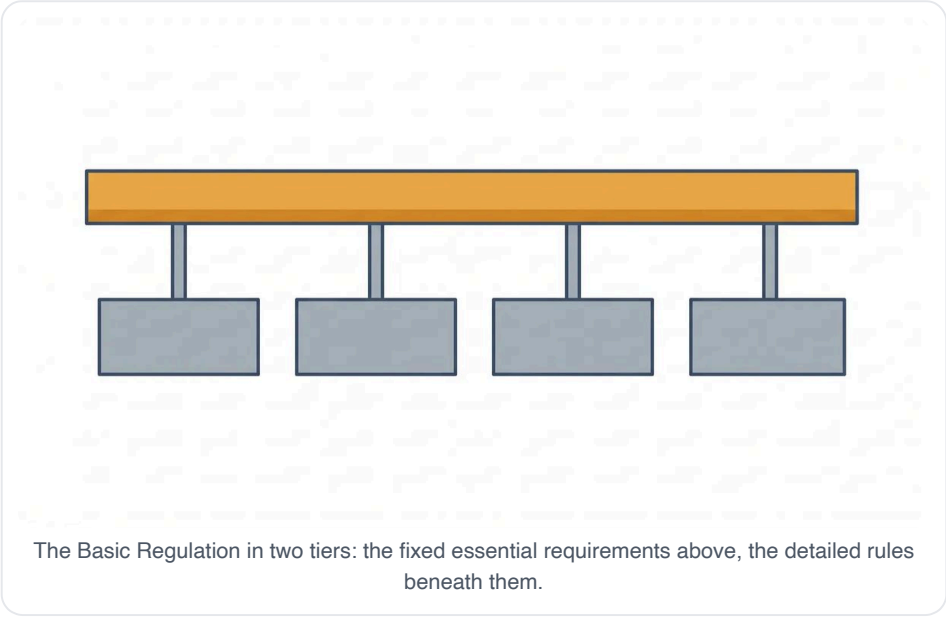
This is the **cornerstone** of EU aviation safety law. It replaced the previous Regulation (EC) 216/2008 and establishes:

- EASA's mandate and powers
- Essential requirements for airworthiness, environmental protection, crew licensing, air operations, aerodromes, and ATM/ANS
- Framework for implementing rules (delegated and implementing acts)
- Ramp inspections and third-country operators

⚠️ **Exam tip:** The Basic Regulation is (EU) 2018/1139. It replaced the older (EC) 216/2008. Questions often test whether you know the correct regulation number.

Two tiers: essential requirements and the detail beneath them

The Basic Regulation is built in two tiers, and understanding the join explains why the detailed rules can be rewritten every couple of years while the framework stays still.



The Basic Regulation in two tiers: the fixed essential requirements above, the detailed rules beneath them.

The upper tier is the essential requirements, and they sit in the Regulation's own Annexes rather than in its articles. Article 9(1) provides that aircraft and "their engines, propellers, parts and non-installed equipment shall comply with the essential requirements for airworthiness set out in Annex II to this Regulation". Article 9(2) covers noise and emissions by reference to Annex 16 to the Chicago Convention, with the essential requirements for environmental compatibility in Annex III applying to the extent the Convention provisions do not contain environmental protection requirements, and it adds that organisations involved in the design, production and maintenance of those products "shall comply with point 8 of Annex III".

The lower tier is the detail, and Article 17(1) is the empowerment that produces it: "in order to ensure the uniform implementation of and compliance with the essential requirements referred to in Article 9", the Commission adopts implementing acts laying down detailed provisions on, among other things, the rules and procedures for issuing, maintaining, amending, limiting, suspending or revoking the approvals and the licences; "the privileges and responsibilities of the holders of the approvals and licences"; "the rules and procedures for the maintenance of products, parts and non-installed equipment"; and "the rules and procedures for the continuing airworthiness management of aircraft". Those items map onto the Parts you work to: organisation approvals (Part-145, Part-CAMO, Part-CAO), personnel licences (Part-66), the privileges attaching to each, the maintenance rules, and the continuing airworthiness management rules. Regulation (EU) No 1321/2014 is Article 17(1) made concrete.

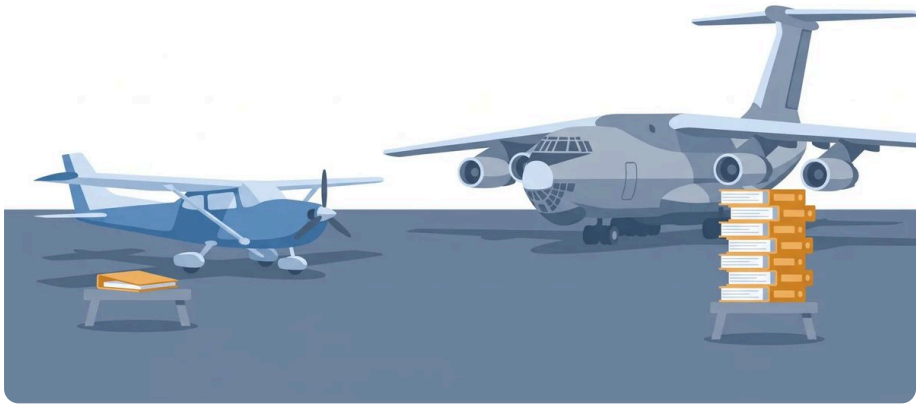
The direction between the tiers matters. An implementing act is measured against the essential requirements, not the other way round: amending Part-145 cannot change what Annex II to the Basic Regulation requires, and Article 17(2) obliges the Commission to "ensure compliance with the essential requirements referred to in Article 9" when it adopts the detail.

Article 4: the shape every measure must take

Article 4 binds the Commission, the Agency and the Member States whenever they take a measure under the Regulation. They must reflect the state of the art and best practice and take account of worldwide aviation experience and technical progress; build on the best available evidence and analysis; allow for immediate reaction to established causes of accidents, serious incidents and intentional security breaches; take account of interdependencies between aviation safety, cyber security and other technical domains; lay down requirements "in a manner which is performance-based and focuses on objectives to be achieved, while allowing different means of achieving compliance" where possible; promote efficient use of resources between authorities; take non-binding measures where possible; and take account of international rights and obligations, including those under the Chicago Convention.

Article 4(2) then requires the measures to "correspond and be proportionate to the nature and risk of each particular activity", taking into account, as appropriate, whether persons other than flight crew are carried and whether the operation is open to the public; the extent to which third parties or property on the ground could be endangered; "the complexity, performance and operational characteristics of the aircraft involved"; the purpose of the flight and the airspace used; "the type, scale, and complexity of the operation or activity"; the extent to which those exposed can assess and control the risks; and "the results of past certification and oversight activities".

Two features of your own working life follow visibly from that article. The lighter regimes — Part-ML for light aircraft continuing airworthiness, Part-CAO for combined organisations, Part 21 Light for design and production of sporting and recreational aircraft — have the shape Article 4(2)(c) and (e) call for, matched to the complexity of the aircraft and the scale of the activity rather than applied uniformly. And Article 4(2)(g), the results of past certification and oversight, is the same principle that Article 62(6) states operationally when it makes oversight risk-based: an organisation's own record is an input to how closely it is watched.



Rules are proportionate: a light, simple aircraft gets a lighter regime than a large, complex one.

What happened to existing certificates when the old regulation was repealed

Article 139(1) repealed Regulation (EC) No 216/2008 "with effect from 11 September 2018". A repeal of that size raises an obvious question for anybody holding a licence or an approval issued under the old regime, and Article 140 answers it.

Article 140(1) provides that the certificates and specific airworthiness specifications issued or recognised, and the declarations made or recognised, in accordance with Regulation (EC) No 216/2008 and its implementing rules "shall continue to be valid and shall be deemed to have been issued, made and recognised pursuant to the corresponding provisions of this Regulation". A Part-66 licence issued in 2015 therefore did not need reissuing, and neither did a Part-145 approval. Nothing had to be done by the holder.

The other transitional provisions are worth knowing because they explain documents you will still meet. Article 140(2) required the implementing rules adopted under the old regulations to be adapted to the new one by 12 September 2023, and set out how to read four defined terms in the meantime. Article 139(4) provides that references to the repealed regulations "shall be construed as references to this Regulation and, where appropriate, read in accordance with the correlation table in Annex X". So an older maintenance organisation exposition, training manual or contract that cites Regulation (EC) No 216/2008 is not thereby wrong: it is read across. It is, however, a reliable sign that the document has not been reviewed for some years, which is a different and usually more interesting finding.

Two of the four implementing regulations in this note still show their age on their title page. Regulation (EU) No 965/2012 and Regulation (EU) No 1178/2011 are each described as "laying down technical requirements and administrative procedures ... pursuant to Regulation (EC) No 216/2008", because both were adopted before the Basic Regulation was replaced. They are current law; the citation in their titles is simply read through Article 139(4).

What the Basic Regulation does not do

It is as useful to know where this regulation stops. It does not itself certify anything: no aircraft, organisation or person is approved by Regulation (EU) 2018/1139. It creates the Agency (Article 75(1)), states the objectives, sets the essential requirements, allocates certification, oversight and enforcement between the Agency and

the national authorities, and empowers the Commission to make the detailed rules. Every certificate you will ever hold or rely on is issued under one of those detailed rules by one of those authorities. When a question asks what the Basic Regulation "establishes", the answer is always a framework, a power or an obligation — never an individual approval.

Key Implementing Regulations

Regulation	Subject	Contains
(EU) No 748/2012	Initial Airworthiness	Part 21 — design, production, certification of aircraft and parts
(EU) No 1321/2014	Continuing Airworthiness	Part-M, Part-145, Part-66, Part-147, Part-ML, Part-CAMO, Part-CAO, Part-T
(EU) No 376/2014	Occurrence Reporting	Mandatory and voluntary reporting, timelines, just culture
(EU) No 965/2012	Air Operations	Commercial and non-commercial flight operations rules
(EU) No 1178/2011	Air Crew	Pilot licensing, medical certification
(EU) 2023/989	Amendment to 1321/2014	Updated Part-66 syllabus (effective 12 June 2024)

Regulation (EU) No 748/2012: Part 21 and Part 21 Light

Article 1(1) of Regulation (EU) No 748/2012 states that it lays down, "in accordance with Articles 19, 58 and 62 of Regulation (EU) 2018/1139", the common technical requirements and administrative procedures for the airworthiness and environmental protection certification or declaration of compliance of products, parts and appliances, and specifies among other things: the issue of type certificates, restricted type certificates, supplemental type certificates and changes to them; the issue of certificates of airworthiness, restricted certificates of airworthiness, permits to fly and **authorised release certificates**; the issue of repair design approvals; the showing of compliance with environmental protection requirements; the issue of noise certificates; the certification of design and production organisations; and **the issue of airworthiness directives**.

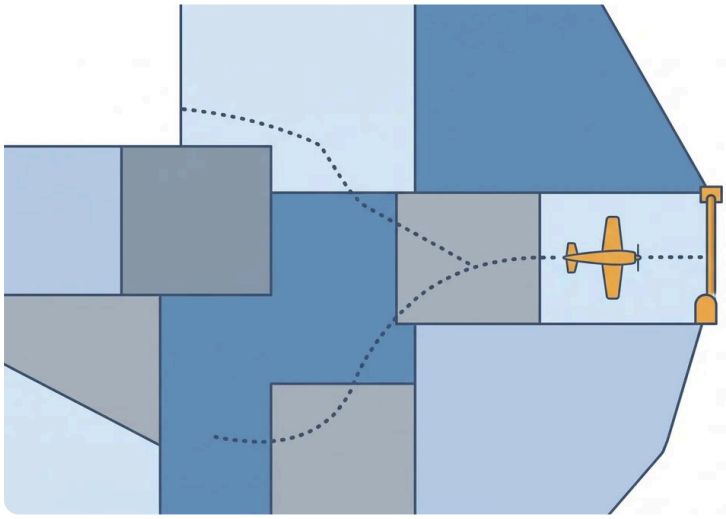
Three items on that list are the ones a maintenance engineer meets most often without necessarily realising which regulation produced them: the authorised release certificate you fill in or receive with a component, the airworthiness directive you embody, and the repair design approval behind a repair scheme. All three originate here, in initial airworthiness, and not in the continuing airworthiness regulation you work to.

Article 1(2) defines the two Annexes. "Part 21" means the requirements and procedures for the certification of aircraft and related products, parts and appliances, and of design and production organisations, laid down in Annex I. "Part 21 Light" means the requirements and procedures for the certification or declaration of design compliance of aircraft **other than unmanned aircraft intended primarily for sports and recreational use** and of related products and parts, and for the declaration of design and production capability of organisations, laid down in Annex Ib.

The relationship between them is one of rule and derogation, and the direction runs one way only. Article 2(1) states the rule: products, parts and appliances "shall be issued with certificates as specified in Annex I (Part 21)". Article 2(2) opens the alternative: "by way of derogation from paragraph 1, certificates may be alternatively issued as specified in Annex I (b) (Part 21 Light)" for a closed list of products — among them an aeroplane of 2 000 kg maximum take-off mass or less with a maximum operational seating configuration of four persons, a sailplane or powered sailplane of 2 000 kg or less, a balloon, a rotorcraft of 1 200 kg or less with a maximum operational seating configuration of four persons, a gyroplane, and a piston engine or fixed pitch propeller intended for installation on one of those. Article 2(3) opens a narrower alternative again: for a shorter list, including an aeroplane of 1 200 kg or less that is not jet-powered with a maximum operational seating configuration of two persons, "a declaration of design compliance may alternatively be made".

Work the direction through with two concrete cases. A large transport aeroplane appears in none of those lists, so only Part 21 is available to it; there is no lighter route and no declaration. A 1 900 kg four-seat piston aeroplane appears in the Article 2(2) list, so its designer may choose a Part 21 type certificate or a Part 21 Light certificate; a 1 100 kg two-seat piston aeroplane appears in the Article 2(3) list as well, and because that paragraph derogates from both of the others, all three routes are open to it. **Eligibility for the lighter route never removes the heavier one; it only adds a choice.** Article 2a even allows movement between them after the event: a holder of a valid type certificate or supplemental type certificate issued under Part 21 could, until 25 August 2025, ask the Agency to maintain the approved type design under Part 21 Light instead, provided the product falls within Article 2(2).

The choice has a consequence that shows up on the aircraft's paperwork. The restricted certificate of airworthiness printed in Annex Ib carries, as a printed restriction, the statement that it "is issued on the basis of a declaration of design compliance made in accordance with Regulation (EU) No 748/2012 and is valid and recognised in all EU Member States without further requirements or evaluation", but that it "does not comply with all of the applicable Standards of Annex 8 to the Convention on International Civil Aviation and therefore may not be valid for international air navigation over non-EU Member States, unless approved by" the State concerned. The lighter route is a genuine simplification inside the Union and a genuine limitation outside it — which is the ICAO Annex 8 point from the opening of this note, made visible on a certificate.



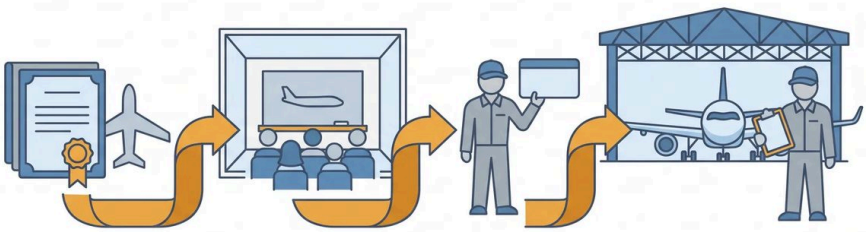
A restricted certificate of airworthiness: recognised in all Member States, possibly not valid beyond them.

Operational suitability data: where your type course comes from

One definition in Regulation (EU) No 748/2012 reaches directly into a maintenance engineer's career. Article 1(2)(k) defines operational suitability data as data "which is part of an aircraft type certificate, restricted type certificate or supplemental type certificate", consisting of the minimum syllabus of pilot type rating training; the scope of aircraft validation source data for simulator qualification; **"the minimum syllabus of maintenance certifying staff type rating training, including determination of type rating"**; the determination of type or variant and type-specific data for cabin crew; and the master minimum equipment list.

So the minimum content of your aircraft type course is fixed by the aircraft's own type certificate, established during initial certification by the design organisation and the Agency. It is not chosen by the training organisation that delivers the course, and it does not vary between training organisations. Regulation (EU) No 1321/2014 closes the loop from the other side: Article 6(3) required type training courses approved before the operational suitability data minimum syllabus was approved for the relevant type to include the relevant elements of the mandatory part of that data "not later than 18 December 2017 or within two years after the operational suitability data was approved, whichever is the latest".

This is the clearest single example in Module 10 of the way the Parts interlock. A decision taken during type certification, under Part 21 in Regulation (EU) No 748/2012, determines the syllabus of a course delivered under Annex IV (Part-147), which qualifies a person licensed under Annex III (Part-66), so that they may release an aircraft maintained under Annex II (Part-145) — and those three Annexes all belong to a single second instrument, Regulation (EU) No 1321/2014. Four Parts, two regulations.



A type certificate decision fixes the type course that qualifies the engineer who releases the aircraft.

Regulation (EU) No 965/2012 and the continuing airworthiness seam

Regulation (EU) No 965/2012 lays down detailed rules for air operations with aeroplanes and rotorcraft, including ramp inspections of aircraft of operators under the safety oversight of another State when they land in the territory of the Member States (Article 1(1)). It also sets the conditions for issuing, maintaining, amending, limiting, suspending or revoking the certificates of operators engaged in commercial air transport, and their privileges and responsibilities (Article 1(2)); the conditions and procedures for declarations by operators engaged in commercial specialised operations or in non-commercial operations of complex motor-

powered aircraft (Article 1(3)); and the conditions under which certain high-risk commercial specialised operations require an authorisation (Article 1(4)).

Its Annexes divide the subject by who is being regulated and what kind of operation is being flown:

Annex	Part	Subject
Annex I	—	Definitions for terms used in the other Annexes
Annex II	Part-ARO	Authority requirements for air operations
Annex III	Part-ORO	Organisation requirements for air operations
Annex IV	Part-CAT	Commercial air transport operations
Annex V	Part-SPA	Specific approvals
Annex VI	Part-NCC	Non-commercial air operations with complex motor-powered aircraft
Annex VII	Part-NCO	Non-commercial air operations with other-than-complex motor-powered aircraft
Annex VIII	Part-SPO	Specialised operations

The relation to continuing airworthiness is precise, and it is a division of labour rather than an overlap. **Regulation (EU) No 965/2012 tells the operator that it must have the function; Regulation (EU) No 1321/2014 says what the organisation performing that function must be.** Point ORO.AOC.135(a) requires the operator to nominate persons responsible for the management and supervision of flight operations, crew member training, ground operations, and "continuing airworthiness or for the continuing airworthiness management contract in accordance with Regulation (EU) No 1321/2014, as the case may be". Point M.A.201(e) of Part-M then states the substance from the other side: for aircraft used by air carriers licensed in accordance with Regulation (EC) No 1008/2008, the operator is responsible for the continuing airworthiness of the aircraft it operates and must take the necessary steps to be approved as a continuing airworthiness management organisation under Part-CAMO as part of its air operator certificate, and to be approved under Part-145 or to conclude a written contract with an organisation that is.

Point ORO.GEN.310(d) handles the case where an aircraft on one operator's AOC is used by another: its continuing airworthiness "shall be managed by the organisation responsible for the continuing airworthiness of the aeroplane or helicopter included in the AOC, in accordance with Regulation (EU) No 1321/2014". The management does not follow the user; it stays with the AOC.

Leasing: the clearest place the two regulations meet

Point ORO.AOC.110 is worth knowing in detail, because it is the boundary case in which air operations, continuing airworthiness and the ICAO Annexes are all in play at once. Its opening rule is that "any lease agreement concerning aircraft used by an operator certified in accordance with this Part shall be subject to prior approval by the competent authority", and that an operator shall not lease in aircraft registered in a State whose operators are subject to an operating ban, or from an operator subject to such a ban.

For a **wet lease-in** from a third-country operator, point ORO.AOC.110(c) requires the applicant to demonstrate to the competent authority that the third-country operator holds a valid AOC issued in accordance with Annex 6 to the Chicago Convention; that its safety standards for continuing airworthiness and air operations are

equivalent to the applicable requirements of Regulation (EU) No 1321/2014 and Regulation (EU) No 965/2012; and that the aircraft has a standard certificate of airworthiness issued in accordance with Annex 8 to the Chicago Convention.

For a **dry lease-in** of an aircraft registered in a third country, point ORO.AOC.110(d) requires the applicant to demonstrate that an operational need has been identified which cannot be satisfied by leasing an EU-registered aircraft; that the duration does not exceed seven months in any 12 consecutive month period; that compliance with the applicable requirements of Regulation (EU) No 1321/2014 is ensured; and that the aircraft is equipped in accordance with the EU air operations regulations.

That third condition is where Part-T comes in. Regulation (EU) No 1321/2014 applies to the continuing airworthiness of third-country registered aircraft dry leased in by a licensed air carrier, and Article 3(6) requires operators to ensure it in accordance with Annex Va (Part-T). Point T.1 then makes the competent authority "the authority designated by the Member State that has issued the Air Operator Certificate to the operator". Follow the whole case through and the arrows are unambiguous: the operator's own authority approves the lease, the operator's own authority oversees the aircraft's continuing airworthiness under Part-T, and the aircraft's State of Registry remains responsible for the certificate of airworthiness the aircraft carries.

Regulation (EU) No 1178/2011, and one word that catches people out

Regulation (EU) No 1178/2011 is the aircrew regulation. Article 1(1) lists what it covers: pilot licences and their ratings and the conditions for issuing, maintaining, amending, limiting, suspending or revoking them; the certification of persons providing flight training or flight simulation training and assessing pilots' skills; medical certificates and the certification of aero-medical examiners; the aero-medical assessment of cabin crew and the conditions for cabin crew attestations; the certification of pilot training organisations and aero-medical centres; the certification of flight simulation training devices and the organisations operating them; and the administration and management system requirements for the Member States, the Agency and those organisations.

None of that is about maintenance staff. It appears in a maintenance module because it completes the family of implementing regulations built on the Basic Regulation, and because it fixes a term that is otherwise easy to misuse. In EU law, **"Approved Training Organisation (ATO)" is the Regulation (EU) No 1178/2011 term for an organisation entitled to provide training to pilots**; its Article 2 defines it that way, and the organisation requirements sit in Annex VII (Part-ORA). The term for an organisation approved to deliver Part-66 basic and type training is *maintenance training organisation*, and that is the term Annex IV (Part-147) uses throughout. The two are different approvals under different regulations for different people.



Pilots train at an ATO; Part-66 trainees at a maintenance training organisation: two different approvals.

The overlap that does exist is operational. Point M.A.201(f) of Part-M carves out, from a set of operator obligations, aircraft used for commercial air transport operations performed "by commercial Approved Training Organisations ('ATO') and Declared Training Organisations ('DTO') referred to in Article 10a of Regulation (EU) No 1178/2011" — so a flying school's aircraft is treated differently from an airline's, and to read that point at all you have to know which regulation those two acronyms come from.

Regulation (EU) 2023/989: what it changed and when it bit

Regulation (EU) 2023/989 does two distinct jobs, and its own structure separates them. Article 1 *amends* Regulation (EU) No 1321/2014, with Annex I of the amending act rewriting parts of Annex III (Part-66) and Annex II rewriting parts of Annex IV (Part-147). Article 2 *corrects* the same regulation, with Annexes III and IV correcting Annex I (Part-M) and Annex Vb (Part-ML).

The correction is instructive on its own. The recitals record that an earlier amending regulation, Implementing Regulation (EU) 2022/1360, "inadvertently deleted point (3) of point M.A.302(d) of Annex I (Part-M) ... and points (3) to (9) of point ML.A.302(c)", which should have been retained, and inadvertently added a duplicate point to M.A.502. A published regulation can contain a drafting error; the error is undone by another regulation, and until that one is adopted the defective text is the law. It is a useful corrective to the assumption that regulation is infallible, and it is a good reason to keep the amendment markers in a consolidated text in view.

The dates matter more than the mechanism. Article 3 provides that the regulation enters into force on the twentieth day following publication, that "it shall apply from 12 June 2024", and that Article 1(1) and Article 2 apply from 12 June 2023. Three provisions inserted into Regulation (EU) No 1321/2014 handle anybody caught mid-course. Article 6(4) sets the finish date for a basic training course, or any part of one, commenced before 12 June 2024, and requires the certificates of recognition to be issued before then too. Article 6(5) provides that those certificates are issued in accordance with the version of the Regulation applicable before 12 June 2024. And Article 6(6) requires the competent authority, for a licence issued or changed after 12 June 2024, to accept an applicant's basic knowledge examination status corresponding to the earlier version as meeting the requirements of the later one — so a candidate part-way through the module examinations did not lose the modules already passed.

For this module specifically, the amendment replaced a seven-paragraph Module 10 with the ten-paragraph structure you are studying. The paragraphs added were **10.4 Independent certifying staff**, **10.8 Oversight principles in continuing airworthiness** and **10.10 Cybersecurity in aviation maintenance**. Paragraph 10.1, this paragraph, is examined at level 1 for category A and at level 1 for categories B1, B2, B2L and B3. Across the whole of Module 10 only two paragraphs are examined differently for category A than for the others: 10.4, which carries a dash for category A and level 3 for the rest, and 10.9, which carries a dash for category A and level 1 for the rest.

Reading a syllabus change: a new paragraph in Appendix I tracks a change in the regulation, not a change of teaching fashion, and paragraph 10.10 shows the pattern plainly. Information security with a potential impact on aviation safety became a regulated subject in its own right through Commission Implementing Regulation (EU) 2023/203, which reaches Part-145 and Part-CAMO organisations directly; the syllabus followed. When you meet an unfamiliar paragraph number, the useful question is which instrument it is answering.

The Parts Within Regulation (EU) No 1321/2014

Regulation 1321/2014 is the most important implementing regulation for maintenance engineers. It contains multiple "Parts" as annexes:

Part	Annex	Scope
Part-M	Annex I	Continuing airworthiness requirements for aircraft other than those within the Part-ML scope, plus any Part-ML-size aircraft listed on the air operator certificate of a licensed air carrier
Part-145	Annex II	Approved maintenance organisation requirements
Part-66	Annex III	Certifying staff — licensing requirements
Part-147	Annex IV	Maintenance training organisation requirements
Part-ML	Annex Vb	Light aircraft continuing airworthiness (simplified Part-M)
Part-CAMO	Annex Vc	Continuing Airworthiness Management Organisations
Part-CAO	Annex Vd	Combined Airworthiness Organisations (maintenance and continuing-airworthiness management under one approval, for aircraft that are not complex motor-powered and are not listed in the air operator certificate of an air carrier licensed in accordance with Regulation (EC) No 1008/2008)
Part-T	Annex Va	Continuing airworthiness of third-country-registered aircraft dry leased-in by an EU licensed air carrier

Why the Annexes are lettered the way they are

The lettering of Annexes Va to Vd looks arbitrary until you see the regulation's own housekeeping. Article 7 of Regulation (EU) No 1321/2014 repeals Regulation (EC) No 2042/2003 and provides that references to the repealed regulation "shall be construed as references to this Regulation and shall be read in accordance with the correlation table in Annex VI". Annex V is the list of the repealed regulation and its successive amendments, and Annex VI is that correlation table. Both are administrative annexes that predate the four newer Parts, and the lettering leaves them where they were: Part-T, Part-ML, Part-CAMO and Part-CAO were added as Annexes Va, Vb, Vc and Vd rather than by renumbering everything after Annex IV.

The practical point is a small one but it saves confusion in the exam: the letter tells you when a Part joined the regulation, not how important it is. Annexes I to IV are the original four; the lettered annexes are the later arrivals.

What decides which Part applies to an aircraft

Article 1 sets the outer boundary. The regulation covers the continuing airworthiness of aircraft, including components for installation on them, which are registered in a Member State — "unless their regulatory safety oversight has been delegated to a third country and they are not used by an EU operator" — or registered in a third country and used by an EU operator where their regulatory safety oversight has been delegated to a Member State. Article 1(b) adds the separate case of third-country registered aircraft dry leased in by a licensed air carrier.

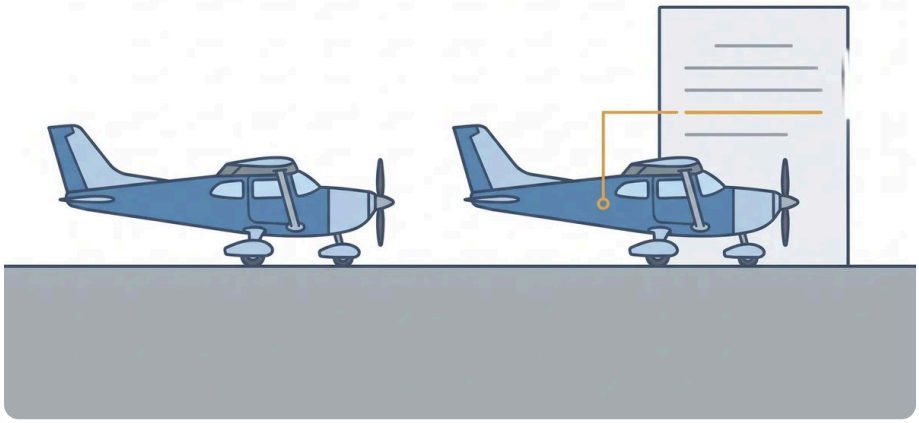
Within that boundary, Article 3 splits the fleet. Article 3(1) makes Part-M the general rule: the continuing airworthiness of Article 1(a) aircraft and their components "shall be ensured in accordance with the requirements of Annex I (Part-M), except for aircraft listed in the first subparagraph of paragraph 2 to which the requirements of Annex Vb (Part-ML) shall apply".

Article 3(2) then defines the Part-ML set, and every entry in it is qualified as "other than complex motor-powered aircraft":

- aeroplanes of 2 730 kg maximum take-off mass or less;
- helicopters of 1 200 kg maximum take-off mass or less, certified for a maximum of up to four occupants;
- other ELA2 aircraft;
- non-conventional aircraft with a maximum take-off mass of 1 200 kg or less if they can maintain zero horizontal speed in flight, or of 2 730 kg or less otherwise.

And then the switch that catches people out. The last subparagraph of Article 3(2) provides: "Where an aircraft referred to in the first subparagraph is listed in the air operator certificate of an air carrier licensed in accordance with Regulation (EC) No 1008/2008, the requirements of Annex I (Part-M) to this Regulation shall apply." Article 3(3) sets three conditions that must be met before such an aircraft can be listed at all: its aircraft maintenance programme has been approved by the competent authority under point M.A.302; the due maintenance required by that programme has been performed and certified under points 145.A.48 and 145.A.50; and an airworthiness review has been performed under point M.A.903 with a new airworthiness review certificate issued under point M.A.901 or M.B.901.

Work the direction through on one airframe. A 2 000 kg four-seat aeroplane in private ownership is inside the Article 3(2) list, so Part-ML applies to it. Put the same aeroplane on the air operator certificate of a licensed air carrier and Part-M applies instead. Nothing about the aircraft changed; the entry on an AOC did. **Neither mass alone nor "it is used commercially" is the test on its own.** The test is the mass-and-class list in Article 3(2) first, and then the AOC listing in its last subparagraph, and the second overrides the first.



The same light aeroplane: Part-ML in private hands, Part-M once it is listed on an air carrier's AOC.

Two further allocations complete the picture. Article 3(4) provides that for an aircraft holding a permit to fly, continuing airworthiness is ensured on the basis of the specific arrangements defined in the permit to fly issued under Annex I (Part-21) to Regulation (EU) No 748/2012 — so a permit-to-fly aircraft follows its permit, not a standard Part. Article 3(6) allocates the Article 1(b) dry leased-in aircraft to Annex Va (Part-T).

What decides which Part applies to an organisation or a person

Article 4(1) covers organisations: those involved in the continuing airworthiness of aircraft and components, including maintenance, "shall be approved, upon their request, by the competent authority in accordance with the requirements of Annex II (Part-145), Annex Vc (Part-CAMO) or Annex Vd (Part-CAO), as applicable to the respective organisations". The words "upon their request" are worth noticing: an approval is applied for, never imposed.

The boundary between the three is set by scope rather than by size, and point CAO.A.010 states it exactly. A combined airworthiness organisation approval covers "the maintenance **and** continuing airworthiness management of aircraft and components for installation thereon" — both activities under one approval — but only "where such aircraft are not classified as complex motor-powered aircraft and are not listed in the air operator certificate of an air carrier licensed in accordance with Regulation (EC) No 1008/2008". So a CAO cannot take on an airline's fleet or a complex motor-powered aircraft, however competent it is; that work needs a Part-145 organisation and a Part-CAMO organisation, whether as two approvals or as one organisation holding both.

Article 5(1) covers people: "certifying staff shall be qualified in accordance with the requirements of Annex III (Part-66)", except as provided for in a closed list of points — M.A.606(h), M.A.607(b), M.A.801(c) and M.A.803

of Part-M; ML.A.801(c) and ML.A.803 of Part-ML; CAO.A.040(b) and (c) of Part-CAO; and point 145.A.30(j) of and Appendix IV to Part-145. The structure is worth reading carefully: **Part-66 is the rule and those points are exceptions to it**, not parallel routes of equal standing. An exception has to be pointed to by number; the rule does not.

Article 6(1) covers training: organisations involved in the training of the personnel referred to in Article 5 "shall be approved in accordance with Annex IV (Part-147)" to be entitled to conduct recognised basic training courses and/or recognised type training courses, to conduct examinations, and to issue training certificates. Point 147.A.05 completes the loop by defining what those courses must contain: Part-147 "establishes the requirements to be met by organisations seeking approval to conduct training and examination **as specified in Annex III (Part-66)**". The content comes from Part-66; Part-147 governs the organisation that delivers it. The arrow does not run the other way — a training organisation cannot define the syllabus it teaches.

One signature, six Parts

The pieces are easiest to hold together on a single concrete case. Take a B1.1 licence holder releasing a transport aeroplane after line maintenance for a licensed air carrier, and ask which Part answers each question that arises.



A B1.1 engineer releases a transport aeroplane after line maintenance: six Parts stand behind that signature.

Question	Part	Where the answer sits
May I hold the licence, and who issued it?	Part-66	Article 5(1) requires Part-66 qualification; the issuing authority is fixed by point 66.1(a)
Where did my basic course and examinations come from?	Part-147	Article 6(1); the content is set by Part-66 through point 147.A.05
Where did my type training syllabus come from?	Part 21	The operational suitability data forming part of the aircraft's type certificate, under Article 1(2)(k) of Regulation (EU) No 748/2012
May my employer release the aircraft?	Part-145	Article 4(1); the competent authority is fixed by point 145.1

Who manages the aircraft's continuing airworthiness?	Part-CAMO	Point M.A.201(e)(2) requires the licensed air carrier to be approved as a CAMO as part of its AOC
Which continuing airworthiness rules does the aircraft follow?	Part-M	Article 3(1), the general rule for aircraft registered in a Member State
And if the same operator dry leases in a third-country registered aircraft?	Part-T	Articles 1(b) and 3(6) move the aircraft out of Part-M altogether

Six Parts and two regulations, before the seventh appears the moment the operator leases an aircraft in. That is why so much of this module is about structure rather than about doing the job: the structure is what makes the signature lawful.

Where responsibility actually sits

Point M.A.201 is the Part-M point that allocates responsibility.

- **The default is the owner.** M.A.201(a): "the owner of the aircraft shall be responsible for the continuing airworthiness of aircraft" and shall ensure no flight takes place unless the aircraft is maintained in an airworthy condition, its operational and emergency equipment is correctly installed and serviceable or clearly identified as unserviceable, the airworthiness certificate is valid, and maintenance is performed in accordance with the approved maintenance programme.
- **Leasing moves it.** M.A.201(b): where the aircraft is leased, the owner's responsibilities transfer to the lessee if the lessee is stipulated on the registration document or the transfer is detailed in the leasing contract; thereafter references to the owner in Part-M cover the lessee.
- **A licensed air carrier takes it on itself.** M.A.201(e): where an aircraft is used by an air carrier licensed under Regulation (EC) No 1008/2008, the operator is responsible for the continuing airworthiness of the aircraft it operates.
- **Doing the work never transfers it away.** M.A.201(c): "any person or organisation performing maintenance shall be responsible for the tasks performed." Contracting a task out does not move the responsibility for the aircraft; it adds a second responsibility for the task.
- **The pre-flight inspection is not maintenance.** M.A.201(d): the pilot-in-command — or, for aircraft used by licensed air carriers, the operator — is responsible for its satisfactory accomplishment, it may be carried out by the pilot or another qualified person, and it "shall not need to be carried out by an approved maintenance organisation or by certifying staff".

Information security sits outside this regulation

One requirement that reaches a Part-145 organisation directly is not in Regulation (EU) No 1321/2014 at all, and knowing that is the point. Commission Implementing Regulation (EU) 2023/203 of 27 October 2022 sets out the requirements to be met by organisations and competent authorities "to identify and manage information security risks with potential impact on aviation safety", to detect information security events and identify those that are incidents with such potential impact, and "to respond to, and recover from, those information security incidents" (Article 1).

Its scope reaches this module's population squarely. Article 2(1)(a) applies it to maintenance organisations subject to Section A of Annex II (Part-145), except those solely involved in the maintenance of Part-ML aircraft; Article 2(1)(b) applies it to continuing airworthiness management organisations subject to Section A of Annex Vc (Part-CAMO), with the same exception; and Article 2(3) applies it to "the competent authority responsible for the issuance, continuation, change, suspension or revocation of aircraft maintenance licences in accordance with Annex III (Part-66)". It applies from 22 February 2026.

Regulation (EU) No 1321/2014 then reaches back across to it. Point 145.B.125(c) requires the competent authority to provide the Agency as soon as possible with safety-significant information stemming from the information security reports it receives "pursuant to point IS.I.OR.230 of Annex II (Part-IS.I.OR) to Implementing Regulation (EU) 2023/203". **Part-IS is not an Annex to Regulation (EU) No 1321/2014; the continuing airworthiness regulation points at it.** Paragraph 10.10 of this module takes the subject further.

Occurrence Reporting — Regulation (EU) No 376/2014

This regulation mandates reporting of aviation safety occurrences. Key requirements:

- **Mandatory reporting** is required for incidents, accidents, and other safety-relevant events by organisations, pilots, engineers, and other defined reporters
- **Initial report** must be submitted within **72 hours** of becoming aware of the occurrence
- **Voluntary reporting** is encouraged for events not captured by mandatory criteria
- Reports are protected under **"just culture"** principles — reporters acting in good faith shall not be penalised
- Information collected is used for **safety improvement**, not for blame or punishment

Key timeline: The initial occurrence report must be filed within **72 hours**. This timeframe is commonly tested in Module 10 examinations.

What counts as an occurrence, and which ones are yours

Regulation (EU) No 376/2014 defines an "occurrence" as "any safety-related event which endangers or which, if not corrected or addressed, could endanger an aircraft, its occupants or any other person and includes in particular an accident or serious incident" (Article 2, point (7)). The definition is deliberately wide, and the phrase "if not corrected or addressed" is the part that catches maintenance work: a defect found and rectified before flight is still an occurrence if, left alone, it could have endangered the aircraft.

Article 4(1) narrows the mandatory duty to occurrences "which may represent a significant risk to aviation safety" falling into listed categories. The maintenance family is Article 4(1)(b), "occurrences related to technical conditions, maintenance and repair of aircraft", and it lists four kinds: structural defects; system malfunctions; maintenance and repair problems; and propulsion problems, "including engines, propellers and rotor systems", and auxiliary power unit problems. Article 4(5) provides that the Commission adopts, by implementing act, a list classifying the occurrences to be referred to when reporting, together with a separate simplified list for aircraft other than complex motor-powered aircraft.

Article 3(2) sets the scope: the Regulation applies to occurrences and other safety-related information involving civil aircraft to which Regulation (EU) 2018/1139 applies, with a carve-out for unmanned aircraft not requiring a certificate or declaration unless the occurrence caused a fatal or serious injury or involved other aircraft. A Member State may extend it to aircraft outside the Basic Regulation's scope.

Who must report, and the two 72-hour clocks

Article 4(6) lists the natural persons who must report, and two of its entries are aimed squarely at this profession:

- point (b): "a person engaged in designing, manufacturing, continuous airworthiness monitoring, maintaining or modifying an aircraft, or any equipment or part thereof, under the oversight of a Member State or of the Agency";
- point (c): "a person who signs an airworthiness review certificate, or a release to service in respect of an aircraft or any equipment or part thereof", under that same oversight.

The duty is personal. It attaches to you as an individual, not to your employer on your behalf, and it is not discharged by assuming that somebody senior will deal with it.

There are then **two separate 72-hour clocks**, and confusing them is a common error:

- **Yours.** Article 4(7): the persons listed in Article 4(6) "shall report occurrences within 72 hours of becoming aware of the occurrence, unless exceptional circumstances prevent this".
- **Your organisation's.** Article 4(8): following notification, an organisation established in a Member State reports the details to the competent authority of that Member State "as soon as possible, and in any event no later than 72 hours after becoming aware of the occurrence". Article 4(9) applies the same clock to an organisation certified or approved by the Agency, but its report goes to the Agency instead.

Note carefully what the clock runs from. It runs from **becoming aware**, not from the event. A crack found during a scheduled inspection may have been growing for months; the 72 hours starts when it is found. And the second clock starts from the organisation becoming aware, so a report filed by an engineer late on a Friday does not consume the organisation's own three days.



A crack found during a scheduled inspection: the 72-hour clock starts now, not when the crack began.

Article 4(2), (3) and (4) build the systems those reports travel through: each organisation established in a Member State, each Member State, and the Agency must each establish a mandatory reporting system. Article 4(6) sets the routing: you report through the system of the organisation that employs, contracts or uses your services or, failing that, through the system of the Member State in which your organisation is established, or through the Agency's system where the Agency approved your organisation. There is always a route, even if your organisation has failed to provide one.

Voluntary reporting is not a lesser version of the same thing

Article 5 requires every organisation, every Member State and the Agency to establish a voluntary reporting system as well, and defines what it is for: "details of occurrences that may not be captured by the mandatory reporting system" and "other safety-related information which is perceived by the reporter as an actual or potential hazard to aviation safety". Article 5(4) states the two gaps it fills — occurrences not subject to mandatory reporting under Article 4(1), and reports "by persons who are not listed in Article 4(6)".

That second limb is the one worth remembering. The mandatory duty binds a closed list of people; the voluntary system is open to anybody. A storesman who notices a labelling practice that will eventually put the wrong part in a technician's hand is not on the Article 4(6) list and has nothing to report under Article 4(1) — and the voluntary system exists precisely for that report.



Mandatory reporting binds a closed list of people; the voluntary system is open to anybody.

Article 2(10) supplies the target: a "hazard" is "a situation or an object with the potential to cause death or injury to a person, damage to equipment or a structure, loss of material, or a reduction of ability to perform a prescribed function". Article 5(5) and (6) require the organisation to pass on what may involve an actual or potential safety risk — to the Agency if the Agency approved it, otherwise to the competent authority — and allow a Member State to require an organisation to report everything it collects voluntarily. Article 5(8) permits information from voluntary and mandatory reporting to be integrated into a single system, so a single reporting form can serve both, with the classification made afterwards.

What happens to the report after you file it

The obligations do not stop at collection, and the deadlines that follow are as specific as the first one.

Step	Who	By when	Provision
Report the occurrence	the reporter	72 hours from becoming aware, unless exceptional circumstances prevent it	Article 4(7)
Report onward to the competent authority, or to the Agency	the organisation	as soon as possible, and in any event no later than 72 hours	Article 4(8) and 4(9)
Analyse to identify the safety hazards, and determine corrective or preventive action	the organisation	—	Article 13(1)
Implement the action in a timely manner and establish a process to monitor its implementation and effectiveness	the organisation	—	Article 13(2)
Transmit the preliminary results of the analysis and any action to be taken	the organisation	30 days from the date of notification of the occurrence by the reporter	Article 13(4) and 13(5)
Transmit the final results	the organisation	as soon as available and, in principle, no later than three months from notification	Article 13(4) and 13(5)
Store the analysis and follow-up in the European Central Repository	the Member State	no later than two months after storage in the national database	Article 13(9)

Article 13(8) closes the loop upwards: the Member State or the Agency has access to the analysis, monitors the action taken, and where it concludes that the implementation and effectiveness of the reported action is inappropriate "shall ensure that additional appropriate action is taken and implemented by the relevant organisation". Article 13(3) closes it downwards: each organisation "shall regularly provide its employees and contracted personnel with information concerning the analysis of, and follow-up on, occurrences for which preventive or corrective action is taken". Feedback to the shop floor is a legal obligation, not a courtesy, and an organisation that never closes that loop is not complying with Article 13(3).

Who handles the report inside the organisation

Article 6(1) requires each organisation to designate one or more persons "to handle independently the collection, evaluation, processing, analysis and storage of details of occurrences", and states the purpose of that independence: the handling must be done "with a view to preventing the use of information for purposes other than safety" and must "appropriately safeguard the confidentiality of the identity of the reporter and of the persons mentioned in occurrence reports, with a view to promoting a 'just culture'". Article 6(2) allows small organisations, by agreement with the competent authority, to use a simplified mechanism and to share those tasks with organisations of the same nature while still complying with the confidentiality and protection rules.

Article 6(3) requires each Member State to designate one or more competent authorities to establish a mechanism to collect, evaluate, process, analyse and store the details independently, and it names the candidates: the national civil aviation authority, the safety investigation authority, or "any other independent

body or entity based in the Union that is entrusted with this function", jointly or separately. So the authority that receives your occurrence report is not necessarily the authority that issued your licence.

Article 7 governs quality: a report must contain at least the information listed in Annex I to the Regulation; reports stored in the organisation, national and Agency databases must include a safety risk classification, reviewed, amended if necessary and endorsed by the competent authority or the Agency under a common European risk classification scheme; and the databases must use formats that are standardised for exchange and "compatible with the ECCAIRS software and the ADREP taxonomy". Article 8(1) puts the Commission at the top of the chain: it "shall manage a European Central Repository to store all occurrence reports collected in the Union", to which the Member States transfer the safety information from their national databases.

Just culture: what it protects, and the two things it does not

Article 2, point (12) defines just culture as "a culture in which front-line operators or other persons are not punished for actions, omissions or decisions taken by them that are commensurate with their experience and training, but in which gross negligence, wilful violations and destructive acts are not tolerated". Both halves of that sentence are operative. Article 1(2) states the object it serves: "the sole objective of occurrence reporting is the prevention of accidents and incidents and not to attribute blame or liability."

Article 15(2) turns that into a restriction on use: information derived from occurrence reports "shall be used only for the purpose for which it has been collected", and Member States, the Agency and organisations "shall not make available or use the information on occurrences in order to attribute blame or liability, or for any purpose other than the maintenance or improvement of aviation safety".

Article 16 protects the source, and the mechanism is more subtle than anonymity:

- Article 16(2): inside the organisation, all personal details are made available to staff other than the persons designated under Article 6(1) "only where absolutely necessary in order to investigate occurrences with a view to enhancing aviation safety", and disidentified information is disseminated within the organisation as appropriate.
- Article 16(3) and (4): "no personal details are ever recorded" in the national database or in the Agency database. The disidentified information is made available to all relevant parties.
- Article 16(6): without prejudice to applicable national criminal law, Member States "shall refrain from instituting proceedings in respect of unpremeditated or inadvertent infringements of the law which come to their attention only because they have been reported".
- Article 16(7): where disciplinary or administrative proceedings are instituted under national law, information contained in occurrence reports "shall not be used against the reporters or the persons mentioned in occurrence reports".
- Article 16(9): employees and contracted personnel who report or are mentioned in reports "shall not be subject to any prejudice by their employer or by the organisation for which the services are provided on the basis of the information supplied by the reporter".

So the protection is not that nobody knows who reported. Your organisation's designated person knows, and has to, in order to investigate. The protection is that your identity does not travel beyond that point, and that what you wrote cannot be turned against you.

And it stops in two named places. Article 16(10) provides that the protection under paragraphs 6, 7 and 9 does not apply "in cases of wilful misconduct", or "where there has been a manifest, severe and serious disregard of an obvious risk and profound failure of professional responsibility to take such care as is evidently required in the circumstances, causing foreseeable damage to a person or property, or which seriously compromises the level of aviation safety". That second limb is a demanding test with four elements — a manifest and severe disregard, of an obvious risk, with a profound failure of professional responsibility, causing foreseeable damage or seriously compromising safety — and it is not met by ordinary error, however costly the outcome.

Two further provisions decide how much protection you actually have, and they run in the same direction. Article 16(8) allows Member States to "adopt or maintain in force legislative provisions ensuring a higher level of protection"; the closing subparagraphs of Article 16(6) and (7) allow them to apply the protection without the Article 16(10) exceptions and to extend it to civil or criminal proceedings. **The Regulation sets a harmonised floor, not a ceiling**, so a statement about your own protection has to be checked against your Member State's law as well as against the Regulation.

Article 16(11) requires each organisation, "after consulting its staff representatives", to adopt internal rules describing how just culture principles are guaranteed and implemented within it — which is the document you should ask for if you want to know how your own employer applies this. Article 16(12) requires each Member State to designate a body responsible for implementing paragraphs 6, 9 and 11; employees and contracted personnel may report alleged infringements of Article 16 to that body, may not be penalised for doing so, and may inform the Commission. Article 15(4) requires competent authorities and the authorities responsible for the administration of justice to cooperate through advance administrative arrangements seeking "the correct balance between the need for proper administration of justice, on the one hand, and the necessary continued availability of safety information, on the other".

⚠ **Exam trap:** just culture is neither immunity nor anonymity. It is not immunity, because Article 16(10) withdraws the protection for wilful misconduct and for a manifest, severe and serious disregard of an obvious risk, and because Article 16(6) operates "without prejudice to applicable national criminal law". It is not anonymity, because the designated person under Article 6(1) necessarily knows who reported — what Articles 16(3) and 16(4) guarantee is that no personal details are ever recorded in the national or Agency databases. An answer offering "complete anonymity" or "immunity from all legal proceedings" is wrong on both counts.

Where the report goes next, and why it never carries your name

Two points in Part-145 show the onward route, and they are a good illustration of how two regulations interlock without duplicating each other. Point 145.B.125(a) requires the competent authority of the Member State to notify the Agency of any significant problems with the implementation of Regulation (EU) 2018/1139 and its delegated and implementing acts "within 30 days from the time the authority became aware of the problems". Point 145.B.125(b) adds that, without prejudice to Regulation (EU) No 376/2014, the competent authority "shall provide the Agency as soon as possible with any safety-significant information stemming from the occurrence reports stored in the national database pursuant to Article 6(6) of Regulation (EU) No 376/2014".

Beyond the deadlines set out above, the route keeps going. The national database entry is classified for risk under the common European scheme and transferred to the European Central Repository, and safety-

significant information from it reaches the Agency. At Union level the Commission, the Agency and the competent authorities exchange and analyse the repository's contents through a network of aviation safety analysts (Article 14), and each Member State publishes an annual safety review containing aggregated and anonymised information on the types of occurrence reported, the trends identified and the action taken (Article 13(11)).

At no point on that route does your name travel with the report. That is not a courtesy of the system; it is the design that makes the system work at all, because a reporting scheme that exposed its reporters would very quickly have nothing to analyse.

Bilateral Agreements

The European Union may conclude **bilateral aviation safety agreements (BASAs)** with third countries. These agreements provide for mutual recognition of certificates and approvals. The most significant is the **EU–US Bilateral Agreement** which covers:

- Mutual acceptance of airworthiness certificates and design approvals
- Maintenance performed under FAA-approved repair stations being accepted in the EU (and vice versa)
- Reduced duplication of oversight activities

Remember: Bilateral agreements are concluded between the *European Union* and the third country, not by EASA. A bilateral agreement that an individual Member State concluded with a third country before Regulation (EC) No 1592/2002 entered into force, and notified to the Commission and the other Member States, also remains a valid basis for accepting that country's certificates - but only in the absence of a Union agreement or delegated act under points (a) and (b), and without prejudice to Article 140(6). EASA provides technical support for the negotiations and conducts the technical assessments and evaluations of third-country laws and foreign aviation authorities.



Under the EU-US agreement, work from an FAA-approved repair station can be accepted in the EU.

Three legal routes, and no fourth

Article 68 of Regulation (EU) 2018/1139 is the provision that makes acceptance of a third-country certificate lawful, and it is worth reading as a closed list rather than as a general permission. The Agency and the national competent authorities may do one of two things: either issue the certificates provided for in the Regulation "on the basis of certificates issued in accordance with the laws of a third country", or "accept certificates and other relevant documentation attesting compliance with civil aviation rules which were issued in accordance with the laws of a third country". But they may do so only "where such possibility is provided for in":

- **Article 68(1)(a)** — "international agreements concerning the recognition of certificates concluded between the Union and a third country". This is the bilateral aviation safety agreement route, and note the parties: the Union and the third country.
- **Article 68(1)(b)** — a delegated act adopted on the basis of Article 68(3). Under that paragraph the Commission may lay down detailed rules on acceptance of third-country certificates and documentation "ensuring an equivalent level of safety to that provided for in this Regulation, including the conditions and procedures for achieving and maintaining the necessary confidence in regulatory systems of third countries".
- **Article 68(1)(c)** — in the absence of both, and without prejudice to Article 140(6), an agreement concerning the recognition of certificates "concluded between a Member State and a third country before the entry into force of Regulation (EC) No 1592/2002" that was notified to the Commission and the other Member States.

Acceptance is therefore an exception, not a default. Where none of the three routes covers a document, there is no power to accept it, however competent the issuing authority may be. The consequence is worth stating plainly: a certificate is not accepted because it is technically equivalent, it is accepted because an instrument provides that it may be.

The two limbs of Article 68(1) are also worth separating, because they produce different pieces of paper. Issuing an EU certificate on the basis of a third-country one leaves you holding an EU certificate. Accepting the third-country document leaves the foreign document in place and gives it effect. Both are provided for; they are not the same transaction.

Who negotiates, who assesses, and what a working arrangement is not

Article 68(2) sets out the Agency's part: "in order to achieve and maintain confidence in the regulatory systems of third countries, the Agency shall be authorised to conduct the necessary technical assessments and evaluations of the laws of third countries and of foreign aviation authorities", and for that purpose "may conclude working arrangements in accordance with Article 90(2)". Article 90(1) adds that the Agency, on request, assists the Commission in its management of relations with third countries and international organisations, contributing in particular "to the harmonisation of rules, the mutual recognition of certificates, in the interest of European industry, and the promotion of European aviation safety standards".

Article 90(2) then draws the line between cooperation and obligation: the Agency may establish working arrangements with third-country authorities and international organisations, in prior consultation with the

Commission, but "those working arrangements shall not create legal obligations incumbent on the Union and its Member States".

Instrument	Concluded by	Effect
International agreement on the recognition of certificates	the Union and the third country	provides the legal possibility of accepting the other side's certificates, under Article 68(1)(a)
Delegated act on acceptance	the Commission, under Article 68(3) and Article 128	lays down detailed rules for acceptance ensuring an equivalent level of safety, under Article 68(1)(b)
Pre-2002 Member State agreement	a Member State and the third country, notified to the Commission and the other Member States	a residual basis for acceptance under Article 68(1)(c), and a closing one
Working arrangement	the Agency, in prior consultation with the Commission	technical cooperation only; expressly creates no legal obligations for the Union or the Member States

The direction is consistent across all four rows. The legal power to recognise is created at Union level; the technical confidence that makes recognition sensible is built by the Agency; and the Agency's own instruments carry no legal obligation. An engineer meets the results of the first row; the last row, the working arrangement, never appears on a certificate at all, and nothing is accepted under Article 68(1) on the strength of one.

Why Member States stopped signing their own

Article 68(1)(c) has an unusual shape for a legal provision: it is a door that is closing. It admits only agreements concluded *before* Regulation (EC) No 1592/2002 entered into force and notified under that regulation or under Regulation (EC) No 216/2008, and it is expressly "without prejudice to Article 140(6)". Article 140(6) is the reason: "Member States shall terminate or adjust existing bilateral agreements that they concluded with third countries for the fields covered by this Regulation as soon as possible following the entry into force of this Regulation and in any event before 12 September 2021."

The logic is straightforward once the tiers of this note are in view. Recognition of certificates is a matter the Union now regulates, so it is the Union that concludes agreements about it. Separate national arrangements would produce as many different answers to the question of whether a component may be fitted as there are Member States — which is precisely the fragmentation the single European aviation safety system in Article 62(1) exists to remove.

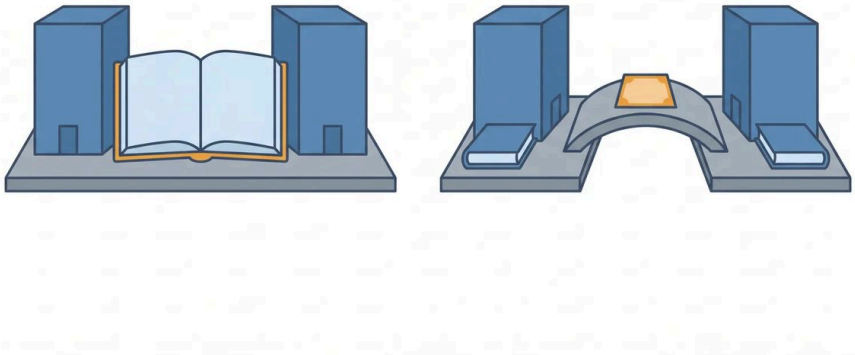
Traces of the older arrangements survive where they were used to certificate products that are still flying. Article 3(1) of Regulation (EU) No 748/2012 provides that a product with a type certificate issued before 28 September 2003 by a Member State is deemed to have a type certificate issued under that Regulation where its type-certification basis was that of a State of design "with which a Member State had concluded a bilateral airworthiness agreement or similar arrangement under which such products have been certificated on the basis of the certification specifications of that State of design" — unless the Agency determines that those specifications, the service experience or that State's safety system do not provide a level of safety equivalent

to that required by Regulation (EU) 2018/1139. An old bilateral agreement can therefore still be the reason a type certificate is valid today, even though no Member State could conclude such an agreement now.

Participation is not recognition

There is a second kind of agreement with a third country, and confusing it with a recognition agreement gets the relationship wrong in both directions. Article 129 provides that "the Agency shall be open to the participation of European third countries which are contracting parties to the Chicago Convention and which have entered into international agreements with the Union pursuant to which they adopted and apply Union law in the fields covered by this Regulation". Such agreements may specify the nature and extent of the country's participation in the Agency's work, including financial contributions and staff, and the Agency may establish working arrangements under Article 90(2) to give them effect.

The two instruments produce opposite results. Under an Article 129 participation agreement the third country **adopts and applies Union law**, so its rules are the same rules. Under an Article 68(1)(a) recognition agreement each side keeps its own law, and the agreement provides that certain of the other side's certificates and findings may be accepted. One harmonises the rules; the other bridges two sets of rules that stay different.



Participation shares one set of rules; recognition keeps two sets and lets certificates cross between them.

⚠ **Exam trap:** a bilateral aviation safety agreement makes certificates and findings of compliance mutually acceptable. It does not make the partner authority's rules applicable in the Union, and it does not give that authority any power over Union operators or organisations. Answers of the form "the partner authority regulates European operators" or "the partner country's requirements replace EU requirements" invert the whole point of the instrument: recognition operates on *certificates*, never on *jurisdiction*.

Module 11 — Aeroplane Aerodynamics, Structures and Systems

11.1(a) — Aeroplane Aerodynamics and Flight Controls

This section covers the fundamental principles of aerodynamics as they apply to aeroplane flight. Understanding how air behaves around an aerofoil, how lift and drag are generated, and how an aircraft is controlled and stabilised is essential knowledge for every aircraft maintenance engineer. These principles underpin every maintenance task — from rigging flight controls to inspecting wing surfaces for contamination.

The Atmosphere

The **atmosphere** is the envelope of air surrounding the Earth. It is composed of approximately **78% nitrogen**, **21% oxygen**, and **1% other gases** (argon, CO_2, water vapour, etc.). The atmosphere has several key properties relevant to flight:

- **Pressure** — decreases with altitude. At sea level, standard pressure is **1013.25 hPa** (29.92 inHg, 14.7 psi). Pressure halves roughly every 18,000 ft.
- **Temperature** — decreases with altitude in the troposphere at approximately **1.98°C per 1,000 ft** (lapse rate). ISA sea-level temperature is **15°C**.
- **Density** — decreases with altitude and increasing temperature. Air density directly affects lift generation and engine performance.
- **Humidity** — water vapour displaces heavier nitrogen and oxygen molecules, reducing air density and therefore reducing lift.

International Standard Atmosphere (ISA)

The **ISA** is a theoretical model of atmospheric conditions used as a reference standard for aircraft performance calculations, altimeter calibration, and engineering design. ISA conditions at sea level:

ISA Sea-Level Values

- Temperature: **15°C** (288.15 K / 59°F)
- Pressure: **1013.25 hPa** (29.92 inHg / 760 mmHg)
- Density: **1.225 kg/m³**
- Lapse rate: **–1.98°C per 1,000 ft** (up to the tropopause at ~36,090 ft)
- Temperature at tropopause: **–56.5°C** (constant above, in the stratosphere)

ISA is used as a baseline. Actual conditions are expressed as deviations: "ISA +10" means the actual temperature is 10°C warmer than ISA standard at that altitude.

Why Density Is the Property That Matters

Density is not measured directly in flight; it is a consequence of pressure and temperature acting together through the gas law. Pressure falls steeply with height while temperature falls comparatively slowly, so the pressure term dominates and density falls with altitude even though colder air is denser at a given pressure.

Density from Pressure and Temperature

$$\rho = \frac{P}{RT}$$

Where P is absolute pressure in pascals, T is absolute temperature in kelvin, and R = 287 J/kg·K for dry air. Temperature must be in kelvin: K = °C + 273.15.

Worked Example — Checking the ISA Values

Sea-level ISA density: $\rho = 101325 \div (287 \times 288.15) = 1.225$ kg/m³ — the standard figure quoted above.

ISA temperature at 20,000 ft: $15 - (1.98 \times 20) = -24.6$°C. At the tropopause: $15 - (1.98 \times 36.09) = -56.5$ °C, which is where the lapse stops. On a day reported as ISA +15, the actual temperature at 20,000 ft would be –9.6°C.

Because pressure halves roughly every 18,000 ft, it has halved twice by the tropopause — about one quarter of the sea-level value.

What Falling Density Does to the Aeroplane

Every aerodynamic force contains the term $\frac{1}{2}\rho v^2$. If an aeroplane flies at the same angle of attack and the same weight at two altitudes, the wing must produce the same lift, so $\frac{1}{2}\rho v^2$ must be the same at both — and since ρ is lower higher up, the **true airspeed must be higher** at the higher altitude. The indicated airspeed, which is a read-out of dynamic pressure rather than of true speed, is unchanged. The same reasoning fixes the stall: the aeroplane always stalls at the same maximum lift coefficient and therefore at the same **equivalent airspeed** for a given weight and configuration, regardless of altitude, while its true airspeed at the stall rises steadily as density falls.

Hot, High and Humid

Density altitude is the pressure altitude corrected for temperature deviation — the altitude at which the ISA would show the density the aircraft is actually experiencing. High airfield elevation, a hot day and high humidity all push it up together, lengthening take-off run, flattening climb gradient and reducing engine thrust, because both the wing and the engine depend on mass flow. An aircraft leaving a hot high airfield can be performing as though it were several thousand feet higher than the altimeter reads.



Elevation, heat and humidity raise density altitude: a longer take-off run and a flatter climb.

Bernoulli's Theorem and Subsonic Airflow

Bernoulli's Principle (Daniel Bernoulli, 1738) states that in a steady flow of an ideal fluid, an increase in velocity produces a decrease in static pressure, and vice versa. The total pressure (the sum of static and dynamic pressure) remains constant along a streamline:

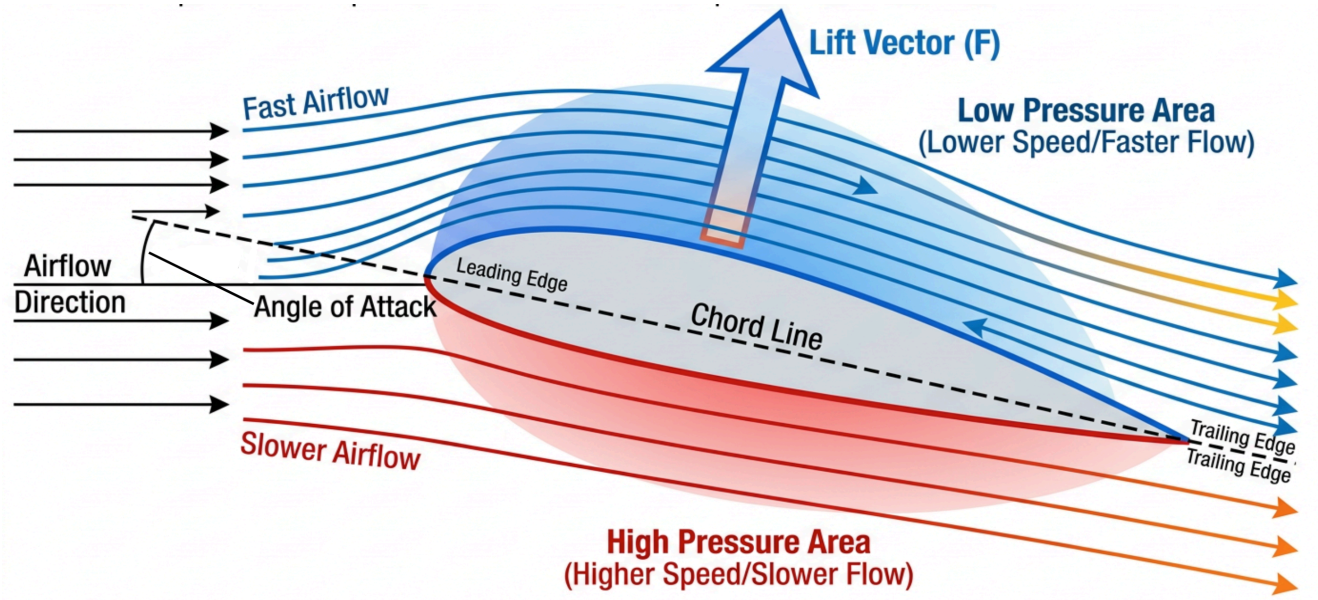
Bernoulli's Equation

$$P_{static} + \frac{1}{2}\rho v^2 = \text{constant}$$

Where: P_{static} = static pressure, ρ = air density, v = velocity

$\frac{1}{2}\rho v^2$ is the **dynamic pressure** (also called q or velocity pressure)

When air flows over an aerofoil, the curved upper surface forces the air to travel a longer path and accelerate. By Bernoulli's principle, this increased velocity reduces the static pressure above the wing. The relatively slower airflow below the wing maintains higher pressure. This pressure difference creates an upward force — **lift**.



✈ Aviation Context — Continuity Equation

Bernoulli's principle works in conjunction with the **Continuity Equation**: $A_1 v_1 = A_2 v_2$. When airflow is constricted (e.g., by the curvature of the upper wing surface), velocity must increase to maintain the same mass flow rate. This increased velocity is what creates the pressure drop described by Bernoulli.

Total, Static and Dynamic Pressure

Bernoulli's equation is the reason the pitot-static system works at all. At the mouth of the pitot tube the flow is brought completely to rest at the **stagnation point**, where velocity is zero and the measured pressure is the full total (pitot) pressure. The static ports, set flush in undisturbed flow along the fuselage, sense static pressure alone. The airspeed indicator capsule is fed both and responds to the difference, so an indicated airspeed reading is really a direct measurement of dynamic pressure $\frac{1}{2}\rho v^2$, not of speed through the air. That is why a wing loaded to a given IAS carries the same aerodynamic load at any altitude, and why the aeroplane's structural and stall speeds are quoted in indicated terms.

Where the Theorem Applies — and Where It Does Not

The equation assumes a fluid that is inviscid and of constant density, so it holds well for subsonic flight at low Mach numbers and becomes progressively less accurate as compressibility takes effect above roughly M 0.3 to 0.4, where the air is measurably squeezed and the simple form must be corrected. It also applies only *outside* the boundary layer: inside it, viscosity dissipates energy, so total pressure is not conserved. That single limitation explains most of the wing's practical behaviour — air that has lost energy in the boundary layer cannot climb the rising pressure towards the trailing edge, which is why separation happens and why devices such as slots, slats and vortex generators work by **adding energy** to that layer rather than by manipulating pressure alone.

The same convergent-passage effect is used deliberately in a wing **slot**, which behaves as a venturi: air is accelerated through the narrowing gap from the high-pressure lower surface and discharged over the upper surface as a fast, high-energy sheet. A **Krueger flap** makes an instructive contrast — it hinges forward and down from the *lower* surface of the leading edge and forms no slot at all, raising the stalling angle by adding leading-edge camber instead of by feeding energised air over the top. Both devices raise the critical angle of attack, so the difference between a slat and a Krueger flap is the presence of the slot, not the effect on the stall.

Dynamic Pressure and the Working Lift Coefficient

Because dynamic pressure varies with the **square** of speed, and because lift must continue to equal weight in level flight, the lift coefficient the wing is actually working at varies inversely with the square of the speed. Speeding up in level flight therefore forces C_L — and with it the angle of attack — to reduce; slowing down forces both to increase. The rigging (incidence) angle between the chord line and the airframe datum is fixed by the structure and never changes in flight, so a change of angle of attack is always a change in the aircraft's attitude or flight path, never a change of incidence.

Worked Examples — the Inverse-Square Rule

Dynamic pressure at 100 m/s at ISA sea level: $q = 0.5 \times 1.225 \times 100^2 = 6125$ Pa.

Doubling the IAS multiplies q by four, so to hold lift equal to weight the lift coefficient must be multiplied by $1/4 = 0.25$.

Flying at 1.3 times the stall speed, q is $1.3^2 = 1.69$ times the value at the stall, so $C_L = C_{L_{max}} \div 1.69 \approx 0.59$ — about 59% of $C_{L_{max}}$.

Constant IAS or Constant Angle of Attack?

Examiners separate two flap questions that look identical. **At constant IAS in straight and level flight**, lift still equals the same weight at the same dynamic pressure, so the *working* C_L is unchanged; what flap deployment raises is the *available* $C_{L_{max}}$, together with the drag. **At constant angle of attack**, nothing balances the extra lift the added camber produces, so lift momentarily exceeds weight and the aeroplane climbs. Note also that adding camber shifts the section's zero-lift angle more negative, so a flapped section behaves as though it were at a larger angle of attack for the same attitude — while the aircraft, holding a constant IAS, ends up flying at a lower nose attitude.

Boundary Layer

The **boundary layer** is the thin layer of air immediately adjacent to the aircraft surface where the air velocity transitions from zero (at the surface, due to viscosity) to the free-stream velocity. It is critical because drag, heat transfer, and surface contamination effects all occur within this layer.

Types of Boundary Layer Flow

Type	Characteristics	Drag
Laminar	Smooth, orderly layers of air sliding over each other; occurs near the leading edge; thin boundary layer	Low skin friction drag
Turbulent	Chaotic mixing of air; occurs further back along the surface; thicker boundary layer; better at resisting flow separation	Higher skin friction drag
Transition point	The location where laminar flow changes to turbulent flow; varies with Reynolds number, surface roughness, and pressure gradient	—

Free-stream flow is the undisturbed airflow outside the boundary layer. **Relative airflow** is the direction of airflow relative to the aerofoil — equal and opposite to the direction of flight.

⚠ Maintenance Significance

Surface roughness (dents, protruding rivet heads, peeling paint, insect residue, ice crystals) causes the laminar-to-turbulent transition to occur earlier, increasing drag and potentially causing flow separation. This is why maintaining **smooth aerodynamic surfaces** is critical during maintenance — even a thin layer of frost can increase drag by 30–40% and reduce lift significantly.

How the Boundary Layer Grows and Why It Separates

The layer starts vanishingly thin at the stagnation point and thickens with distance aft as more and more air is dragged along by friction. Ahead of the point of maximum thickness of the aerofoil the flow is accelerating into falling pressure, which is a stable, helpful condition. Behind it the pressure rises again towards the trailing edge — an **adverse pressure gradient** — and the air nearest the surface, already slowed by viscosity, must climb it on the momentum it has left. When that momentum is exhausted the flow near the surface stops, reverses, and the main stream lifts away: this is **separation**, and it is what a stall, a spoiler and a contaminated wing all have in common.

The two flow types behave differently for exactly this reason. A laminar layer is thin and carries little momentum close to the surface, so although it produces low skin friction it gives way to an adverse gradient early. A turbulent layer continuously mixes fast air from above down towards the wall, so it holds on much further aft at the cost of higher friction. Real wings are therefore a compromise, and every boundary layer control device on the aircraft is an attempt to buy the separation resistance of turbulent flow without paying for it over the whole surface.

Reynolds Number

$$Re = \frac{\rho v L}{\mu}$$

The ratio of inertia forces to viscous forces, where L is a characteristic length (chord, or distance from the leading edge) and μ is the dynamic viscosity of the air. A high Reynolds number means inertia dominates and the flow trips into turbulence readily; this is why the transition point on a full-size wing sits much further forward, as a fraction of chord, than on a small model of the same shape.

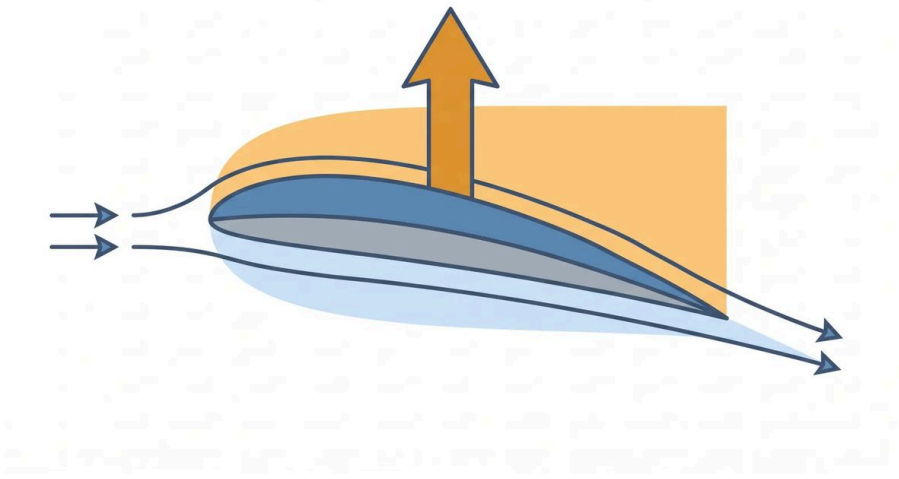
The Stagnation Point and Angle of Attack

The stagnation point is where the oncoming flow divides, and it is not fixed. At a high angle of attack it lies well back under the leading edge, so the air that goes over the top must whip around a tighter corner, producing a sharp suction peak and a correspondingly steep adverse gradient behind it — the reason separation starts near the leading edge at high angles. As the aircraft accelerates in level flight and the angle of attack reduces, the stagnation point migrates forward towards the leading edge and that suction peak flattens out. That acceleration also raises the Reynolds number, but it does not follow that the transition point simply marches forward with speed: transition is set by the pressure gradient as well as by Reynolds number, and the falling angle of attack moves the point of minimum pressure aft, which extends the favourable gradient and works against any forward movement. What is certain in a level acceleration is that lift must continue to equal weight while the dynamic pressure rises, so the lift coefficient must reduce.

Because the layer is thinnest near the leading edge, a surface defect there is far more damaging than the same defect further aft. A proud rivet head, a step at a repair doubler, an unfaired sealant bead, filler that has shrunk, or leading-edge erosion all sit inside a layer only a fraction of a millimetre thick and can trip it or thicken it out of proportion to their size. Repairs in the forward part of the chord are consequently held to tighter contour and flushness limits than repairs further aft, and the structural repair manual will normally say so explicitly.

↑ Generation of Lift

Lift is the aerodynamic force generated perpendicular to the relative airflow. It is produced by the pressure difference between the upper and lower surfaces of the wing, combined with the deflection of airflow (Newton's Third Law — the wing pushes air down, and the air pushes the wing up).



Lift comes from lower pressure above the wing, higher pressure below, and air deflected downward.

Lift Equation

$$L = C_L \times \tfrac{1}{2}\rho v^2 \times S$$

Where:

- L = lift force (Newtons)
- C_L = lift coefficient (depends on aerofoil shape and angle of attack)
- ρ = air density (kg/m³)
- v = airspeed (m/s)
- S = wing planform area (m²)

Angle of Attack (AoA)

The **angle of attack (α)** is the angle between the **chord line** of the aerofoil and the **relative airflow**. As AoA increases, C_L increases (more lift is generated) — up to the **stall angle** (typically 15–18° for conventional aerofoils), beyond which C_L drops sharply as the airflow separates from the upper surface.

Centre of Pressure (CP)

The **centre of pressure** is the point on the chord line where the total aerodynamic force (resultant of lift and drag) effectively acts. As AoA increases, the CP moves forward; as AoA decreases, the CP moves aft. This movement affects the aircraft's pitching moment and longitudinal stability.

Aerofoil Types

Type	Characteristics	Typical Use
Symmetrical	Upper and lower surfaces are mirror images; zero lift at zero AoA; CP does not move	Tailplanes, aerobatic aircraft
Cambered (asymmetric)	Upper surface more curved than lower; generates lift at zero AoA; higher max C_L	Most transport wings
Supercritical	Flattened upper surface delays shock wave formation; higher critical Mach number	Modern transport aircraft (A320, B737)
Laminar flow	Maximum thickness further aft; maintains laminar flow over larger area; lower drag	High-performance gliders, some GA aircraft

↔ Drag

Drag is the aerodynamic force that opposes the aircraft's motion through the air. Total drag is the sum of two main components:

Drag Equation

$$D = C_D \times \tfrac{1}{2}\rho v^2 \times S$$

Total Drag = Parasite Drag + Induced Drag

Parasite Drag

Parasite drag (also called **zero-lift drag**) is drag that is not directly associated with lift production. It **increases with speed squared**. Components:

- **Form drag (pressure drag)** — caused by the shape of the aircraft; the pressure difference between front and rear surfaces. Streamlining reduces form drag.
- **Skin friction drag** — caused by air viscosity in the boundary layer rubbing against the aircraft surface. Smooth surfaces reduce skin friction.
- **Interference drag** — caused by the interaction of airflows at junctions (wing-fuselage, engine pylon-wing). Fairings reduce interference drag.

Induced Drag

Induced drag is drag that is a direct consequence of lift production. High-pressure air below the wing tip flows around to the low-pressure area above, creating wing-tip vortices. These vortices deflect the airflow downward (downwash), tilting the lift vector backward and creating an aft-acting component — induced drag. Key points:

- Induced drag **decreases with speed** (at higher speeds, less AoA is needed for the same lift, producing weaker vortices)
- Induced drag **increases with AoA** and weight
- High aspect ratio wings produce less induced drag (less tip vortex relative to wing area)
- **Winglets** reduce induced drag by limiting the spanwise flow at the wing tip

Ground Effect

When the aircraft is within approximately **one wingspan of the ground**, the ground disrupts the formation of wing-tip vortices, significantly reducing induced drag. This is called **ground effect**. It makes the aircraft feel like it is "floating" during the landing flare and can cause the aircraft to become airborne at a speed below normal take-off speed. Ground effect reduces induced drag by up to **50%** when very close to the surface.

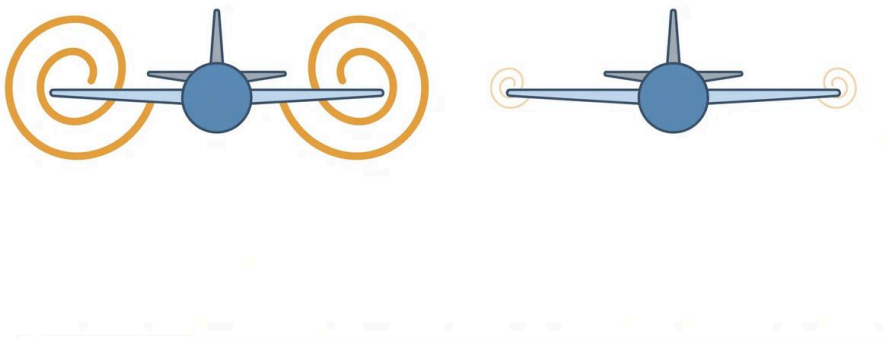
⚖️ Lift/Drag Ratio and Aircraft Polar Diagram

The **lift/drag ratio (L/D)** is a measure of aerodynamic efficiency. It represents how many units of lift are generated for each unit of drag. The maximum L/D ratio occurs at a specific AoA and speed — this is the speed for maximum range and best glide performance.

- Typical L/D max for a modern transport aircraft: **15:1 to 20:1**
- Typical L/D max for a glider: **30:1 to 60:1**
- The **aircraft polar diagram** plots C_L against C_D. A tangent line from the origin to the curve gives the angle of attack for maximum L/D ratio.

The Drag Curve — Why L/D Max Sits Where It Does

The two families of drag pull in opposite directions with speed. Parasite drag grows with the square of speed, while induced drag falls off as speed rises, because less angle of attack is needed to carry the weight and the tip vortices weaken. Adding the two curves gives a total-drag curve shaped like a shallow bucket, and its lowest point — where parasite drag and induced drag are **equal** — is the minimum drag speed, V_{IMD}. Since lift equals weight throughout, minimum drag and maximum lift/drag ratio are the same condition, reached at one angle of attack and, for a given weight, one speed.



Slow flight at high angle of attack gives strong tip vortices; at higher speed they weaken.

Speed region	Dominant drag	Behaviour
Below V_{IMD}	Induced (high angle of attack, strong vortices)	Drag *increases* as speed decreases — the "back of the drag curve"; speed is not self-correcting and thrust must be added to accelerate back out
At V_{IMD}	Induced equals parasite	Minimum total drag, maximum L/D, best glide range, minimum thrust required for level flight
Above V_{IMD}	Parasite (rising with the square of speed)	Drag increases with speed in the ordinary way; speed is stable and self-correcting

Anything that reduces induced drag lowers the induced curve, so the crossing point moves to a **lower** speed and the peak L/D rises. Increasing aspect ratio does exactly this: it reduces induced drag and therefore reduces the speed for minimum drag. Increasing weight has the opposite effect, raising the induced curve and pushing V_{IMD} up in proportion to the square root of weight, while leaving the best lift/drag *ratio* itself unchanged, since that is set by the angle of attack.

Worked Example — Glide Distance

In a steady glide the lift/drag ratio is the glide ratio: distance travelled divided by height lost. A transport aircraft gliding at an L/D of 17 from 30,000 ft covers $17 \times 30{,}000 = 510{,}000$ ft in still air, which is $510{,}000 \div 6076 \approx 84$ nautical miles.

A heavier aircraft glides the same distance, but must fly faster to do it — the angle is unchanged, only the speed at which that angle is achieved. Lowering the landing gear, extending flap or deploying spoilers cuts the achievable L/D sharply and the distance with it.

On the polar diagram, the point of *minimum drag coefficient* is not the point of best L/D: minimum C_D lies at the leftmost point of the curve, whereas maximum L/D is found where a straight line drawn from the origin just

touches the curve. Reading the corresponding C_L and C_D at that tangent point gives both the best ratio and the angle of attack that produces it.

Stalling

A **stall** occurs when the angle of attack exceeds the **critical angle** (stall angle, typically 15–18°). The airflow can no longer follow the upper surface contour, separates from the wing, and lift decreases dramatically while drag increases sharply.

Key Stall Facts

- A stall is an **angle of attack** event, NOT a speed event — an aircraft can stall at ANY speed if the critical AoA is exceeded
- However, the **stall speed** is the minimum speed at which the aircraft can maintain level flight at 1g (with maximum C_L)
- Stall speed increases with: **weight** (heavier = more lift needed = higher speed), **load factor** (in turns), **altitude** (lower density), but that applies to **true airspeed** only — the stalling IAS/EAS for a given weight and configuration does not change with altitude — and **forward CG** (more tail down-force needed)
- Stall speed decreases with: use of **flaps/slats** (increase $C_{L_{max}}$)

Stall Warning Systems

- **Stall warning vane (angle of attack sensor)** — a small vane on the fuselage side that measures the local AoA; triggers warnings before the stall angle is reached
- **Stick shaker** — vibrates the control column to give a tactile warning of approaching stall
- **Stick pusher** — automatically pushes the control column forward to reduce AoA and prevent a full stall (used on T-tail aircraft where a "deep stall" is irrecoverable)

Load Factor and Stall Speed

In a level turn the wing must support more than the weight, because the lift vector is tilted and only its vertical component opposes gravity. The ratio of lift produced to weight carried is the **load factor**, and since lift depends on the square of speed, the stall speed rises with the square root of the load factor — not in direct proportion to it, and not with weight or lift as such.

Stall Speed in a Turn

$$V_{S_{turn}} = V_S \sqrt{n} \qquad n = \frac{1}{\cos \phi}$$

Where n is the load factor and ϕ the angle of bank in a level turn. Weight enters the same way: V_S varies with the square root of weight, so a 21% weight increase raises the stall speed by about 10%.

Worked Examples — Turning Stall Speed

An aeroplane stalls at 100 kt in level flight. In a **45° banked level turn**, $n = 1 \div \cos 45^\circ = 1.41$ and $\sqrt{1.41} = 1.19$, so the stall speed becomes about **119 kt**.

At a **load factor of 1.5**, $\sqrt{1.5} = 1.22$, giving about **122 kt**. A load factor of 2 would be needed to reach 141 kt — the value candidates often pick by mistake.

Two further influences are easy to miss. A **forward centre of gravity** lengthens the arm between the CG and the tailplane, so more download is needed on the tail to hold the nose up; the wing must carry that download as well as the weight, and the stall speed rises. A **reduction in thrust** removes the small vertical component of thrust that was helping to support the aircraft at high nose attitudes, so the wing must supply more lift and again stalls sooner. Both of these raise stall speed, while additional flap and lower weight lower it.

The Stalling Angle Never Changes

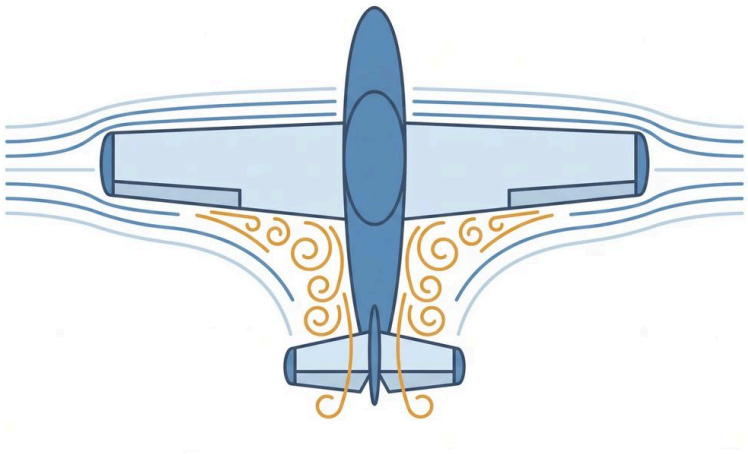
Every one of those factors changes the *speed* at which the wing reaches the critical angle; none of them changes the *angle* itself, which is fixed by the shape of the aerofoil section. Gross weight, altitude, bank angle and a steady climb all leave the stalling angle where it was. This is precisely why certified stall protection senses angle of attack rather than airspeed: one threshold remains valid at every weight, altitude and bank angle, whereas a speed threshold would need constant correction. It also explains why the stalling equivalent airspeed for a given weight and configuration is the same at all altitudes even though the true airspeed at the stall keeps rising.

The centre of pressure reaches its **most forward position at the stalling angle**. Beyond that angle the leading-edge suction peak collapses as the flow separates and the centre of pressure moves sharply rearward, producing a nose-down pitching moment that helps the aeroplane lower its nose and recover.

How the Stall Spreads Across the Span

A wing does not stall everywhere at once. Designers deliberately arrange for the **root to stall first** — by washout, by a different root section, or by stall strips — so that the outboard wing keeps flying and the ailerons remain effective while the separated wake from the root beats over the tailplane and gives the pilot airframe buffet as a natural warning. If the tip stalls first the opposite happens: lift is lost outboard and behind the centre

of gravity, which on a swept wing produces a wing drop combined with a **pitch-up** just as control authority is disappearing. On a T-tail aircraft, a fully developed stall can leave the tailplane immersed in the wing's separated wake, so that the elevator becomes ineffective and the aircraft is locked into a deep stall — the reason those types are fitted with a stick pusher rather than a shaker alone.



Root stalls first: the outboard wing and ailerons keep flying while the root's wake buffets the tailplane.

⚠ Stall Warning — Maintenance Points

The angle of attack vane measures the **local** flow direction at its mounting position and is calibrated for that position, so a vane installed on the wrong side, indexed to the wrong datum, or obstructed by tape, paint or a badly faired repair will give a false margin. Vane heaters must be serviceable before flight in icing conditions — a frozen vane can sit at a benign angle and never trigger. The stall warning, stick shaker and, where fitted, stick pusher have scheduled functional tests, and the pusher circuit normally has an inhibit or disconnect facility that must be confirmed in the correct state before release to service.

❄ Aerofoil Contamination

Contamination of the wing surface — even a very thin layer — disrupts the boundary layer and degrades aerodynamic performance. This is one of the most safety-critical maintenance considerations.

Contaminant	Effect
Frost	Even a thin layer of frost roughens the surface, causing early boundary layer transition. Can reduce lift by up to **30%** and increase stall speed. Aircraft must be de-iced before departure — the "clean aircraft concept."
Ice	Changes the aerofoil shape, reducing lift and increasing drag. Can add significant weight. Rime ice (rough, opaque) is worse than glaze ice (smooth, clear) for drag increase.

Contaminant	Effect
Rain (heavy)	Water film increases surface roughness, adds weight, and can cause premature boundary layer transition. Typically **5–15% lift reduction** in very heavy rain.
Insects	Insect residue near the leading edge disturbs the boundary layer. On laminar-flow wings, can cause significant drag increase. Aircraft operating in tropical regions require regular leading-edge cleaning.

⚠ Clean Aircraft Concept

EASA regulations require that an aircraft must be **"clean" — free of all frost, ice, snow, and slush** — before take-off. This is known as the **clean aircraft concept**. Ground de-icing and anti-icing procedures (using heated Type I/II/III/IV fluids) are applied to ensure compliance. The **holdover time** is the period during which anti-icing fluid remains effective after application; if exceeded before take-off, the aircraft must be re-treated.

Why So Little Contamination Does So Much

The damage is done in the first few per cent of chord, where the boundary layer is at its thinnest and where the suction peak that produces most of the lift is generated. Roughness there trips the layer to turbulent flow immediately, thickens it, and leaves it with far less energy to climb the adverse pressure gradient behind the point of maximum thickness. Separation therefore begins earlier along the chord and at a **lower angle of attack than on a clean wing**, so the aeroplane loses not only maximum lift but part of its margin to the stall.

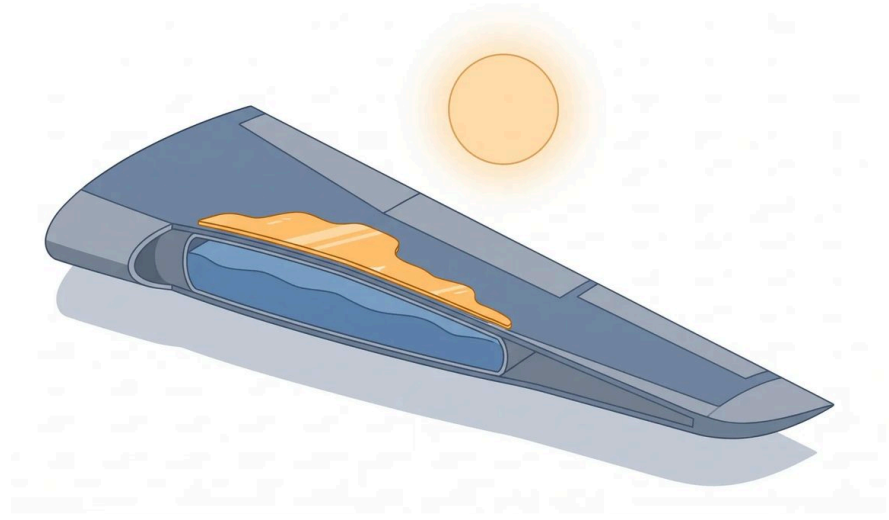
⚠ The Warning May Not Come First

Stall protection is calibrated to the **clean** aerofoil. If contamination has lowered the angle at which the wing actually separates, the aircraft can stall *before* the angle of attack vane reaches its trigger angle, so the stick shaker fires late or not at all. Contamination is also rarely symmetrical — one wing shaded from the sun, or one wing over a fuel tank still cold from cruise — and the more contaminated wing stalls first, giving an uncommanded roll at low altitude with the ailerons already close to their limit. This combination, not the raw lift loss alone, is what makes wing contamination lethal.

Where Contamination Hides

- **Cold-soaked fuel frost** — after a long sector the fuel in the tanks is far colder than the ambient air, and frost or clear ice forms on the upper wing skin above the tanks even on a warm, dry day with no precipitation. It can be almost invisible and is found by touch.
- **Clear ice on the lower surface** — forms over the same cold tank skins; upper-surface clear ice is the more serious case because it is where the lift is made and it is the hardest of all contamination to see.
- **Quiet areas** — hinge lines, control surface gaps, the fin and stabiliser leading edges, and the wing-to-fuselage fairing collect frozen deposits and dried thickened fluid that can re-hydrate and refreeze, in the worst case restricting a control surface.

- **Probes, vanes and drains** — the pitot heads, static ports, total air temperature probe, angle of attack vanes and drain masts must all be clear, and must not be sprayed directly during de-icing.



Cold-soaked fuel chills the upper wing skin above the tanks, so frost forms there even on a warm, dry day.

How the Fluids Work

The two functions are distinct. **De-icing** removes what is already on the aircraft, and is done with a hot, thin, unthickened fluid whose heat does the work; it leaves almost no protection behind. **Anti-icing** puts down a thickened fluid that stays in place as a protective film and absorbs falling precipitation. Thickened fluids are pseudoplastic: they are viscous at rest but shear-thin under the airflow of the take-off roll and flow off the wing before rotation — which is why they are only cleared for aircraft whose rotation speed is high enough to shed them, and why the wing must not be walked on or wiped after application. When both are needed the treatment is done as a two-step process, and the holdover clock starts at the beginning of the *final* application, not when the truck drives away.

Maintenance rules follow from the same physics. Frozen deposits are never chipped, scraped or knocked off, because the damage left in the skin and the leading edge is itself an aerodynamic defect and can also mask corrosion; approved fluid or warm hangarage is the only method. Only fluids to the correct specification and mix ratio may be used, and dilution is checked with a refractometer against the type approved for the conditions. Insect residue and dried fluid are removed with approved cleaners rather than abrasives, since polishing through the paint on a laminar-flow leading edge trades one contamination problem for a permanent surface-finish one.

Stability

Stability is the tendency of an aircraft to return to its original state after being disturbed (e.g., by a gust). There are two aspects:

- **Static stability** — the initial tendency after a disturbance:
  - **Positive** — tends to return to original state (nose drops after gust pitches it up)
  - **Neutral** — remains in the new state (no tendency to return or diverge)
  - **Negative** — tends to diverge further from original state (undesirable)
- **Dynamic stability** — describes the motion over time:
  - **Positive** — oscillations decrease in amplitude over time (returns to equilibrium)
  - **Neutral** — oscillations continue at constant amplitude (neither increasing nor decreasing)
  - **Negative** — oscillations increase in amplitude (divergent — dangerous)

Three Axes of Stability

Axis	Motion	Primary Stabilising Surface	Key Factors
Longitudinal (pitch)	Nose up/down about lateral axis	Horizontal stabiliser (tailplane)	CG position relative to CP; tailplane area and moment arm
Lateral (roll)	Wing up/down about longitudinal axis	Wing dihedral, swept wings	Dihedral angle, wing sweep, high/low wing position
Directional (yaw)	Nose left/right about vertical axis	Vertical stabiliser (fin)	Fin area and moment arm; fuselage side area

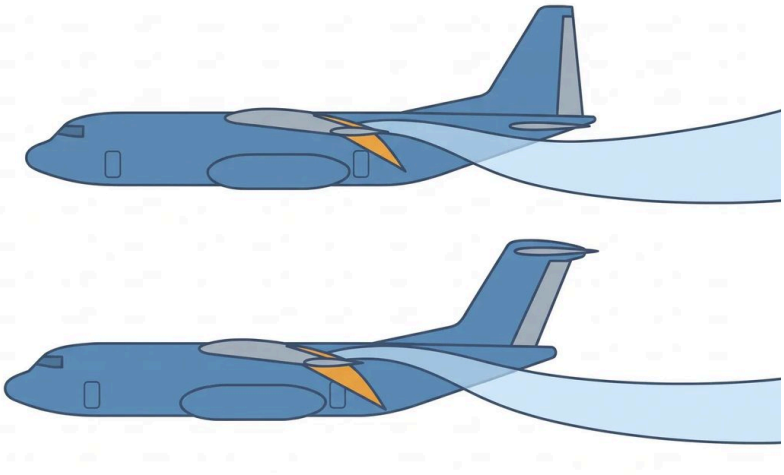
Longitudinal Stability: Centre of Gravity and Neutral Point

The restoring moment in pitch comes from the tailplane. A gust that raises the nose increases the angle of attack of the wing *and* of the tailplane; the tailplane's extra lift acts on a long arm behind the centre of gravity and produces a nose-down moment that opposes the disturbance. Whether the aircraft as a whole is stable depends on where the centre of gravity lies relative to the **neutral point** — the position at which the wing and tail contributions exactly cancel. With the CG ahead of the neutral point the aircraft is statically stable in pitch, and the distance between them, expressed as a percentage of the mean aerodynamic chord, is the **static margin**.

This is why the loading limits matter to the engineer as much as to the loadmaster. A forward CG gives a large static margin: the aircraft is very stable, but stick forces are heavy, more tail download is carried, the stall speed is higher and more elevator authority is consumed in the landing flare. An aft CG gives light stick forces and less trim drag, but a small static margin, sluggish natural recovery from a disturbance and, at the limit, a nose-up tendency at the stall that the elevator may be unable to overcome.

Configuration changes disturb this balance and the trim system has to absorb them. Lowering large trailing-edge flaps adds camber and, in the case of a Fowler flap, moves the flap physically aft, so the wing's centre of pressure shifts rearward and generates a **nose-down pitching moment**. On a conventional low-set tailplane

the increased downwash behind the flapped wing strikes the tail and produces a compensating nose-up effect, so the net change the pilot feels is modest. On a **high-set T-tail** that downwash largely passes below the tailplane, nothing offsets the wing's moment, and the aircraft is left with a net nose-down pitch change on flap extension that the stabiliser trim must take out.



Flaps down: downwash strikes a low tailplane and offsets the nose-down moment, but passes below a T-tail.

How Lateral and Directional Stability Are Produced

Lateral stability works through sideslip, not through roll directly. When a wing drops, the aircraft slides towards the low wing; with dihedral, the low wing then meets that sideways component of airflow at a greater effective angle of attack than the raised wing, generates more lift and rolls the aircraft level again. Wing sweep produces the same effect by a different route — the wing angled into the slip presents less effective sweep and a longer effective span to the flow — and a high wing adds a further contribution of its own. In a sideslip the fuselage deflects the oncoming flow upwards over the windward wing root and downwards over the leeward one, raising the effective angle of attack on the wing that is into the slip and lowering it on the other, which rolls the aircraft back towards level. The side (keel) area of the fin and the upper fuselage adds to this because the net side area lies **above** the centre of gravity, so the side force the slip generates acts on a moment arm that rights the aircraft; net side area acting **below** the centre of gravity does the opposite and rolls the aircraft further into the slip, which is why a deep lower fuselage or a large ventral fin subtracts from lateral stability rather than adding to it. The pendulum argument often quoted alongside the high wing is a separate one about **mass** rather than area — the fuselage and payload hang below a high wing — and the two should not be run together: it is the mass that has to be low, and the side area that has to be high. Modern swept, high-wing transports frequently need **anhedral** to subtract some of this effect, because too much lateral stability is as troublesome as too little.

Directional stability comes from the fin and the fuselage side area behind the centre of gravity. Because the fin is a symmetrical section it develops no side load at all while it is aligned with the airflow; a side load appears as soon as it meets the flow at an angle, and that angle can be produced **either by the aircraft yawing into a**

sideslip or by deflecting the rudder, which cambers the section. Both routes are valid, which is what allows the fin to act as a weathercock in a sideslip and as a control surface support at the same time.

Dynamic Modes

Mode	Motion	Typical character
Phugoid	Slow exchange of height and speed at almost constant angle of attack	Long period, lightly damped; easily controlled by the pilot, so weak damping is acceptable
Short-period pitch oscillation	Rapid oscillation in angle of attack at almost constant speed	Must be heavily damped and die out quickly, or the aircraft is unflyable
Dutch roll	Combined yaw and roll oscillation — strong lateral stability against weak directional stability	Characteristic of swept-wing transports; controlled by a yaw damper acting on the rudder
Spiral	Slow divergence in a steepening banked descent — strong directional stability against weak lateral stability	Often mildly divergent by design; slow enough for the pilot or autopilot to correct

Dutch roll and spiral instability sit at opposite ends of the same design compromise, so a swept-wing aircraft is deliberately given a slightly divergent spiral mode in exchange for a Dutch roll that the yaw damper can handle. That makes the yaw damper a required item on many types, and dispatch with it inoperative is normally limited by the minimum equipment list to reduced altitudes and speeds. For the engineer, the practical consequences of stability are rigging consequences: a wing repaired to the wrong dihedral, a fixed trim tab bent to an unapproved setting, a mis-rigged control cable tension or a stabiliser trim indication out of calibration all change the aircraft's natural behaviour before any pilot input is made.

Wing Planforms

Feature	Definition	Effect
Aspect ratio	Span² ÷ wing area (or span ÷ mean chord)	High AR = less induced drag, better L/D (gliders); Low AR = better manoeuvrability (fighters)
Wing sweep	Angle of the leading edge from perpendicular to fuselage	Increases critical Mach number; provides directional stability; increases stall speed; tip stall tendency
Taper	Reduction in chord from root to tip (taper ratio = tip chord ÷ root chord)	Reduces structural weight; approximates elliptical lift distribution; reduces induced drag
Washout (twist)	Wing tip has lower incidence angle than wing root	Root stalls before tip — preserves aileron control during stall; prevents tip stall on swept wings

Feature	Definition	Effect
Mean Aerodynamic Chord (MAC)	Chord of a rectangular wing with the same area, same pitching moment and same lift	Used as reference for CG position (expressed as % MAC)

Working with the Geometry

Planform Relationships

$$\text{mean chord} = \frac{S}{b} \qquad AR = \frac{b^2}{S} = \frac{b}{\text{mean chord}}$$

Where b is the span and S the wing area. Centre of gravity position is quoted against the mean aerodynamic chord:

$$\%\,\text{MAC} = \frac{\text{distance of CG aft of the MAC leading edge}}{\text{MAC}} \times 100$$

Worked Example — Span, Area, Chord and Aspect Ratio

A wing spans 50 ft and has an area of 200 ft². Its mean chord is $200 \div 50 = 4$ ft, and its aspect ratio is $50^2 \div 200 = 2500 \div 200 = 12.5$ — the same answer as span divided by mean chord, $50 \div 4 = 12.5$.

If that aircraft's MAC were 4 ft and the centre of gravity were found 1 ft aft of the MAC leading edge, the CG would be at $(1 \div 4) \times 100 = 25\%$ MAC.

Sweep — What It Buys and What It Costs

Only the component of the airflow perpendicular to the leading edge shapes the pressure distribution over the section; the spanwise component simply slides along the wing. Sweeping the wing therefore presents the section with an effective velocity reduced by the cosine of the sweep angle, so the aircraft can fly considerably faster before any point on the wing reaches the local speed of sound. The price is paid at low speed: the same cosine factor reduces the lift the section produces, so a swept wing has a lower maximum lift coefficient and needs more high-lift device area to achieve an acceptable approach speed.

Two different spanwise flows have to be kept apart here, because they run in opposite directions and both are examined. The first exists on **any** wing of finite span and is driven by the pressure difference between the surfaces. The lower surface is at higher pressure and the upper surface at lower pressure, and the two are free to meet around the tip, so on the **lower** surface the flow drifts outward, towards the tip, where it spills around into the tip vortex, while on the **upper** surface it drifts the other way, inward towards the root. Anything that increases the pressure difference strengthens both drifts, which is why **deploying large trailing-edge flaps increases the outward spanwise flow on the lower surface** and makes the tip vortices visibly stronger — the increase is on the lower surface, not the upper one.

Sweep adds a second, quite separate spanwise flow, and this one lives inside the boundary layer on the upper surface. Because the isobars on a swept wing run along the span rather than across it, there is a pressure gradient acting outboard, and the slow air deep in the boundary layer has too little momentum to resist it, so it is dragged steadily towards the tip as it travels aft. It arrives there thick and low in energy, so the tip is the first place to separate — and because the tip of a swept wing lies behind the centre of gravity, losing its lift produces a nose-up pitching moment at the very moment the ailerons are losing their grip. That single mechanism is the reason swept wings carry washout, wing fences, saw-tooth leading edges, vortex generators and stall strips, and it is why those items are not cosmetic. There is no contradiction between the two flows: the outboard drift belongs to the retarded air inside the upper-surface boundary layer on a swept wing, while the root-ward drift described above belongs to the main flow outside it.

Twist and Tip Devices

Washout can be built in two ways. **Geometric washout** is a physical twist in the structure that sets the tip at a smaller angle of incidence than the root, so as the aircraft is slowed the root reaches the stalling angle while the tip is still several degrees below it. **Aerodynamic washout** achieves the same end without twist, by using a section at the tip with a higher stalling angle than the section at the root. Either way the intent is identical: the root stalls first and the ailerons keep working.

Wingtip devices attack the tip vortex itself rather than the lift distribution. By obstructing the flow spilling from the lower surface to the upper surface at the tip, a winglet weakens the vortex and reduces induced drag; as a side effect the extra side area above the wing gives a small increase in static lateral stability. What it does not do is create an elliptical lift distribution — that is the job of planform taper and twist. A winglet also lengthens the effective load path, increasing the bending moment at the wing root, which is why a retrofit is a structural modification requiring reinforcement and a revised inspection programme, not simply a bolt-on aerodynamic part.

Control Around Three Axes



The three axes are mutually perpendicular and all three pass through the **centre of gravity**. That is the point about which the aeroplane actually rotates in flight, wherever the control force happens to be applied, because in free flight there is nothing else for it to pivot around. A control surface works by changing the camber of the surface it is attached to, which changes the aerodynamic force that surface produces; that force acts a long way from the centre of gravity, and the resulting moment is what turns the aircraft.

Control Moment

$$M = F \times d \qquad F \propto \tfrac{1}{2}\rho v^2 \times \text{area} \times \text{deflection}$$

The moment available depends on the force generated by the surface and on d, its distance from the centre of gravity. Because the force contains the dynamic pressure term, control effectiveness varies with the square of the indicated airspeed.

Two practical consequences follow. At low indicated airspeed the same full deflection produces only a weak moment, which is why the controls feel soft on the approach and why the aircraft is flown with larger, earlier inputs; and at high indicated airspeed a small deflection produces a very large moment and a very large structural load, which is why maximum control travel is restricted at speed on many types. A control input in roll is also different in kind from one in pitch or yaw: a held aileron deflection produces a steady *rate* of roll, because roll damping from the wings rises with roll rate until it balances the input, whereas a held elevator or rudder deflection settles the aircraft at a new attitude.

Primary and Further Effects

Input	Primary effect	Further effect and mechanism
Aileron	Roll about the longitudinal axis	Yaw away from the turn (adverse yaw): the down-going aileron adds lift and therefore induced drag on the rising wing. Once banked, the tilted lift vector also turns the aircraft.
Rudder	Yaw about the vertical axis	Roll in the same direction as the yaw: the outer wing travels faster through the air and generates more lift. A yaw to the right therefore tends towards a right-wing-low attitude.
Elevator	Pitch about the lateral axis	Change of speed and of angle of attack; in a sustained turn, back pressure is also what converts bank into turn rate rather than into a descent.

⚠ Full, Free and Correct Sense

After any work that disturbs a flying control system the check is not simply that the surface moves. It must move through its **full** range against the stops, **freely** without binding or interference throughout that range, and in the **correct sense** for the cockpit input — verified by an engineer watching the surface, not deduced from the cockpit. Reversed rigging is one of the classic fatal maintenance errors, and it is why control systems are subject to a duplicate inspection by a second qualified person after assembly, adjustment or repair.



Control sense is verified by an engineer watching the surface itself, not deduced from the cockpit.

Primary Flight Controls

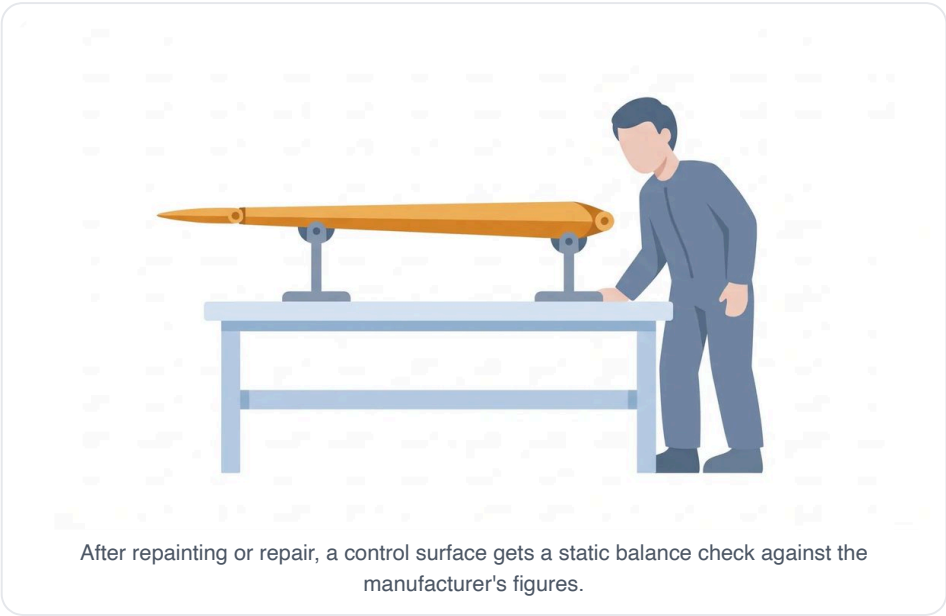
Control	Location	Axis	Movement	Cockpit Control
Ailerons	Wing trailing edge (outboard)	Longitudinal (roll)	Move differentially — one up, one down	Control wheel/stick left/right
Elevators	Horizontal stabiliser trailing edge	Lateral (pitch)	Move together — both up or both down	Control column forward/backward
Rudder	Vertical stabiliser trailing edge	Vertical (yaw)	Moves left or right	Rudder pedals left/right

Hinge Moment, Balance and Feel

The force a deflected surface generates acts at some distance behind its hinge line, and the product of the two is the **hinge moment** the pilot or the actuator must overcome. It grows with dynamic pressure and with deflection, so an unassisted control that is pleasant at circuit speed can be immovable at cruise. Designers reduce it aerodynamically by placing part of the surface ahead of the hinge, either as a **horn balance** at the tip or as a **set-back (inset) hinge** along the span, so that the air load on the forward portion helps to drive the

surface in the direction of movement. Too much balance is as bad as too little: an over-balanced surface tends to run away from neutral of its own accord.

Mass balance is a different thing entirely and must not be confused with aerodynamic balance. Weights are fitted ahead of the hinge line so that the centre of gravity of the surface itself lies on or slightly ahead of the hinge, preventing the inertial coupling between wing bending and surface rotation that drives **flutter** — a destructive divergent oscillation that can remove a surface in seconds. This is why a control surface may not be repainted, repaired, re-skinned or have a balance weight disturbed without a static balance check to the manufacturer's figures afterwards, and why water trapped inside a surface is a serious airworthiness defect rather than a nuisance. On fully powered controls the pilot is disconnected from the air loads altogether, so an **artificial feel** system provides the missing cue, typically increasing feel force with dynamic pressure and adding a centring spring.



After repainting or repair, a control surface gets a static balance check against the manufacturer's figures.

Adverse Yaw and the Ways It Is Cured

- **Differential ailerons** — the linkage geometry gives the up-going aileron a larger angular travel than the down-going one, so the **down-going aileron moves through the smaller angle**. This raises the drag on the descending wing and reduces the drag difference that causes adverse yaw.
- **Frise ailerons** — the up-going aileron's blunt leading edge projects below the wing into the airflow, deliberately generating drag on the down-going wing, while the down-going aileron's nose stays shrouded.
- **Aileron-rudder interconnect** — a mechanical or electrical link applies a proportional amount of rudder with aileron so the yaw is cancelled at source.
- **Roll spoilers** — because a spoiler rises only on the down-going wing it adds drag exactly where it is wanted, so spoiler-assisted roll produces no adverse yaw at all.

Roll and Pitch Control at High Speed

Large transports rarely rely on a single pair of ailerons across the whole envelope. Outboard ailerons give a long moment arm and are ideal at low speed, but at high speed their loads can twist the outer wing enough to oppose the roll they are commanding — in the extreme, **aileron reversal**. Such types therefore lock out the outboard ailerons above a threshold speed, usually as a function of flap position, and use inboard ailerons together with roll spoilers instead. In pitch, the elevator handles manoeuvring while a **trimmable horizontal stabiliser** carries the steady load, giving far greater trim authority than a tab could and leaving the full elevator range available for manoeuvre; on smaller types the whole surface may move as an all-flying tailplane with an anti-balance tab to restore feel. On the rudder, a travel limiter or ratio changer progressively restricts available deflection as speed rises, protecting the fin from loads it was never designed to take.

✈ Which Way Does the Trim Tab Go?

Work it out in two steps and the answer is never in doubt. **First**, decide which way the *control surface* must move to fix the problem. **Second**, put the tab the *opposite* way, because the tab is a servo whose air load drives the surface.

If the aircraft is **yawing to the left**, the rudder must move right, so the rudder trim tab is set to the **left**.

If the aircraft is flying **left wing low**, the left aileron must move down to make more lift on that side, so the left aileron trim tab is moved **up** — and in the cockpit the aileron trimmer is moved up, which drives that tab up and the aileron down.

↔ Spoilers, Lift Dumpers and Speed Brakes

A spoiler panel works by **deliberately separating the flow**. Raised into the airstream, it destroys the low-pressure region over the wing behind it, so the suction that was producing lift disappears and the wake it leaves behind adds a large amount of form drag. That combination — lift down and drag up at the same time, with no change to the wing's speed capability — is what makes one family of panels able to do three quite different jobs.

One Panel Set, Three Functions

On a transport aircraft the roll, speed-brake and ground-spoiler demands are summed in a **mixer** before they reach the actuators. With the speed brake already partly out, a roll demand drives the panels on the down-going wing further up while retracting those on the up-going wing towards faired, so roll authority is preserved throughout the speed-brake range. Roll spoilers normally have a small deadband around the neutral control-wheel position so that they stay stowed in cruise, and they are valuable precisely where ailerons are weakest — at high speed, where aileron loads twist the outer wing, and in any situation where adverse yaw is unwelcome, since a spoiler adds its drag on the wing that is going down.

In flight the same panels used symmetrically act as **speed brakes**, letting the crew steepen the descent at a constant speed or decelerate in level flight without shutting the engines to idle and losing bleed and anti-ice supply. Some airframe buffet with speed brake extended is normal and expected. Deployment is typically restricted or automatically inhibited with flaps extended, at low airspeed, at high angle of attack and close to the ground, because dumping lift there would be exactly the wrong thing to do.

Why Lift Dumping Shortens the Landing Roll

The braking force a tyre can generate is the product of the friction coefficient and the **normal load** pressing that tyre onto the runway. Immediately after touchdown the wing is still carrying a large part of the aircraft's weight, so the normal load is small and heavy braking would simply spin the wheels down into anti-skid release. Dumping the remaining lift transfers that weight onto the wheels within a second or two, which is what makes the wheel brakes and the anti-skid system effective, and the panels' own drag contributes as well. The same logic applies to a rejected take-off, where automatic deployment is armed for the take-off roll for exactly that reason. Deployment is normally armed by the crew before landing and triggered by evidence that the aircraft is on the ground — landing gear compression, wheel spin-up, thrust lever position, or a combination of these, depending on type.

⚠ Spoiler System Hazards and Defects

Failure to deploy on landing lengthens the stopping distance far more than the lost drag alone would suggest, because the brakes cannot bite until the weight is on the wheels. **Asymmetric deployment** produces an uncommanded roll, which is why these systems carry comparators, hydraulic isolation or mechanical interconnects to detect and contain it. In maintenance, treat every panel as live: a powered spoiler can slam closed with enough force to sever a hand, so systems must be depressurised and ground locks or pins fitted to the AMM procedure before anyone works in the panel bay. Panels are usually composite honeycomb, so they are vulnerable to water ingress, disbond and edge-seal damage and are checked by tap test or the approved NDT method; a droop with the system depressurised is normal on many types, but a panel that floats up in flight, or does not sit within the rigging tolerance of the wing contour, is a drag and handling defect in its own right.

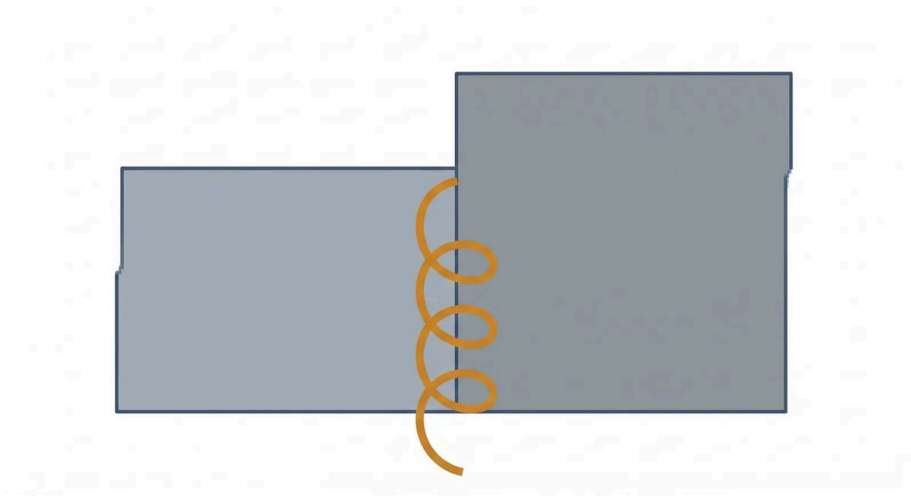
Wing Fences and Saw-Tooth Leading Edges

Both devices attack the outboard boundary-layer drift described under Wing Planforms above, and both do it at one station part way out along the span, upstream of the aileron, rather than by altering the section or the twist. That is what separates them from washout. Washout accepts the drift and simply arranges for the root to reach the stalling angle first; a fence or a saw-tooth interrupts the drift itself, so the outboard panel starts its boundary layer again and keeps it thin and energetic instead of merely stalling later than the root. The distinction matters on a wing that already carries as much washout as its cruise drag budget will allow, because there the only way left to protect the aileron is to stop the drift rather than to compensate for it.

How Each One Works

A **fence** is a physical barrier standing proud of the surface and running roughly in the flow direction. It dams the spanwise migration, so the boundary layer outboard of the fence effectively starts again from scratch and is thin and energetic where it matters — normally just inboard of the aileron. At higher angles of attack the fence also sheds a vortex from its upper edge that reinforces the effect. Its cost is a small permanent drag penalty at every speed, since the plate is there whether it is needed or not.

A **saw-tooth** (dog-tooth) leading edge achieves the same barrier aerodynamically. The step where the outboard leading edge extends forward of the inboard one sheds a strong streamwise vortex that behaves like an invisible fence, and that vortex simultaneously mixes high-energy air into the outboard boundary layer. Because there is no plate standing in the airflow, the cruise drag penalty is smaller; the discontinuity also gives the outboard panel a slightly different section and chord, which shifts its stalling behaviour in the desired direction. Some types achieve a similar effect with small pylon-like **vortilons** under the leading edge, which are inert at cruise angles and only generate their vortex when the angle of attack rises.



The forward step of the outboard leading edge sheds a vortex that acts like an invisible fence.

✈ Maintenance Relevance

These are aerodynamic devices with a certificated shape, not brackets. A fence that has been dented, distorted, shortened during a repair or refitted with an unapproved fillet no longer dams the flow as tested, and the effect appears only where it is least welcome — near the stall. The same applies to erosion or repeated blending that rounds off the sharp step of a saw-tooth, since the strength of the vortex depends on that sharp discontinuity. Restore the original profile to the SRM, keep the leading edge clean and free of contamination in this region, and never fair over or "tidy up" a step that appears to serve no purpose.

Boundary Layer Control: Vortex Generators and Stall Strips

These two devices are often mentioned in the same breath but they pull in opposite directions. A vortex generator is fitted to **delay** separation where it is unwanted; a stall strip is fitted to **provoke** separation where it is wanted first. Both work by manipulating the boundary layer rather than by adding lift, and both are small enough to be overlooked on a walk-round while being entirely load-bearing for the aircraft's certified handling.

Vortex Generators — Mechanism and Placement

Each vane is set at an angle to the local flow, so it works like a tiny wing and sheds a trailing vortex from its tip. That vortex **takes energy from the free stream and mixes it down into the boundary layer**, which is the whole point: the re-energised layer can then carry on against the adverse pressure gradient instead of stalling out against it. Note carefully what they do *not* do — they do not reduce the adverse pressure gradient itself and they do not redirect the flow; they simply give the boundary layer the momentum it needs to cope.

Vane height is chosen to be comparable with the local boundary layer thickness: too short and the vane sits buried in slow air with nothing to mix, too tall and it adds drag without doing useful work at the surface. They are installed in rows just upstream of where separation would otherwise begin, either all set the same way (co-rotating) or in mirrored pairs (counter-rotating), and their use extends well beyond the wing upper surface — ahead of control surfaces to keep them effective at large deflections, on engine nacelles and strakes to organise the flow onto the wing at high angles of attack, and on aft fuselage upsweeps to reduce separation drag. The price is a small parasite drag penalty that is paid at every speed, which is why they are used where the separation problem justifies it rather than everywhere.

Stall Strips — Mechanism and Placement

A stall strip presents a sharp edge to the flow at the inboard leading edge. At low angles of attack the flow negotiates it with almost no penalty, but as the angle of attack rises the flow can no longer turn the sharp corner and separates cleanly from that point, so the inboard wing gives up first. That is exactly the intent: the resulting separated wake beats over the tailplane and gives the pilot honest airframe buffet as a stall warning, while the outboard wing and its ailerons are still flying. The angle at which the strip trips is governed by how far it is positioned around the curve of the leading edge, which is why its location and its symmetry between left and right wings are held to tight tolerances.

⚠ Small Parts, Certificated Effects

A row of vortex generators with several vanes missing, bent or flattened by ground equipment will let the flow separate early at that spot, showing up as buffet, a higher stall speed, or a wing that drops at the stall. Replacements must go back in the **exact position and angular orientation** using the manufacturer's template or drawing dimensions — never by eye, never realigned to look tidy — and the count must be checked against the drawing after any leading-edge or panel replacement. Stall strips deserve the same discipline: a strip refitted a little further around the leading edge, or fitted to one wing and not the other after a bird-strike repair, changes the stalling angle asymmetrically and can turn a straight-ahead stall into a wing drop. Neither device may be filled, faired, thickened with successive paint schemes or "improved", and both must be confirmed present and correctly positioned after painting and after any leading-edge repair.

Secondary Flight Controls

Trim Tabs and Tab Types

Tab Type	Purpose	Operation
Trim tab	Relieves control force in steady flight	Pilot-adjustable; moves opposite to control surface
Balance tab	Reduces hinge moment (lightens control feel)	Mechanically linked; moves opposite to control surface automatically
Anti-balance tab	Increases hinge moment (adds feel/resistance)	Moves in SAME direction as control surface
Servo tab	Pilot moves tab only; aerodynamic force moves main surface	Used on large control surfaces; reduces pilot effort significantly
Spring tab	Provides aerodynamic assistance at high speeds only	Spring preload means tab only activates when forces exceed threshold

High-Lift Devices

Leading edge devices and **trailing edge devices** are used to increase $C_{L_{max}}$, allowing the aircraft to fly at lower speeds during take-off and landing.

Device	Location	Mechanism	Effect
Slats	Leading edge	Extend forward and down, creating a slot	Re-energise boundary layer; increase stall angle; increase $C_{L_{max}}$
Slots (fixed)	Leading edge	Permanent gap in the leading edge	Direct high-energy air to upper surface; delay stall
Plain flap	Trailing edge	Hinged portion deflects downward	Increases camber; increases C_L and drag

Device	Location	Mechanism	Effect
Split flap	Trailing edge (lower surface)	Only lower portion deflects down	Increases C_L; very high drag increase
Slotted flap	Trailing edge	Creates slot between wing and flap when deployed	Re-energises boundary layer; higher $C_{L_{max}}$ than plain flap
Fowler flap	Trailing edge	Extends aft AND deflects, increasing wing area	Highest $C_{L_{max}}$ of all flap types; used on most transport aircraft

Drag-Inducing Devices

- **Spoilers (flight spoilers)** — panels on the upper wing surface that deploy upward into the airflow, reducing lift and increasing drag. Used in flight for roll assistance (differential spoilers) and descent control (speed brakes).
- **Ground spoilers (lift dumpers)** — deploy automatically or manually after landing to "dump" remaining lift, putting the full aircraft weight on the wheels for effective braking. Typically deploy to a higher angle than flight spoilers.
- **Speed brakes** — may be wing-mounted spoilers or fuselage-mounted devices. Increase drag to control airspeed during descent without reducing power excessively.

Boundary Layer Control Devices

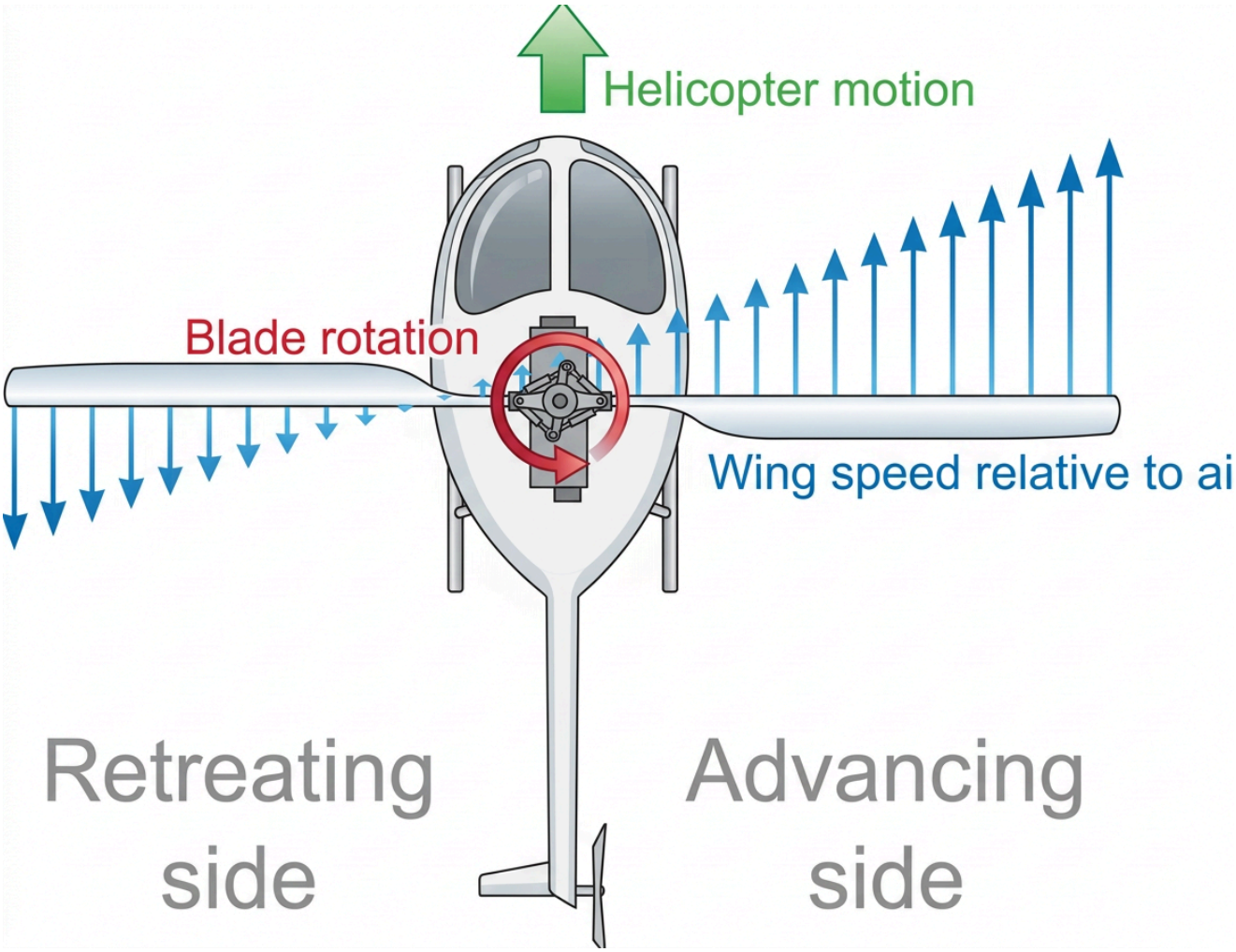
- **Vortex generators** — small metal tabs on the wing upper surface that create small vortices, mixing high-energy free-stream air into the boundary layer to delay separation. Common on light aircraft and on engine nacelle surfaces.
- **Wing fences** — vertical plates on the upper wing surface that prevent spanwise flow (airflow migrating from root to tip on swept wings), delaying tip stall.
- **Saw-tooth leading edge** — a sharp notch in the wing leading edge that creates a controlled vortex at high AoA, delaying stall at the outboard wing section.
- **Stall strips** — small triangular strips on the inboard leading edge that cause the root to stall first, ensuring the ailerons remain effective during a stall approach.

Module 12 — Helicopter Aerodynamics, Structures and Systems

12.1 — Theory of Flight — Rotary Wing Aerodynamics

Rotary wing aerodynamics is fundamentally different from fixed-wing flight. A helicopter generates lift by spinning rotor blades — effectively rotating wings — through the air. This creates unique aerodynamic phenomena that do not exist on aeroplanes: dissymmetry of lift, gyroscopic precession, blade flapping, Coriolis effect on blades, and the ability to hover, fly sideways, and autorotate. Mastering these concepts is essential for every helicopter technician.

Terminology



Term	Definition
Rotor disc	The circular area swept by the rotor blades as they rotate. Think of it as a translucent disc created by the spinning blades.
Tip path plane (TPP)	The imaginary plane described by the blade tips as they rotate. In a hover, the TPP is nearly horizontal. In forward flight, the TPP tilts forward (disc tilt).

Term	Definition
Advancing blade	The blade moving in the same direction as the helicopter's forward motion (ψ = 0°–180° for CCW rotation viewed from above, right side). Its airspeed = rotational speed + forward speed.
Retreating blade	The blade moving opposite to the helicopter's forward motion (ψ = 180°–360°, left side). Its airspeed = rotational speed – forward speed.
Blade azimuth (ψ)	Angular position of a blade measured from the nose (ψ = 0° over the nose, 090° over the advancing side, 180° over the tail, 270° over the retreating side).
Blade span	The length of the blade from root to tip.
Blade chord	The distance from the leading edge to trailing edge of the blade aerofoil cross-section.
Pitch angle (θ)	The angle between the blade chord line and the plane of rotation (the hub plane). Controlled by the pilot through collective and cyclic inputs.
Angle of attack (α)	The angle between the blade chord line and the *relative airflow* at any given blade section. Determines lift production. Not the same as pitch angle.
Coning angle	The upward angle of the blades from the hub due to lift forces bending the blades upward, balanced by centrifugal force pulling them outward. A spinning rotor forms a shallow cone shape.

Disc Geometry: Area, Disc Loading and Solidity

Three derived quantities turn the vocabulary above into numbers an engineer can work with. The **disc area** is the area of the circle the tips sweep, so it depends only on rotor radius. The **disc loading** is the aircraft weight carried by each square metre of that circle, and it is the single most useful figure for predicting how much downwash a rotor will produce and how much power it will need to hover. **Solidity** is the fraction of the disc that is actually blade.

Disc area: $A = \pi R^2$ (R = rotor radius, tip-to-tip diameter $= 2R$)
Disc loading: $DL = W/A$ (weight per unit disc area)
Solidity: $\sigma = \dfrac{\text{total blade area}}{\text{disc area}} = \dfrac{Nc}{\pi R}$ (N = number of blades, c = blade chord)

Solidity is raised by **adding blades** or by **widening the chord** of the blades already fitted, because both put more blade area inside the same swept circle. Changing the blade *angle* does not change solidity at all — pitch alters how hard each blade works, not how much blade there is. A higher-solidity rotor can carry more thrust before the blades reach their stalling angle, which is why heavy-lift and high-speed designs use more blades rather than simply more pitch, but the extra blade area also drags more air round with it and costs profile power in the cruise.

Worked example — disc loading and downwash: take a rotor of 10 m diameter ($R = 5$ m) on an aircraft weighing 2,000 kg.
Disc area $A = \pi \times 5^2 \approx 78.5\ \mathrm{m}^2$.
Disc loading $DL = 2{,}000 / 78.5 \approx 25\ \mathrm{kg/m}^2$, i.e. about 245 N/m².
Halving the rotor diameter to 5 m would cut the disc area to about 19.6 m² and **quadruple** the disc loading to roughly 100 kg/m² — because area varies with the *square* of the radius. That is why a small rotor blows a far more violent downwash and needs far more power to hover than a large one lifting the same weight.



Same weight, half the rotor diameter: a quarter of the disc area, four times the disc loading and downwash.

Further Terms You Will Meet on the Ramp

Term	Meaning and why it matters
Plane of rotation (hub plane)	The plane perpendicular to the rotor shaft. It is fixed to the aircraft, unlike the tip path plane, which is free to tilt as the blades flap. When the blades are not flapping relative to the shaft the two planes coincide; once cyclic is applied, or once dissymmetry of lift starts the blades flapping, the tip path plane tilts away from the hub plane and the angle between them is the flapping angle.
Rotational velocity	The speed of a blade section due to rotation alone, $V = \Omega r$. It is zero at the axis and greatest at the tip, so an untwisted blade at a single pitch angle would try to produce almost all of its lift in the outer third of the span.
Tip speed	Rotational velocity at the tip, ΩR. It is the design constraint that fixes rotor RPM: high enough for adequate centrifugal stiffening and stored energy, low enough that the advancing tip stays clear of compressibility in forward flight.
Blade twist (washout)	A built-in reduction of blade pitch from root to tip. It compensates for the rise in rotational velocity along the span and spreads the lift more evenly, which lowers induced power in the hover and reduces tip loading. Twist is manufactured into the blade — it is not adjustable in service.

Term	Meaning and why it matters
Induced flow (downwash, inflow)	The column of air the rotor accelerates downwards through the disc. It is the *consequence* of producing thrust, and it deflects the relative airflow at every blade section downwards, reducing the angle of attack the section actually sees.
Inflow angle (ϕ)	The angle by which induced flow tilts the relative airflow below the plane of rotation at a given blade station. Angle of attack is what is left of the pitch angle after the inflow angle has been subtracted.
Total reaction	The single resultant aerodynamic force on a blade section. Resolved perpendicular to the relative airflow it is lift; resolved parallel to it, it is drag. Resolved instead *in the plane of rotation*, it is the force that either helps drive the blade round or holds it back — the idea the whole of autorotation rests on.
Feathering	Rotation of the blade about its own spanwise axis to change pitch, through the pitch-change (feathering) bearing. This is what the collective and cyclic actually command.
Flapping	Vertical movement of the blade about the horizontal (also called lateral or flapping) hinge, or the equivalent flexing of a hingeless blade root.
Dragging (lead-lag)	Horizontal, in-plane fore-and-aft movement of the blade about the vertical (drag) hinge. Do not confuse it with flapping: flapping is vertical about a horizontal hinge, dragging is horizontal about a vertical hinge.
Reverse flow region	A circle of disc near the retreating blade root where the aircraft's forward speed exceeds the local rotational velocity, so the air arrives over the trailing edge. It contributes no useful lift and grows as forward speed rises.
Blade station	A defined radial position along the blade, quoted as a distance from the centre of rotation. Balance charts, damage limits and repair schemes are all written against blade stations.



A blade station is a radial position measured from the centre of rotation; damage is recorded against it.

Pitch Angle and Angle of Attack — the Distinction the Exam Tests

These two angles are constantly confused and the difference between them is examined directly. **Pitch angle is mechanical**: it is the angle the blade has been physically twisted to by the control system, measured against the rotor's own reference plane, and the pilot sets it — equally on all blades with the collective, and cyclically once per revolution with the cyclic. **Angle of attack is aerodynamic**: it is the angle between the chord line and the airflow the section is actually meeting, and nobody sets it directly. It is what is left when the inflow angle is subtracted from the pitch angle.

Angle of attack at a blade section:

$$\alpha = \theta - \phi \qquad \text{where} \qquad \phi = \arctan\left(\frac{V_i}{\Omega r}\right)$$

θ = pitch angle, ϕ = inflow angle, V_i = induced (and any flapping) velocity through the disc, Ωr = rotational velocity at that station.

Worked example — why the two angles differ: a blade station is turning at $\Omega r = 150$ m/s with an induced velocity through the disc of 12 m/s.
Inflow angle $\phi = \arctan(12/150) = \arctan(0.08) \approx 4.6^\circ$.
With the collective set so that the pitch angle at that station is $\theta = 10^\circ$, the section flies at $\alpha = 10^\circ - 4.6^\circ \approx 5.4^\circ$.
Now let the blade **flap upward** at 3 m/s. The blade's own upward motion adds to the air arriving from above, so the total flow through the section becomes 15 m/s, ϕ rises to about 5.7°, and the angle of attack **falls** to roughly 4.3°. The pitch angle has not changed at all — the control system has been told nothing — yet the section is now producing less lift. That single arithmetic step is the whole mechanism of blade flapping.

Two consequences follow and both are examinable. First, **lift depends on angle of attack, never on pitch angle directly** — pitch only reaches the lift through the angle of attack it produces. Second, because induced velocity is not the same everywhere on the disc, two blades set to identical pitch by the same swashplate can be flying at quite different angles of attack at the same instant. This is why a rotor is tracked and balanced dynamically, in the air, and not simply set to matching pitch angles on the ground.

Aerodynamic Principles of Rotary Wing Flight

How a Rotor Blade Actually Produces Lift

A rotor blade is an aerofoil and it makes lift exactly as a wing does. As the section moves through the air the flow accelerates over the upper surface and slows over the lower surface, so the static pressure on top **falls** while the pressure underneath stays comparatively **high**. Integrated over the whole blade area, that pressure

difference is the lift force, and the larger share of it comes from the suction on the upper surface rather than from any push underneath. The rotor is *not* a screw boring its way into the air, and it does not lift by shovelling air downwards: the downwash is the accompanying reaction to the lift, not its cause.

The same relationship governs a blade section as governs any aerofoil, and it is worth committing to memory because almost every rotor phenomenon in this note is a consequence of it.

Lift of a blade section:

$$L = \tfrac{1}{2}\,\rho\,V^2\,S\,C_L$$

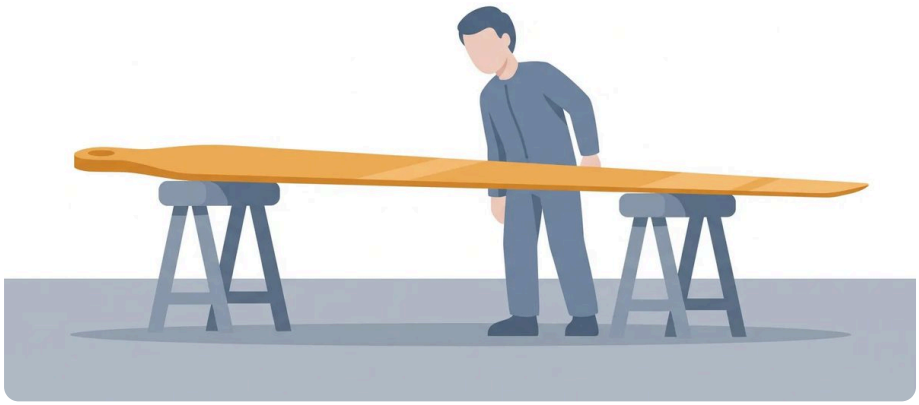
ρ = air density, V = the **relative airflow** at that section (not the aircraft's airspeed), S = section area, C_L = lift coefficient, which rises with angle of attack up to the stall.
Lift therefore varies with the **square of the relative airflow** and with the **angle of attack** — those are the two variables that matter, and every other rotor effect works by changing one of them.

Note carefully which velocity goes into that expression. For a rotor it is the velocity of the *blade section* through the air, which in the hover is purely rotational (Ωr) and in forward flight is the vector sum of rotational velocity, aircraft airspeed and the induced flow. The aircraft's airspeed on its own tells you nothing about what a given blade section is doing, which is why a helicopter can be stationary over a spot with its blades flying at several hundred knots.

Why Rotational Velocity Varies Along the Blade — and What Is Done About It

Every section of a blade completes a revolution in the same time, so a section at twice the radius travels twice as far and moves twice as fast. Because lift goes with the square of that velocity, a section at 80% radius meets four times the dynamic pressure of one at 40% radius. Left uncorrected, an untwisted blade would produce a lift distribution heavily concentrated near the tip: heavy bending loads at the root, a strong tip vortex, and poor efficiency because the air is being accelerated hard through a small annulus of the disc rather than gently through all of it.

Designers answer this with **blade twist** (washout), building a higher pitch angle into the root than into the tip. The reduced pitch outboard trims back the lift the high velocity would otherwise generate there, and the increased pitch inboard recovers lift from a region that would otherwise be nearly idle. The result is a flatter spanwise lift distribution, lower induced power in the hover, and a lower tip loading — which also means a weaker tip vortex for the following blade to fly through. Twist is rolled or moulded into the blade in manufacture; a technician cannot adjust it, and a blade found twisted outside its repair-scheme limits is scrap, not something to be corrected on the aircraft.



Twist is built into the blade in manufacture; a blade twisted beyond its repair-scheme limits is scrap.

Total Reaction, and the In-Plane Component

It is often more revealing to treat the aerodynamic force on a blade section as a single **total reaction** and then resolve it whichever way answers the question in front of you. Resolved perpendicular to the relative airflow it is lift; resolved along the relative airflow it is drag. But a rotor blade is constrained to travel in a circle, so the resolution that determines whether the rotor speeds up or slows down is a third one: the component of total reaction lying **in the plane of rotation**.

In powered flight that in-plane component acts **backwards**, against the direction of blade travel, and the engine has to supply torque continuously to overcome it. Tilt the relative airflow the other way — by putting air *up* through the disc instead of down through it — and the total reaction tilts forward of the rotor axis, so its in-plane component now acts **forwards**, in the direction of rotation. That forward in-plane component of the total reaction, opposing the blade's profile drag and keeping the rotor turning without engine power, is what is meant by the **autorotative force**. It is an aerodynamic force generated by the airflow, not a control input the pilot applies and not a lever position.

Aerofoil Choice: Keeping the Centre of Pressure Still

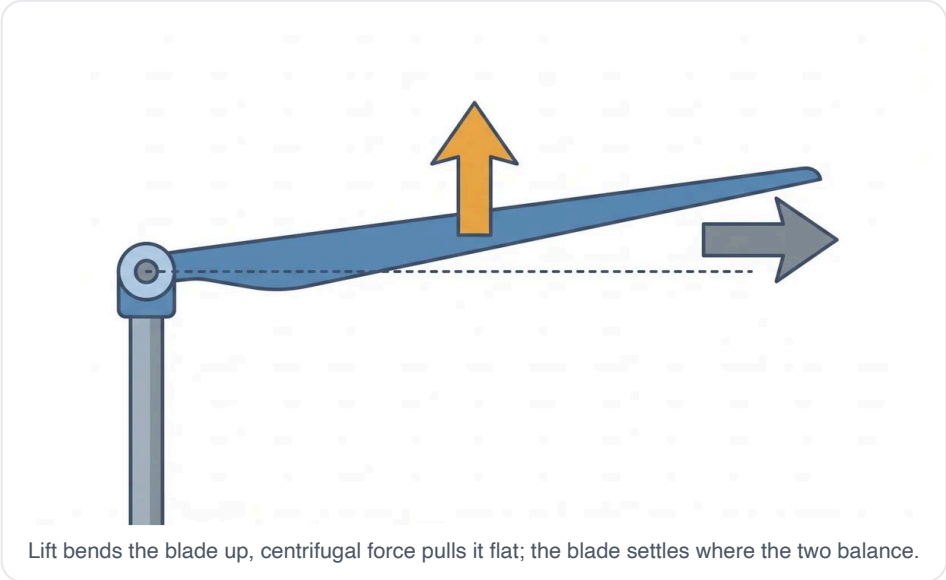
A rotor blade has a design problem a fixed wing does not: it is a long, slender, torsionally flexible beam whose angle of attack changes every revolution, and it is held at the root by a control system that must not be fed large fluctuating twisting moments. On a cambered aerofoil the centre of pressure **migrates along the chord** as angle of attack changes, and on a rotating blade that migration turns into a cyclic pitching (feathering) moment which twists the blade, feeds loads back into the pitch links, and drives control-system wear and vibration.

The classic solution, and the one the examination expects, is the **symmetrical aerofoil**: with equal curvature above and below the chord line the centre of pressure stays essentially fixed near the quarter-chord as angle of attack varies, so the section produces almost no pitching moment of its own. The design objective is best stated as a property rather than a shape — *the best rotor blade is one whose centre of pressure does not move* — and modern blades often reach that same objective with a carefully shaped cambered section whose pitching moment is engineered close to zero. A strongly cambered section with a wandering centre of pressure would impose exactly the torsional loading the head is designed to avoid.

Centrifugal Force, Coning and the Choice of Rotor RPM

The dominant force on a turning rotor is not aerodynamic at all. Centrifugal force on a blade rises with the square of rotational speed, $F_c = m\,\Omega^2 r$, and on a rotor at operating RPM it is very much larger than the lift the same blade produces. That force is what makes a set of thin, flexible blades behave as a rigid disc, and it is why rotor RPM is guarded so closely: the stiffness of the rotor is bought with speed.

The **coning angle** is the compromise between the two. Lift acts upwards along the blade and bends it up; centrifugal force acts outwards and pulls it back into the plane of rotation. The blade settles where the two moments about the flapping hinge balance, and the pair of forces that fixes the coning angle is therefore **lift and centrifugal force** — centrifugal force alone would hold the blades flat, and lift alone would fold them upwards. Working the two ends of that balance gives the directions to remember:



Lift bends the blade up, centrifugal force pulls it flat; the blade settles where the two balance.

- **More lift at constant RPM — coning angle increases.** Raising the collective, increasing gross weight, or pulling g in a turn all bend the blades further up.
- **Lower rotor RPM at constant lift — coning angle increases.** Centrifugal force falls with the square of RPM, so the same lift bends the blades further up. A low-RPM, high-weight condition gives the greatest coning of all, which is why rotor droop is treated as an emergency.

- **Excessive coning reduces the effective disc area**, because the blades are no longer sweeping the full circle. The rotor produces less thrust for the same pitch, which tempts more collective, which cones the blades further — the mechanism behind the classic low-RPM/overpitching trap.

The design RPM of a rotor is fixed chiefly by the **blades themselves — their mass and their radius** — and not by the engine or the gearbox, which are matched afterwards to whatever speed the rotor needs. The blades must turn fast enough to generate the centrifugal force that stiffens them and to store enough rotational kinetic energy ($E = \frac{1}{2} I \Omega^2$, where the inertia I depends on blade mass and the square of its radius) to give the pilot usable time and flare energy after a power loss. At the other end, tip speed must stay far enough below the speed of sound that the advancing tip in forward flight does not run into compressibility. Between those two limits sits a narrow band of usable rotor speed, and everything else in the drive train is designed around it.

Momentum Theory: Where the Power Goes in a Hover

Viewed as a whole rather than blade by blade, the rotor is a device that draws air in from above and pushes it out below at a higher velocity; the thrust equals the rate at which it changes the momentum of that air. This simple accounting gives a remarkably useful result for the hover, because it links induced velocity directly to disc loading — nothing else about the rotor need be known.

Induced velocity in the hover:

$$T = 2\rho A v_i^2 \quad \Longrightarrow \quad v_i = \sqrt{\frac{T}{2\rho A}} = \sqrt{\frac{DL}{2\rho}}$$

Induced power: $P_i = T\, v_i$, so $P_i \propto \frac{T^{3/2}}{\rho A}$.

Worked example — disc loading drives hover power: take the 25 kg/m² (245 N/m²) disc loading from the terminology section, at sea-level density $\rho = 1.225$ kg/m³.
$v_i = \sqrt{245/(2 \times 1.225)} = \sqrt{100} = 10$ m/s of induced flow through the disc.
Now quadruple the disc loading to 100 kg/m² by halving the rotor diameter. $v_i = \sqrt{980/2.45} = 20$ m/s — the downwash **doubles**, and since induced power is thrust multiplied by induced velocity, the power needed simply to hold the hover **doubles** as well for the same weight.
Direction to take away: a larger rotor for a given weight means lower disc loading, lower induced velocity and lower hover power. A small rotor is compact and structurally cheap but pays for it in fuel burn, downwash violence and hover performance.

The same expression explains the effect of air density, which is examined directly. Density appears under the square root, so as ρ **falls** — on a hot day, at high elevation, or both — the induced velocity for a given thrust **rises** and the power required to hover **rises** with it. At the blade, the immediate effect is simpler still: with less

dense air each section produces less lift at the same angle of attack, so to hold position the pilot must **increase the collective pitch**, raising the pitch of all blades equally until total rotor thrust is restored. Rotor RPM is not the lever for this — it is held constant by the governor — and cyclic only tilts the disc to change direction rather than changing total lift. The cost of that extra collective is extra torque, so the anti-torque system and the power margin are both squeezed on exactly the day the aircraft is least able to spare them.

Total power required: the power a rotor demands is the sum of three parts, and knowing which one dominates where explains most of the helicopter's performance curve. **Induced power** pays for making thrust and is largest in the hover, falling as forward speed rises. **Profile power** pays for dragging the blades round through the air and changes comparatively little until the blades approach their limits. **Parasite power** pays for pushing the fuselage, rotor head and undercarriage through the air, is negligible in the hover and grows very rapidly with airspeed. Add them and the total is a U-shaped curve with a distinct minimum in the middle of the speed range — the physical reason a helicopter both hovers and cruises less efficiently than it flies at its best-endurance speed.

Dissymmetry of Lift and Blade Flapping

Dissymmetry of lift is the unequal production of lift between the advancing and retreating sides of the rotor disc in forward flight. It is the most important aerodynamic phenomenon unique to helicopters.

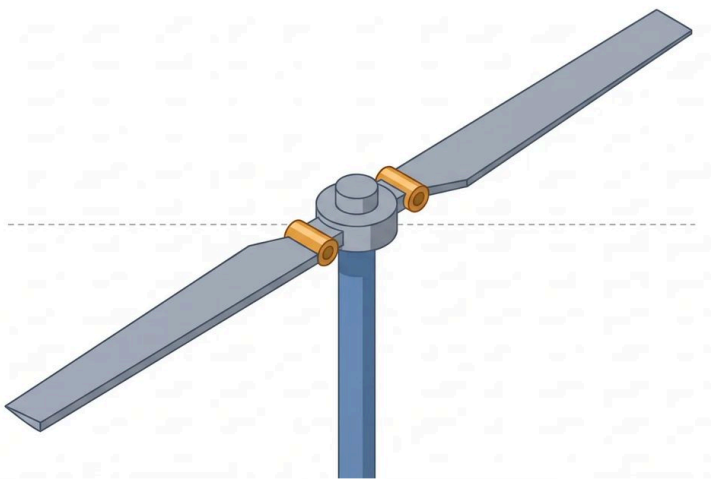
In a hover, all blades experience the same airspeed (rotational velocity only) at any given radial station, so lift is distributed symmetrically. But in forward flight, the advancing blade sees rotational speed **plus** forward speed, while the retreating blade sees rotational speed **minus** forward speed. For example, at 150 knots forward speed and 400 knots tip speed:

- Advancing blade tip speed: 400 + 150 = **550 knots**
- Retreating blade tip speed: 400 – 150 = **250 knots**

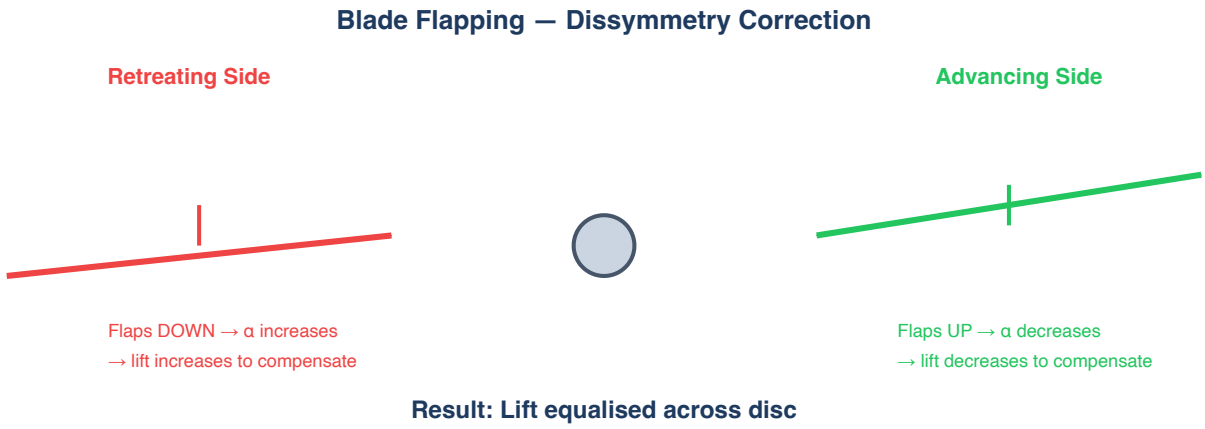
Since lift is proportional to velocity squared ($L \propto V^2$), the advancing blade would produce far more lift than the retreating blade. If uncorrected, this would roll the helicopter violently toward the retreating side.

Blade Flapping — The Solution

The solution is **blade flapping**. The blades are attached to the hub via **flapping hinges** (in a semi-rigid head the blades flap together about a teetering hinge, and in a bearingless head the blade root flexes). When the advancing blade generates excess lift, it flaps *upward*. Flapping up changes the relative airflow angle, **reducing** the blade's angle of attack and therefore reducing its lift. Conversely, the retreating blade flaps *downward*, increasing its angle of attack and increasing its lift. The result is that lift is **equalised** across the disc — dissymmetry of lift is automatically corrected by flapping.



Flapping hinges let the advancing blade rise and the retreating blade drop, evening out lift across the disc.



⚠ **Key Concept:** Blade flapping causes the **tip path plane to tilt**. Because of gyroscopic precession (see below), the flapping response is offset 90° from the aerodynamic input. The advancing blade reaches maximum flap-up not at ψ = 090° (where airspeed is maximum) but at approximately ψ = 000° (over the nose). Similarly, maximum flap-down occurs at approximately ψ = 180° (over the tail). This means the TPP tilts **rearward** in forward flight — the pilot must apply forward cyclic to compensate (blow-back effect).

Torque Reaction and Directional Control

Newton's Third Law dictates that when the engine drives the rotor clockwise (viewed from above), the fuselage will try to rotate **counter-clockwise** (torque reaction). Without a means to counteract this torque, the helicopter would spin uncontrollably. Three methods are used:

System	How It Works	Examples
Conventional tail rotor	A small variable-pitch rotor mounted vertically at the end of the tail boom. It produces a lateral thrust that opposes the main rotor torque. Yaw pedal inputs change the tail rotor blade pitch to increase or decrease anti-torque thrust, controlling helicopter heading. Consumes **8–15%** of engine power.	Most helicopters: Bell 206, Airbus H125, Sikorsky S-76
Fenestron (shrouded/ducted tail rotor)	A multi-blade fan enclosed within a duct (shroud) in the vertical tail fin. Functions like a conventional tail rotor but with safety and noise advantages: the shroud protects against contact with the blades, reduces noise (by shielding blade tips), and improves efficiency in crosswind and low-speed flight. The duct provides an additional aerodynamic benefit — it acts as a short-chord wing producing side force.	Airbus Helicopters: H135, H155, H160
NOTAR (NO TAil Rotor)	Uses a fan inside the tail boom to blow air through a slot along the boom surface (Coandă effect). The main rotor downwash flowing over the boom interacts with this blown air to create a lateral aerodynamic force (circulation control). A direct-jet thruster at the tail provides additional yaw control, especially in hover and low speed. No exposed tail rotor blades — safest system for ground personnel.	MD Helicopters: MD 520N, MD 600N, MD 902 Explorer

How Big Is the Torque, and What Changes It

Torque reaction is not a fixed quantity to be trimmed out once. The torque the transmission delivers to the rotor is set by the power being absorbed and the speed it is turning at, $Q = P/\Omega$, and because the governor holds rotor RPM essentially constant, **torque follows power almost directly**. Raise the collective and the blades take a bigger bite of air; power rises, torque rises, and the anti-torque system must produce a proportionally larger opposing moment or the nose swings. Lower the collective and the reverse happens. This is why the pedals are worked continuously in the hover and barely touched in the cruise, and why the pedal position that holds heading changes every time the power changes.

What the tail rotor supplies is a **moment**, not just a force: thrust multiplied by the distance from the tail rotor to the main rotor axis. That long lever is the reason a tail rotor absorbing a modest share of engine power can hold a large main rotor's torque, and it is also why tail boom length, tail boom straightness after a strike, and correct rigging of the whole pedal run are airworthiness items rather than convenience items.

Direction of Rotation, Yaw Tendency and Tail Rotor Drift

Once the tail rotor is producing side thrust, it is doing two things at once, and only one of them is wanted. The *moment* about the main rotor axis is what holds the heading. The *force* itself has nowhere to go: it acts on the whole aircraft and pushes it sideways. That unwanted sideways push is **tail rotor drift**, also called the translating tendency, and it acts toward whichever side the tail rotor is thrusting — regardless of whether the

tail rotor is mounted on the left or the right of the fin, because it is the direction of thrust and not the mounting position that produces the force. Working both directions of rotation through gives:

Main rotor direction (viewed from above)	Fuselage torque reaction	Tail rotor thrust must act	Resulting drift
Clockwise	Fuselage tries to turn anti-clockwise, i.e. the nose yaws to **port**.	To **port**, so that the tail is pushed left and the nose is held to starboard against the torque.	Aircraft drifts to **port**.
Anti-clockwise	Fuselage tries to turn clockwise, i.e. the nose yaws to **starboard**.	To **starboard**, so that the tail is pushed right and the nose is held to port against the torque.	Aircraft drifts to **starboard**, whichever side of the fin the tail rotor is mounted on.

The drift must be cancelled or the pilot would have to hold a permanent lateral cyclic offset just to hover over a spot. Two design fixes are common. The first is to **tilt the main rotor mast laterally** by a small angle, toward the side *opposite* the drift, so that the rotor thrust vector gains a small horizontal component pointing back against the tail rotor thrust; the aircraft then hovers with the fuselage level and the cyclic near neutral. The second is to build an equivalent lateral bias into the rigging of the cyclic control system so that a centred stick already commands the required disc tilt. Note precisely what this correction is for: it addresses the sideways **drift**. It does not counter torque — that remains the tail rotor's job — and it is not a roll correction.

The Rolling Couple, and Why Tail Rotor Height Matters

Tail rotor thrust and the opposing horizontal component of main rotor thrust are two side forces acting in opposite directions. If they act at **different heights**, they form a couple, and that couple rolls the aircraft. In the air the pilot simply trims it out and it is felt only as a one-wheel-low or one-skid-low hover attitude; on touchdown it shows up as a distinct lateral tilt as the aircraft settles, and on a slope or a moving deck it is genuinely unwelcome. The design answer is to arrange the tail rotor so that its **thrust line lies in line with the main rotor hub**, which removes the vertical separation and therefore removes the rolling couple altogether. This is why tail rotors on many types sit high on a raked fin or on a pylon rather than low on the boom, and why an unapproved change to fin or pylon geometry is never a minor modification.

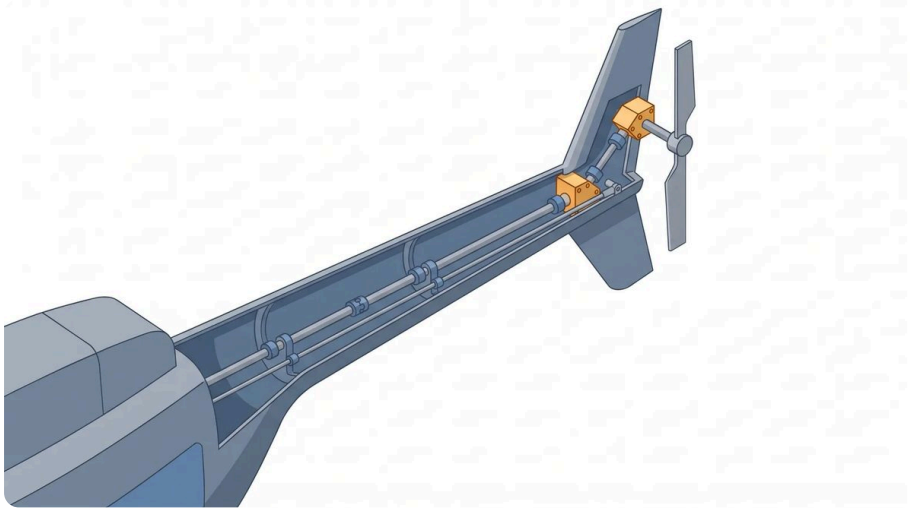
Tail Rotor Rigging and Aerodynamic Behaviour

- **Neutral pedals give positive pitch.** The tail rotor is rigged so that with the pedals central the blades already stand at a positive pitch angle and are producing anti-torque thrust, because in normal powered flight there is always torque to oppose. A tail rotor rigged to zero or negative pitch at neutral pedal would leave the torque reaction uncountered and the aircraft would yaw as soon as power was applied. The pedal range then runs from a small negative pitch — needed for autorotation and for yawing against the torque — through neutral to a large positive pitch for maximum-power conditions.

- **The tail rotor experiences dissymmetry of lift too.** It is a rotor working in a crossflow whenever the aircraft has sideways or forward speed, so its advancing blade sees a higher relative airflow than its retreating blade exactly as the main rotor's does. Tail rotors are therefore built with flapping freedom — usually a teetering (see-saw) head — and are commonly given a **delta-three hinge**, an inclined flapping hinge that mechanically reduces blade pitch as the blade flaps up. Because reducing pitch reduces angle of attack and therefore lift, this coupling damps the flapping response and keeps flapping angles small in a very confined space.
- **The pedal that increases thrust costs power.** Anti-torque thrust is bought from the same engine as rotor thrust. Whichever pedal calls for *more* tail rotor pitch takes a larger slice of available power away from the main rotor, so unless collective or throttle is added the helicopter will begin to sink; the opposite pedal releases power back to the main rotor and the aircraft tends to climb. On a governed engine the response is automatic, but on a piston type with a manually managed throttle it is a handling technique the pilot must apply deliberately.
- **Loss of tail rotor effectiveness.** At low airspeed and high power, particular combinations of wind direction can put the tail rotor into disturbed air — the main rotor's own vortices, its own wake, or a weathercocking flow that pushes the nose round — so that the yaw it produces no longer matches the pedal. Recovery is flown by regaining airspeed to get the fin and tail rotor into clean air. For the technician the point is that a genuine loss of effectiveness has an aerodynamic cause, whereas an uncommanded yaw at ordinary power settings is a rigging, control-run or drive fault to be investigated.
- **Forward speed unloads the tail rotor.** As airspeed builds, the vertical fin generates its own side force, so less tail rotor thrust is needed to hold heading and the pedal moves toward the neutral end of its range. Many fins are deliberately cambered or offset so that at cruise speed the fin alone carries most of the anti-torque duty.

The Three Systems Compared in Service

Beyond how each system generates its side force, the practical differences an engineer meets are in inspection burden and in where the failure modes lie. A **conventional tail rotor** puts an exposed, high-inertia rotor at the end of a long, slender drive train: hanger bearings, flexible couplings, an intermediate and a tail gearbox with their own oil levels and chip detectors, and a control run that may be cable, push-rod or a mixture. It must be tracked and balanced, and a tail rotor out of balance announces itself as a characteristic high-frequency vibration felt through the pedals rather than through the airframe.



A conventional tail rotor ends a long, slender drive train of shafts, bearings and two gearboxes.

A **Fenestron** trades that exposed rotor for a many-bladed fan of small diameter turning faster inside a duct, so the blades are shorter, lighter and shielded. The duct itself is aerodynamically active: the suction generated around the intake lip contributes a real share of the total side force, which is why duct-lip contour, surface condition and the very small blade tip clearance are inspection items and why erosion or a dented lip is more than cosmetic. The blades are commonly spaced unevenly around the hub, which spreads the acoustic energy across several frequencies instead of concentrating it in one tone — this is a noise measure, and it means a blade set must be installed in its designated positions.

A **NOTAR** installation removes the exposed rotor entirely, at the cost of depending on the main rotor's downwash flowing over the boom to work its circulation control. That dependence has a direction worth knowing: the boom slots are most effective in the hover and at low speed, where the downwash over the boom is strongest, while at higher forward speeds the downwash is swept aft and the vertical stabilisers plus the variable direct jet thruster take over the yaw duty. Maintenance attention shifts accordingly — from a tail rotor track and balance to fan condition and balance, slot cleanliness and profile, boom surface condition, and the rigging of the thruster's rotating nozzle.

Gyroscopic Precession and Phase Lag

The spinning rotor acts as a **gyroscope**. A fundamental property of a gyroscope is **precession**: when a force is applied to a spinning disc, the resulting displacement occurs **90° later in the direction of rotation**.

This has a critical consequence for helicopter control. If the pilot wants to tilt the rotor disc forward (to fly forward), the control system must change the blade pitch at a point **90° before** the desired disc tilt. For a counter-clockwise rotating rotor (viewed from above), to tilt the disc forward, the blade pitch must be increased at ψ = 270° (retreating side) and decreased at ψ = 090° (advancing side). The resulting flapping response occurs 90° later — maximum flap-up at ψ = 180° (over the tail), tilting the disc forward as desired.

Phase Lag Rule: Input 90° before → Response at desired point.
To tilt the disc **forward**: increase pitch at **retreating side** (ψ = 270°), decrease at **advancing side** (ψ = 090°).
The swashplate mechanism handles this 90° offset automatically — the pilot simply pushes the cyclic forward.

Two Ways of Explaining the Same 90 Degrees

The gyroscopic account above is the one to quote in an examination, but it is worth knowing the aerodynamic account as well, because it explains why the lag is close to but not always exactly 90°. A flapping blade is a spring-mass system: it is pulled back toward the plane of rotation by centrifugal force, which acts like a very stiff spring, and its natural frequency in flap works out at almost exactly **one cycle per revolution**. A cyclic pitch input forces that system at precisely its own natural frequency, and any lightly damped system driven at resonance responds a quarter of a cycle behind the force that drives it. A quarter of a revolution is 90°. The gyroscopic explanation and the resonance explanation are two descriptions of the same physical result, and both predict the same answer: **after a pitch change, the blade reaches its maximum flapping displacement roughly 90° further round the disc, measured in the direction of rotation.**

Notice what is being lagged. It is not the lift that lags — the lift responds to the pitch change immediately. What lags is the *displacement*. The blade begins to accelerate upward the instant its angle of attack increases, keeps accelerating for as long as the extra lift is present, and therefore reaches its highest point when the extra lift has fallen back to zero, which happens a quarter of a revolution later. Thinking of it as "force now, maximum movement a quarter turn later" gets the direction right every time.

Worked Case: Flying Backwards

Take the mirror image of the forward-flight case above, on the same anti-clockwise rotor viewed from above, and reason it through step by step rather than trying to memorise a result.

Requirement: move the helicopter rearward, so the tip path plane must tilt **rearward** — blades highest over the nose end of their travel and lowest over the tail.
Step 1 — where must maximum flap-up occur? For the disc to tilt back, the blade must reach its highest point **over the nose** (ψ = 000°).
Step 2 — where must the pitch input go? 90° earlier, measured in the direction the rotor is turning. With anti-clockwise rotation viewed from above, the blade passes over the **starboard side** immediately before it reaches the nose.
Step 3 — answer. The pitch increase is fed in on the starboard side, that is, to the **right of the longitudinal axis** (ψ = 090°), with the matching pitch decrease diametrically opposite on the retreating side.
Check: this is the exact opposite of the forward-flight case, where the pitch increase goes in on the retreating side — which is what it should be, since the two commands tilt the disc in opposite directions.

The same three steps solve any variant of this question. Decide which way the disc must tilt, decide where the blade must be at its highest point to produce that tilt, then count back 90° against the direction of rotation to find where the pitch goes in. The only piece of aircraft-specific information you need is which way the rotor turns.

How the Swashplate Hides the Offset from the Pilot

If the pilot had to apply the cyclic 90° away from the direction he wanted to go, helicopters would be unflyable. The offset is therefore built into the mechanism. The swashplate assembly has a **non-rotating (stationary) star**, moved directly by the cyclic and collective control runs, and a **rotating star** which turns with the mast on a bearing between the two and carries the pitch links up to the blade pitch horns. Tilting the non-rotating star tilts the rotating star with it, and each pitch link then rises and falls once per revolution, giving each blade a sinusoidal pitch variation whose maximum occurs at a fixed azimuth.

The 90° is introduced geometrically, by where the pitch horn sits relative to the blade and where the pitch link attaches to the rotating star. That built-in angular offset is called the **control phase angle** or advance angle. Designers may split it between pitch horn offset and a mechanical mixing unit, but the result is always the same: the swashplate is commanded in one direction and the disc tilts in the direction the pilot actually asked for, so cyclic forward gives disc forward and cyclic left gives disc left.

⚠ **Rigging Trap:** because the phase angle is geometric, an error in reassembly changes it. Pitch links fitted to the wrong pitch horns, a swashplate driver or scissor assembly indexed one position out, or a control mixing unit rigged without the correct rig pins, will all shift the phase angle away from its design value. The symptom is **control cross-coupling**: a purely fore-and-aft cyclic input produces a roll as well as a pitch response, or a lateral input produces pitch. Cross-coupling that appears after a rotor head or control-run disturbance is a rigging fault to be traced through the phase angle, not a handling characteristic to be trimmed out.

When the Lag Is Less Than 90 Degrees

The clean 90° result assumes the blade is free to flap about a hinge lying on the axis of rotation. Real heads differ, and the difference has a consistent direction.

- **Teetering (see-saw) heads** have effectively zero flapping hinge offset, since the pair of blades pivots about a single point on the mast axis. The phase lag is essentially the full 90°, and the head can transmit almost no moment to the mast — which is why these types are sensitive to low-g manoeuvres, where the rotor is lightly loaded and the tail rotor's side force is no longer opposed by rotor thrust.
- **Articulated heads with offset flapping hinges** place the hinge a short distance out from the axis. Centrifugal force acting on the flapped blade now has a moment arm about the mast, so the head itself

transmits a moment to the fuselage — that is what gives these types crisp control response. The flapping response builds slightly earlier in azimuth, so the phase lag becomes **somewhat less than 90°**.

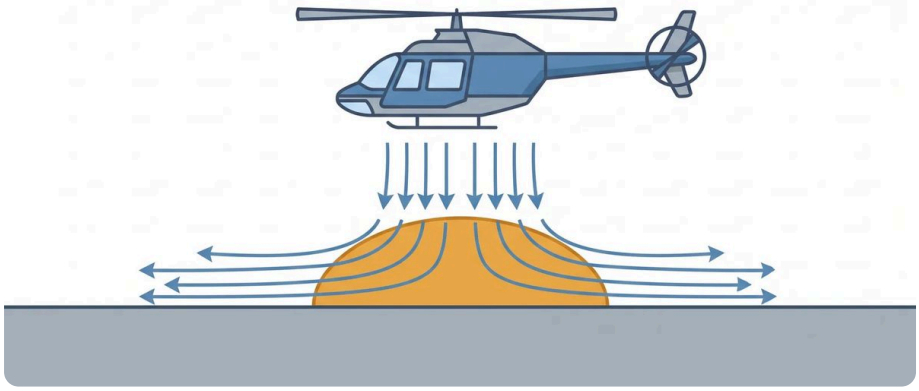
- **Hingeless and bearingless heads** replace hinges with a flexible root that behaves like a large effective hinge offset. Control power is higher again and the phase lag is noticeably less than 90°. The designer compensates by setting the control phase angle to match the actual lag of that head — another reason the phase angle is a design figure to be rigged to the manual, never a number to be assumed.

One further consequence of the lag is worth carrying into the sections that follow: *any* once-per-revolution disturbance of blade lift — not only a deliberate cyclic input — produces its maximum flapping displacement 90° later. The dissymmetry of lift produced by forward speed, the uneven inflow across the disc at low speed, and a single mis-tracked blade all obey the same rule, which is why a rotor fault so often shows itself at an azimuth a quarter of a turn away from where it originates.

Ground Effect and Translational Lift

Ground Effect

When a helicopter hovers close to the ground (within approximately **one rotor diameter** of height), the rotor downwash cannot fully develop and is deflected outward by the ground surface. This reduces the induced velocity at the rotor disc, which reduces **induced drag** — the largest component of drag in a hover. The result is that less power is required to hover *in ground effect* (IGE) than *out of ground effect* (OGE). Typically, hovering IGE requires **10–15% less power** than hovering OGE. Ground effect diminishes rapidly above one rotor diameter height and is negligible above 1.5 rotor diameters.



Close to the ground the downwash is turned outward and a cushion of raised pressure builds under the rotor.

Translational Lift

Translational lift is the increase in rotor efficiency that occurs as the helicopter transitions from hover to forward flight (typically becoming noticeable at **15–24 knots**). In a hover, the rotor operates in its own downwash — the air is recirculated and disturbed. As the helicopter moves forward, the rotor encounters undisturbed (clean) air, which is more efficient at producing lift. Additionally, the induced velocity at the rotor disc decreases with increasing forward speed (the disc is sweeping through a larger volume of air per second), reducing induced drag. The pilot feels translational lift as a noticeable "bump" or climb tendency during the transition.

Ground Effect at the Blade: Pressure Below, Angle of Attack Above

It is worth following ground effect all the way down to the blade section, because the examination asks about it from two different directions and both answers must be available. Seen from underneath, the descending column of air cannot escape freely when the surface is close, so it spreads outward and stagnates, and a region of **raised static pressure — the ground cushion — builds up beneath the rotor disc** and helps support the aircraft. Seen at the blade, the restricted flow means the air is not being drawn down through the disc as quickly, so the **induced velocity falls**. A smaller induced velocity means a smaller inflow angle, and since $\alpha = \theta - \phi$, a smaller inflow angle leaves a **larger angle of attack** at the same collective setting. The blade therefore produces **more lift for the same pitch**, or equivalently the same lift for less power. Both statements describe one effect: the ground increases the lift available.

There is a second, smaller contribution. The reduced induced velocity also tilts the total reaction of each section further forward, so the in-plane component holding the blade back is smaller and the rotor takes less torque to turn at the same thrust. That is the induced-drag reduction described above, and it is the part the pilot notices as a lower torque indication when hovering close to the surface.

What Weakens or Destroys Ground Effect

- **Height.** The effect is strongest with the skids or wheels just clear of the surface and weakens continuously as height increases, becoming negligible beyond about one and a half rotor diameters. This is why flight manuals publish two separate hover ceilings, in and out of ground effect, and why a load that can be lifted from a hard pad may be beyond the aircraft over an obstruction.
- **Surface texture.** Long grass, standing crops, brush, deep snow and broken or rocky ground absorb and disperse the downwash instead of turning it cleanly outward, so the pressure cushion that builds up is weaker than over a hard, smooth surface. Water is a special case: the surface deforms and the downwash digs a depression, so ground effect over water is reduced compared with a solid surface.
- **Slope and obstructions.** A sloping surface, a lip, a wall or a trench alongside the aircraft lets the deflected downwash escape in one direction more easily than another, so the cushion becomes asymmetric and the aircraft may need cyclic to stay level.

- **Wind.** Any appreciable wind, whether from the aircraft's own translation or from the weather, carries the cushion away downwind before it can build. Ground effect is fundamentally a *still-air, low-speed* benefit.

Recirculation — the Trap That Looks Like Loss of Ground Effect

A helicopter holding a low hover in a strong crosswind loses lift, and the reason is examined precisely because the obvious answer is the wrong one. The aircraft is still only a couple of metres up, so it has **not** climbed out of ground effect — the cushion is still there. What the crosswind does is carry the outflowing downwash sideways along the surface and then blow it back **up into the rotor disc** on the upwind side. The blades then meet air that is already moving downwards before it reaches them. That extra downward velocity increases the inflow angle at the affected sections, which **reduces their angle of attack and reduces their lift** at unchanged collective. The loss is caused by **recirculation**, not by the removal of the ground cushion.



In confined air the rotor's downwash returns into the disc: lift is lost to recirculation, not lost cushion.

Recognising the mechanism matters operationally as well as in the examination, because the same recirculation appears whenever a rotor is worked in confined air: hovering close to a hangar wall or a line of trees, operating on a small helideck, or hovering in a hollow. It also explains the classic "hovering in a hole" experience in a confined-area landing, where the aircraft needs noticeably more power than the same hover would take in the open.

Translational Lift: Why Induced Velocity Falls with Speed

Two things happen as the aircraft accelerates, and they are worth separating. The first is the escape from the rotor's own wake, described above: below roughly 15 knots the rotor is still flying through the disturbed, descending air it has just processed, and above that speed it has outrun the recirculating vortices and is meeting undisturbed air. The second is a continuous change in the mass flow through the disc. In the hover the rotor must accelerate a relatively small mass of air to a high velocity to generate its thrust; in forward flight

it sweeps through a much larger volume of air every second, so it obtains the same thrust by giving a **much smaller velocity change to a much greater mass**.

Induced velocity at speed (forward flight approximation):

$$v_i \approx \frac{T}{2\rho A V}$$

V = forward speed. Induced velocity is therefore **inversely proportional to forward speed** once the aircraft is well clear of the hover, so as V rises, v_i falls, the inflow angle falls, and the **induced power $P_i = T v_i$ falls with it**. This is the mathematics behind the descending left-hand branch of the power-required curve.

Because a falling induced velocity means a falling inflow angle, the angle of attack of every blade section **rises** at unchanged collective — which is the aerodynamic origin of the climb tendency the pilot feels as the aircraft accelerates through the transition. Left alone, the helicopter climbs; the correct response is to lower the collective slightly, and the aircraft then accelerates on less power than it needed to hover.

☁ **Two power changes, one transition — get the order right:** leaving the hover for forward flight, the pilot eases the cyclic forward. Tilting the disc diverts part of the rotor thrust into propulsion, so the **vertical component supporting the weight is reduced** and the aircraft would sink; collective and therefore **engine power must be increased** to hold height while it accelerates. A few seconds later, passing roughly 15 knots, effective translational lift arrives and the rotor becomes markedly more efficient, so **power can now be reduced** for the same lift. Both statements are correct — they belong to different moments in the same manoeuvre. The examination asks about the first when it describes moving from the hover into forward flight, and about the second when it specifies acceleration above about 15 knots.

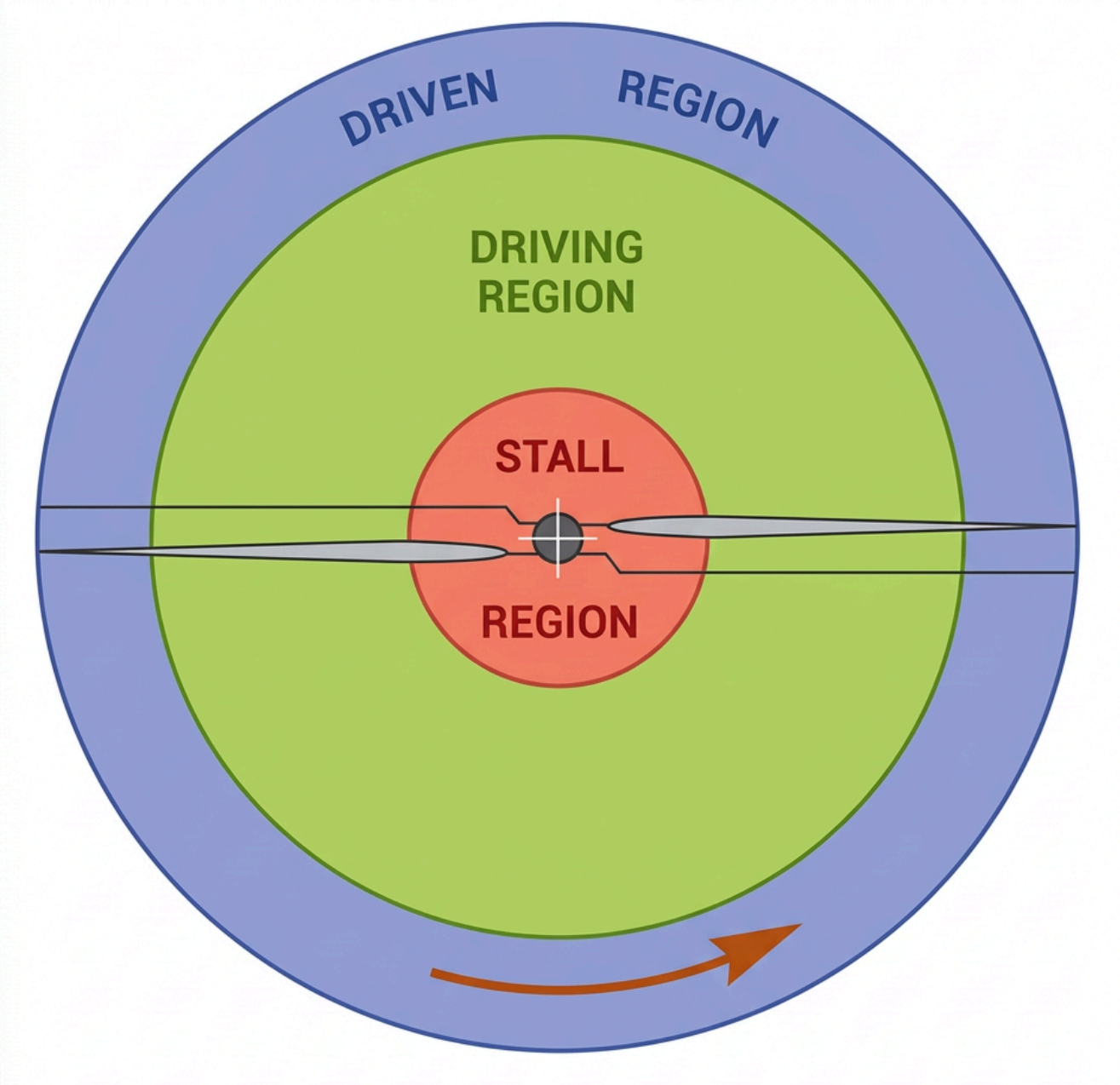
Two Effects That Accompany the Transition

Transverse flow effect. At low forward speed the disc is only partly clear of its own downwash: air entering over the front of the disc has not yet been accelerated downwards, while air leaving over the rear has passed through the disc and carries a substantial induced velocity. The front and rear of the disc therefore work at different inflow angles and different angles of attack, so lift is unevenly distributed fore and aft. Because the flapping response to any once-per-revolution lift variation appears 90° later in the direction of rotation, this fore-and-aft imbalance shows up as a *lateral* disc tilt, felt as a distinct vibration and a roll tendency in the speed band just below effective translational lift. Which way the roll acts follows from the direction of rotor rotation for the type.

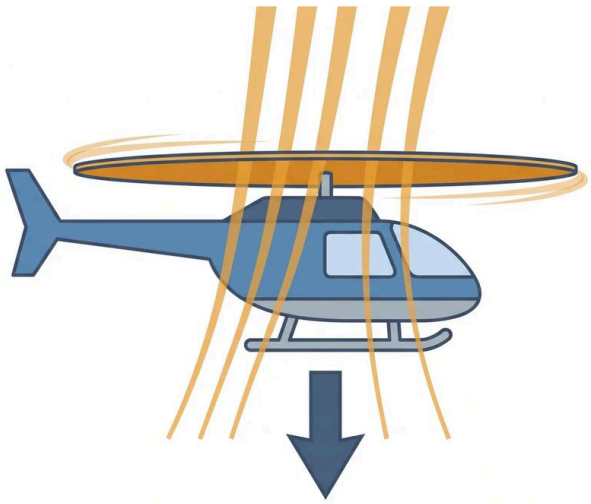
Tail rotor efficiency. The tail rotor gains from the cleaner airflow at the same time the main rotor does, and the vertical fin begins to contribute its own side force. The pedal position needed to hold heading therefore changes noticeably as the aircraft passes through the transition, and the aircraft will yaw if the pilot does not follow it. It is the same clean-air mechanism, applied to a smaller rotor at the other end of the aircraft.

↓ Autorotation

Autorotation is the condition of flight where the rotor is driven entirely by aerodynamic forces resulting from the helicopter descending through the air — no engine power is used. It is the helicopter equivalent of a glide in a fixed-wing aircraft and is the primary emergency procedure after an engine failure.



During autorotation, the helicopter descends and air flows *upward* through the rotor disc. This upward airflow changes the relative airflow at each blade section, tilting the total aerodynamic reaction force forward (in the direction of rotation). The forward component of this force **drives** the rotor, maintaining RPM without engine power. The rotor blade can be divided into three regions:



In autorotation the helicopter descends and air flows upward through the rotor disc, keeping it turning.

Region	Location	Role in Autorotation
Driven (stalled) region	Inner ~25% of blade (near root)	The angle of attack is very high (often stalled). This region produces **drag** that tends to slow the rotor. It is "driven" by the autorotative region.
Driving (autorotative) region	~25–70% of blade (mid-span)	The total aerodynamic force tilts forward of the axis of rotation. The forward component **exceeds drag**, providing a net driving force that sustains rotor RPM. This is the engine of autorotation.
Propeller (driven) region	Outer ~30% of blade (near tip)	The angle of attack is relatively low. This region produces lift but the total force tilts slightly behind the rotation axis, producing a small **drag component**. It produces most of the overall lift.

Aviation Context: Autorotation is practised regularly during helicopter pilot training. For maintenance technicians, understanding autorotation is critical because: (1) it explains why rotor RPM must be maintained within limits — too low and the rotor stalls, losing autorotative capability; (2) it affects blade pitch rigging tolerances; (3) the freewheel unit (sprag clutch) in the transmission must disengage instantly for autorotation to work — testing this is a key maintenance task.

Vortex Ring State, Blade Stall and Compressibility

Vortex Ring State (VRS) / Settling with Power

Vortex ring state occurs when the helicopter descends into its own downwash at a rate of descent greater than about **300 ft/min** at low forward speed (below ~30 knots translational lift speed), typically with power applied. The rotor downwash is recirculated back up through the rotor disc, creating a doughnut-shaped vortex ring around the disc. The rotor becomes inefficient — increasing collective (more power) may not arrest the descent because the additional downwash is immediately recirculated. Symptoms: high vibration,

uncommanded yaw, loss of collective effectiveness, rapidly increasing rate of descent. Recovery: lower collective, increase forward speed (or autorotate) to fly out of the recirculating air mass. Power settling, better described as settling with insufficient power, is a different condition: a descent that the engines cannot arrest within their power limits, although vortex ring state may develop from it.

Retreating Blade Stall

As forward speed increases, the retreating blade must increase its angle of attack to compensate for lower airspeed (to equalise lift). At a critical forward speed, the retreating blade's angle of attack exceeds the **stall angle**, and the blade stalls. This begins at the blade tip (where the velocity deficit is greatest) and progresses inboard. Symptoms: vibration, nose pitch-up, and roll toward the retreating side. Retreating blade stall sets the **maximum forward speed (VNE)** of the helicopter.

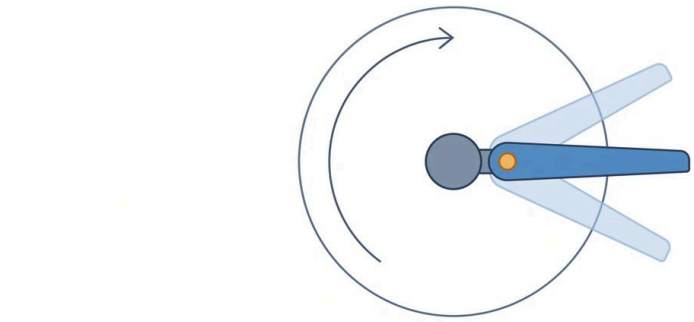
Compressibility Effects

At high forward speeds, the **advancing blade tip** can approach or exceed the speed of sound (Mach 1). At this point, shock waves form on the blade surface, causing a sudden increase in drag, loss of lift, and vibration. Combined with retreating blade stall on the other side, compressibility limits the helicopter's maximum speed. This is why conventional helicopters are limited to approximately **170–200 knots** — the advancing blade tip is near Mach 0.9 and the retreating blade is near stall simultaneously.

🔁 Coriolis Effect — Blade Lead-Lag

The **Coriolis effect** in rotary wing flight causes blades to accelerate or decelerate in the plane of rotation (lead or lag) as they flap up and down. This is a direct application of the conservation of angular momentum.

When a blade flaps **upward**, its centre of mass moves closer to the axis of rotation (the effective radius decreases). To conserve angular momentum ($L = I\omega$), the blade speeds up (leads ahead). When a blade flaps **downward**, the centre of mass moves further from the axis, and the blade slows down (lags behind). This lead-lag motion occurs once per revolution and must be accommodated by the rotor head design — typically via **lead-lag hinges** (drag hinges) on fully articulated heads, or by flexible elements on hingeless/bearingless heads. **Lead-lag dampers** are fitted to prevent excessive oscillation and the dangerous condition known as **ground resonance**.



Lead-lag: each blade swings a little ahead of and behind its normal position, within the plane of rotation.

The Three Blade Motions — Do Not Confuse Them

A rotor blade on a fully articulated head is free to move in three independent ways, each about its own axis and each provided for by its own hinge or bearing. Examination questions routinely offer one motion as the distractor for another, so it is worth fixing the geometry precisely: the hinge is named after the direction of its *axis*, while the motion it permits is at right angles to that axis.

Motion	About which axis	Blade moves	Driven by
Flapping	Horizontal hinge, also called the lateral or flapping hinge	**Vertically**, up and down relative to the plane of rotation	Aerodynamic: any change in blade lift, whether from dissymmetry, cyclic input or a gust
Dragging (lead-lag)	Vertical hinge, also called the drag or lead-lag hinge	**Horizontally**, fore and aft in the plane of rotation	Coriolis effect from flapping, plus the once-per-revolution variation of blade drag in forward flight
Feathering	Spanwise (pitch-change) axis, through the feathering bearing	**Rotates** about its own length, changing pitch angle	The pilot, through collective and cyclic acting on the swashplate and pitch links

Say it once in the form the question asks for: **blade dragging is the horizontal movement of each blade about its vertical hinge.** Vertical movement about a horizontal hinge is flapping, and it is a different motion entirely. Note also that the motions are coupled, not independent in service: a blade that flaps also changes its angle of attack and, through the resulting change in lift, its drag — which is why a flapping blade experiences a change in both angle of attack and drag force even though the pilot has commanded no pitch change at all.

Putting a Number on the Coriolis Effect

Conservation of angular momentum is easy to state and easy to under-appreciate, so it helps to see the size of the effect. A blade's moment of inertia about the rotor axis depends on the *square* of the radius at which its mass sits, $I = \sum mr^2$, and the product $I\Omega$ is what must be conserved when no in-plane force is applied. Because of that square, a modest change in the radius of the blade's centre of mass produces a doubled proportional change in inertia.

Worked example — why the hinge is needed: a blade flaps up far enough that the radius of its centre of mass falls by 2%.
Moment of inertia falls by roughly $2 \times 2\% = 4\%$, because I varies with r^2.
To keep $I\Omega$ constant, the blade's angular velocity must therefore **rise** by about 4% — the blade tries to **lead**, running ahead of its normal position.
Flap the blade back down and the radius, and the inertia, are restored; the blade must slow again and it **lags**.
Now recognise that in forward flight this happens **once every revolution, on every blade, continuously**. A rotor at normal operating speed turns several times per second, so a blade attached rigidly in the plane of rotation would be asked to absorb that acceleration and deceleration as a fully reversed bending load at that frequency — a fatigue problem no reasonable root structure could carry for long. That, not comfort, is why the drag hinge exists.

The associated in-plane force follows the standard Coriolis form, $F = 2m\Omega v_r$, where v_r is the rate at which the mass is moving radially inward or outward. Two directions follow from it. The force is **larger at higher rotor RPM**, and it is **larger when the blade flaps quickly** — so the most demanding case is a rapid, large flapping excursion at full rotor speed, which is exactly the condition produced by an abrupt cyclic input or a strong gust rather than by steady cruising flight.

Head Designs: Hinge, Flexure, or Design the Effect Out

- **Fully articulated heads** give each blade its own drag hinge and its own damper. This is the general solution and it is used wherever there are three or more blades.
- **Hingeless and bearingless heads** replace the hinge with a flexible root section that bends in-plane. The motion still occurs, it is simply accommodated elastically, and damping is provided by elastomeric elements built into the root region rather than by a discrete damper unit.
- **Underslung teetering heads** take a different approach: rather than accommodating the Coriolis force, they largely avoid generating it. The teetering hinge is placed *above* the blades' centre of mass so that, as the see-saw tilts, each blade's centre of mass swings very nearly along an arc that keeps its distance from the axis of rotation constant. If the radius of the centre of mass barely changes, the moment of inertia barely changes, and there is almost no lead-lag tendency to accommodate. This is the reason two-blade underslung rotors can be built with no drag hinges and no lead-lag dampers at all — the geometry has removed the problem rather than absorbing it.

Ground Resonance and Why the Dampers Are Airworthiness Items

Ground resonance is a **mechanical** instability, not an aerodynamic one, and it is the reason lead-lag damping is treated so seriously. In normal running the blades are evenly spaced and the rotor's centre of gravity sits on the axis. If something displaces one blade in lead-lag — a heavy or one-wheel-first touchdown, taxiing over rough ground, or a shock through the undercarriage — the blades become unevenly spaced and the rotor's centre of gravity moves off the axis. A rotating out-of-balance force is then applied to the airframe. If the frequency of that force is close to a natural frequency of the fuselage bouncing on its undercarriage, and the aircraft is **in contact with the ground** so that the undercarriage can feed energy back into the rotor, the oscillation reinforces itself and can build to destructive amplitude within a very few seconds.

Two things break the loop, and both belong to maintenance as much as to flying. The first is **lead-lag damping**: dampers absorb the energy of blade in-plane motion so the disturbance decays instead of building. The second is the **landing gear**: correct oleo servicing and correct tyre pressures set the airframe's natural frequency and provide their own damping, so an under-serviced strut or a soft tyre changes the very frequencies the design relied on. A serviceable damper on an incorrectly serviced undercarriage does not give a safe aircraft, and the flight manual's ground-resonance limitations assume both. The pilot's escape follows the same logic: if rotor RPM is within its normal range the fastest cure is to **lift off**, because leaving the ground removes the ground reaction path that feeds the oscillation; if rotor RPM is too low for that, the throttle is closed and the collective lowered to stop the rotor and take the energy out of it.

Hingeless and bearingless heads can suffer a related instability **in flight**, known as air resonance, where the in-plane blade motion couples with the aircraft's rigid-body roll and pitch modes instead of with the undercarriage. It is dealt with by the same means — adequate in-plane damping designed into the root or the damper elements — which is why damper condition, freedom of blade lead-lag movement and correct damper matching within a set are inspection items, and why a leaking hydraulic damper or a cracked elastomeric element is never deferred as a cosmetic defect.

Flight Regimes

Regime	Rotor Disc	Key Aerodynamic Points
Hovering	Horizontal, producing vertical thrust equal to weight.	Symmetrical lift distribution. Maximum induced power required. Ground effect reduces power needed. Tail rotor must counteract full torque.
Forward flight	Tilted forward (via cyclic). Thrust vector has forward and vertical components.	Dissymmetry of lift corrected by flapping. Translational lift improves efficiency. Induced power decreases. Parasite drag increases. Total power has a U-shaped curve — minimum at ~60–80 kts.
Climbing	Tilted forward with increased collective (more pitch angle on all blades).	Requires more power than level flight. Excess power (above that needed for level flight) provides climb capability. Rate of climb = excess power ÷ weight.

Regime	Rotor Disc	Key Aerodynamic Points
Descending	Reduced collective. Airflow has upward component through disc.	Reduced power required. Risk of vortex ring state at low speed/high descent rate. Autorotation is the extreme case (zero power).
Turning flight	Disc tilted in the direction of the turn (via cyclic).	More collective (and therefore more power) required because the rotor must produce both vertical and horizontal components of thrust. Load factor increases in turns ($n = 1/\cos\phi$). Steeper turns increase power demand significantly.

The Power-Required Curve Behind the Table

The U-shaped total-power curve referred to in the forward-flight row is the sum of three separate demands, each with its own behaviour against airspeed. Knowing which component dominates where turns the curve from something to be memorised into something that can be reasoned out.

Component	What it pays for	Behaviour as airspeed increases
Induced power	Accelerating air downwards through the disc to make thrust.	**Falls** steeply, because induced velocity is inversely proportional to forward speed. Largest single item in the hover; nearly insignificant at cruise speed.
Profile power	Dragging the blades themselves round through the air.	Roughly constant through the low and middle speed range, then **rises** as the advancing blade meets higher dynamic pressure and the retreating blade needs greater pitch.
Parasite power	Pushing the fuselage, rotor head, undercarriage and any external stores through the air.	Negligible in the hover, then **rises very rapidly** — drag with the square of speed and the power to overcome it faster still. It dominates the high-speed end.

Add the three and the minimum of the total sits in the middle of the speed range, at the speed the retained table quotes. Two performance speeds fall straight out of the curve. The speed for **minimum power** gives the greatest excess power over what level flight demands, so it is the speed for maximum rate of climb and for maximum endurance. The speed for the best ratio of speed to power — found where a line from the origin just touches the curve — gives the best specific air range and the best glide-equivalent distance in autorotation. Anything bolted to the outside of the aircraft moves the parasite branch upward and both of those speeds change with it, which is why performance data are quoted per configuration.

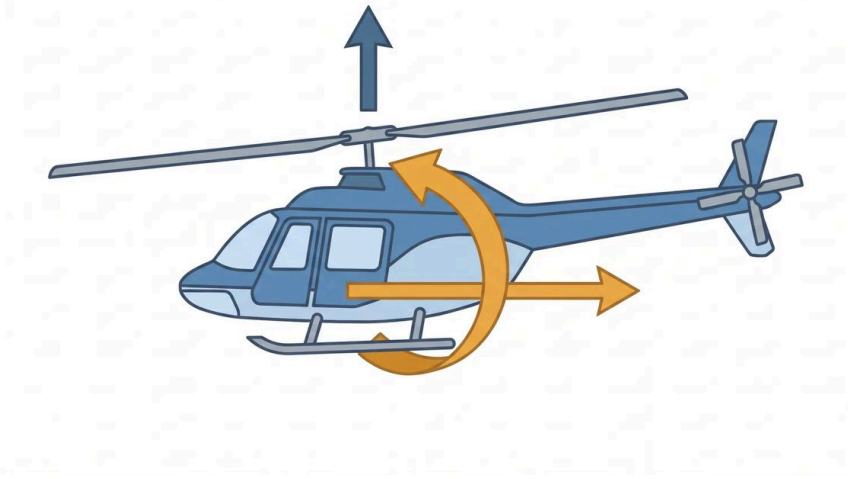
Hover to Forward Flight: What the Cyclic Actually Does

Horizontal flight in any direction is produced by **tilting the rotor disc in the direction of travel** with the cyclic, so that the single rotor thrust vector gains a component along the intended path. The collective does not do this; raising it changes the size of the thrust vector, not its direction. That distinction is the definition of translational flight and it is examined directly.

Follow the vector through and the immediate consequence appears. Rotor thrust is one force of fixed magnitude at any instant. Tilt it forward and it now has a forward component that did not exist before, so the **vertical component supporting the weight must be smaller** than it was. Easing the cyclic forward from a hover therefore **reduces vertical lift**, and the aircraft begins to descend as it accelerates unless the collective is raised to restore the vertical component. In a hover close to the ground that is felt immediately; from a high hover it is felt as an initial sink.

Forward Flight: Fuselage Drag and Attitude

Parasite drag acts on the fuselage at or below the centre of gravity, while the rotor thrust that opposes it acts at the rotor head well above. The two form a couple, and the direction of that couple is **nose-down**. Because parasite drag grows with the square of airspeed, the couple strengthens as speed increases, so the helicopter's natural attitude becomes progressively more nose-down the faster it flies — a tendency the pilot balances with aft cyclic and which designers reduce with a horizontal stabiliser producing a download at the tail. Test the direction on a worked case: at rest there is no parasite drag and no couple; at high cruise the drag force is large and acts low, so the tail rises and the nose drops. The tendency exists only with forward speed and increases with it, not with a reduction in speed.



Drag acting low on the fuselage and thrust at the rotor head form a couple that pitches the nose down.

Descent, Powered and Unpowered

In a **power-on descent** the engine is still driving the rotor, so all four of the classic forces are present at once: weight always acts, the rotor produces lift, thrust is present because the rotor is being driven and the disc is tilted, and drag opposes the motion through the air. It is worth stating plainly because the examination contrasts it with the power-off case, where the engine contributes nothing and the descent is sustained by the airflow alone.

In autorotation the rotor is free-wheeling and lift continues to be produced by **freely rotating blades**, driven by the air coming up through the disc rather than by the engine. Directional control is retained throughout — cyclic, collective and pedals all still work, since none of them depends on engine power — and the manoeuvre is the response to a power loss, not a cause of one. Two behaviours matter to the engineer:



In the final cushioning pull, the energy comes from the rotational inertia stored in the spinning blades.

- **Rotor RPM in a steady autorotation is about the same as in powered flight.** Lowering the collective reduces blade pitch and therefore blade drag, while the upflow through the descending disc drives the rotor; both act in the same direction and the rotor tends to accelerate. The pilot regulates it with the collective — raising the lever increases pitch and drag and **slows** the rotor, lowering it lets the rotor **speed up** — and holds it within the flight manual power-off range, normally in the green arc, never letting it build too high or decay.
- **The flare converts speed into rotor energy.** Raising the nose near the ground increases the upflow through the disc, which accelerates the rotor and simultaneously increases rotor thrust, reducing both the rate of descent and the groundspeed before the collective is finally raised to cushion the touchdown. The energy for that final pull comes from the rotational inertia stored in the blades, which is why blade mass appears in the rotor RPM design decision described earlier in this note.

The **height-velocity diagram** in the flight manual maps where a successful power-off landing is achievable. Its avoid regions are the low-height, low-speed corner — too little height to trade for rotor RPM and too little speed to flare with — and a high-hover region where the descent would build up before translational airflow could establish. Both are consequences of the same energy accounting, not arbitrary limits.

Turning Flight and the Load Factor

In a balanced level turn the rotor must produce a vertical component equal to weight *and* a horizontal component to provide the centripetal force, so total rotor thrust must exceed weight by the load factor $n =$

$1/\cos\phi$. The arithmetic is worth doing once because the numbers are not intuitive.

Worked example — what a bank angle costs:
At $\phi = 30°$: $n = 1/\cos 30° = 1.15$ — the rotor must produce 15% more thrust than in level flight.
At $\phi = 45°$: $n = 1.41$.
At $\phi = 60°$: $n = 1/0.5 = 2.0$ — the rotor must produce **twice** the thrust, and the aircraft and every blade root fitting carry twice their level-flight load.
Consequences that all follow from that extra thrust: more collective and therefore more power is required, so a turn steepened in the cruise can run into the power limit; the coning angle **increases** because lift has risen while rotor RPM has not; and the retreating blade needs a **higher** angle of attack to carry its share, which **reduces** the margin to retreating blade stall. A steep turn at high speed and high weight brings all three effects together, which is why manoeuvre limits tighten as speed and weight increase.

Why the Speed Limit Is Not a Single Number

Never-exceed speed is placarded as a function of density altitude and weight rather than as one figure, and both variations follow from the rotor limits described in the sections above. At **higher density altitude** the same indicated airspeed corresponds to a higher true airspeed, so the advancing blade tip approaches its compressibility limit sooner; at the same time the thinner air requires greater blade pitch for the same thrust, so the retreating blade is already closer to the stall. Both effects push the limiting speed **down**. At **higher weight** the second effect appears on its own: more thrust means more blade pitch, so the retreating blade stalls at a lower forward speed and the limit again comes **down**. An engineer signing off a placard change, a stabiliser modification or an external-stores installation is therefore working on a limitation with real aerodynamic content behind it, not an arbitrary number.

Module 13 — Aircraft Aerodynamics, Structures and Systems

13.1(a) — Aeroplane Aerodynamics and Flight Controls

This section covers the fundamental principles governing how fixed-wing aircraft generate lift, produce drag, and are controlled in flight. Although B2 avionics engineers work primarily with electronic systems, understanding aerodynamic principles is essential for correctly interpreting flight data, troubleshooting sensor inputs, and appreciating why avionics systems behave as they do.

The practical reason this matters is that almost every avionics box on a modern aeroplane is either measuring an aerodynamic quantity, computing one, or acting on one. An air data computer turns two pressures into altitude, airspeed and Mach number using an assumed atmospheric model. A stall warning computer compares a measured angle of attack against a threshold that itself moves with flap position. A yaw damper exists only because a swept-wing aeroplane has a particular unstable oscillation mode. When one of those systems misbehaves, the engineer who understands the underlying aerodynamics can tell the difference between a sensor fault, a rigging fault and an aircraft that is simply doing what physics requires.

The Atmosphere and International Standard Atmosphere (ISA)

The atmosphere is composed of approximately 78% nitrogen, 21% oxygen, and 1% other gases (argon, CO_2, water vapour). Its properties — pressure, temperature, and density — decrease with altitude, directly affecting aircraft performance and the readings of pressure-based instruments.

In the round figures the examiner tends to use, air is about **one-fifth oxygen and four-fifths nitrogen**. Those proportions stay essentially constant with height throughout the region in which aeroplanes operate, because the atmosphere is continually mixed by convection and wind. What changes with altitude is not the mixture but the *amount* of it in a given volume, which is why a cabin at 40 000 ft still contains 21% oxygen yet cannot sustain an unpressurised occupant: the partial pressure of that oxygen has collapsed along with the total pressure.

Structure of the Atmosphere

The lowest layer, the **troposphere**, contains almost all the water vapour and weather, and is the layer in which temperature falls steadily with height. Above it lies the **tropopause**, the boundary at which the temperature stops falling, and then the **stratosphere**, in which temperature is initially constant with height. The tropopause is not at a fixed altitude in the real atmosphere: it sits at roughly 55 000 ft over the equator and can be below 30 000 ft over the poles, so a transport aircraft in cruise may be in the troposphere on one sector and the stratosphere on the next. The ISA model fixes it at 36 089 ft as a compromise so that performance charts and instrument calibrations have one agreed reference.

ISA Sea-Level Values

- **Temperature:** 15 °C (288.15 K)
- **Pressure:** 1013.25 hPa (29.92 in Hg)
- **Density:** 1.225 kg/m³
- **Lapse rate:** –1.98 °C per 1 000 ft (up to 36 089 ft / tropopause)
- **Tropopause temperature:** –56.5 °C (constant above tropopause)
- **Reference latitude:** mean conditions at 45° North
- **Speed of sound at sea level:** approximately 340 m/s (about 661 kt)

The word **lapse rate** has a precise meaning that the exam tests directly: it is the rate at which *temperature* changes with altitude. Pressure and density also fall with height, but those gradients have their own names and are not called lapse rates. The ISA figure of –1.98 °C per 1 000 ft is usually rounded to 2 °C per 1 000 ft for mental arithmetic, and it applies only up to the tropopause — above that height the model holds temperature constant at –56.5 °C.

The ISA is a *model*, not a measurement of today's air. It was built from mean conditions measured at 45° North latitude, deliberately chosen as a mid-latitude average rather than equatorial or polar data, so that it represents the temperate regions where most aviation takes place. Because every altimeter, air data computer and performance chart in the world is calibrated to the same model, two aircraft flying the same indicated flight level are vertically separated even when neither is at the true geometric height the model implies.

How Pressure Falls With Height

Pressure at any level is simply the weight of the air column above that level, so it falls approximately exponentially rather than linearly. A useful rule the exam expects you to know is that **pressure halves for roughly every 18 000 ft of climb**: from 1013 hPa at sea level to about 500 hPa near 18 000 ft, and to about a quarter of sea-level pressure near 36 000 ft. This is also why the pressure change per 1 000 ft is roughly 1 hPa per 27 ft near sea level but far less per foot at altitude — the altimeter's capsule is working against a much weaker pressure gradient up high, which is one reason altimetry tolerances widen with altitude.

Density depends on both pressure and temperature through the gas law $\rho = P/(RT)$. Climbing reduces pressure, which reduces density; but a hot day reduces density at constant pressure. The two combine into **density altitude** — the ISA altitude at which the prevailing density would be found. High, hot and humid conditions all reduce density, and because lift, drag and engine thrust all depend on ρ, density altitude is the single number that predicts degraded take-off performance. Humidity lowers density because a water molecule is lighter than the nitrogen or oxygen molecule it displaces, which is the opposite of most people's intuition.

Aviation Context

Air data computers (ADCs) use ISA as the baseline model. They compute altitude, airspeed, and temperature deviations (ISA + / ISA –) by comparing measured static and total pressures to the ISA model. Accurate calibration of pitot-static probes is therefore critical for every computed air data parameter.

Why the Model Matters in the Hangar

Because the ADC applies a fixed model, any error in the sensed static pressure is converted into an altitude error by the model's own gradient. A static port left partially blocked by masking tape or paint over-reads or under-reads altitude by an amount that grows with height, and the same corrupted static feeds the airspeed and Mach computations — so a single static defect can produce simultaneous altitude, airspeed, Mach and vertical-speed anomalies that look like multiple system failures. Always suspect a common pneumatic source before condemning several computers.



A static port partly blocked by masking tape corrupts the static pressure feeding the air data computer.

Bernoulli's Theorem and Subsonic Airflow

Bernoulli's theorem states that in a steady, incompressible airflow with no energy added or removed, the total energy (the sum of static pressure, dynamic pressure, and potential energy) remains constant along a streamline:

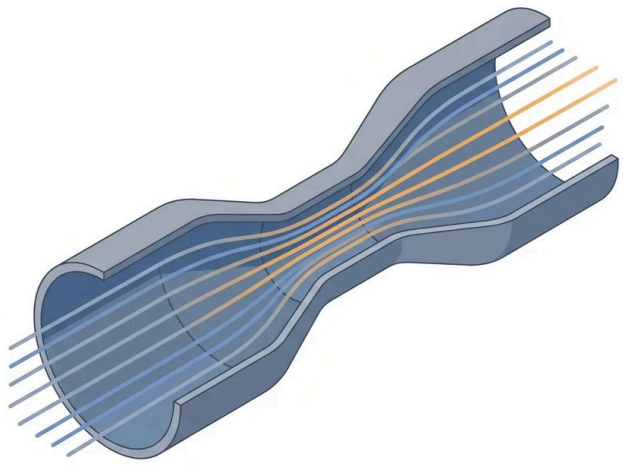
Bernoulli's Equation

$$P_{\text{static}} + \tfrac{1}{2}\rho V^2 + \rho g h = \text{constant}$$

Where P_{static} is static pressure, ρ is air density, V is velocity, and h is height. In level flight, the height term is neglected, giving:

$$P_{\text{total}} = P_{\text{static}} + \tfrac{1}{2}\rho V^2$$

Bernoulli's theorem cannot act alone — it needs the **continuity equation** to tell it where the air speeds up. Continuity says that for a steady flow through a duct or streamtube, mass flow is constant, so $\rho A V$ is the same at every station. In subsonic flow density is nearly constant, so a reduction in area forces an increase in velocity. That is the venturi: the throat is narrow, so the air accelerates, and Bernoulli then demands that its static pressure fall. A cambered aerofoil works as half a venturi, with the curved upper surface and the free stream above it forming the constriction.



In a venturi the narrow throat forces the air to speed up, and its static pressure falls.

Continuity and Dynamic Pressure

$$\rho_1 A_1 V_1 = \rho_2 A_2 V_2 \qquad \text{and} \qquad q = \tfrac{1}{2}\rho V^2$$

The term q is **dynamic pressure** — the pressure recovered when moving air is brought to rest. It is the quantity a pitot probe senses in addition to static pressure, and the quantity that scales every aerodynamic force on the airframe.

An aerofoil is shaped so air accelerates over the upper (cambered) surface. As velocity increases, static pressure decreases. The pressure difference between the lower and upper surfaces produces an upward net

force — **lift**. This is also supported by Newton's Third Law: the wing deflects air downward (downwash), and the equal and opposite reaction acts upward on the wing.

These two explanations are not rivals; they are two accounts of the same event. The pressure field that Bernoulli describes is exactly what turns the air downward, and the momentum change that Newton describes is exactly what the pressure field must produce. Ahead of the wing the air is turned **upwards** (upwash) before being turned down behind it, and at the leading edge there is a **stagnation point** where the flow is brought completely to rest and static pressure reaches its maximum — the total pressure. The stagnation point is not fixed: as angle of attack increases it migrates rearwards and downwards along the lower surface, which is precisely what a leading-edge stagnation-point stall sensor is built to detect.

Worked Example — Dynamic Pressure

Take ISA sea-level density $\rho = 1.225\ \mathrm{kg/m^3}$ and a speed of 100 m/s. Then $q = \frac{1}{2} \times 1.225 \times 100^2 = 6125\ \mathrm{Pa}$, so a pitot probe would sense a total pressure about 6125 Pa above the static pressure of 101 325 Pa. Double the speed to 200 m/s and q becomes 24 500 Pa — four times as much, because q follows V^2. This square law is why a small overspeed produces a disproportionately large structural load, and why airspeed indicator scales are compressed at the low-speed end.

Limits of the Incompressible Assumption

The simple form of Bernoulli's equation assumes constant density. That assumption is acceptable up to roughly M 0.3 and is progressively wrong above it, so air data computers apply a compressibility correction when converting the pitot-static pressure difference into calibrated and true airspeed. This is why the difference between indicated, calibrated, equivalent and true airspeed grows with both speed and altitude.

It also fixes the order of the air data chain, which is a favourite examination trap. **Mach number is derived from impact pressure and static pressure alone** — it is a function of the ratio q_c/P_s, and temperature is not an input to it. The dependency runs the other way round: the computer takes that Mach number, combines it with the measured **total air temperature** to strip out the ram rise and recover **static air temperature**, and only then can it obtain the local speed of sound and turn Mach into true airspeed. So an ADC that has lost its total air temperature input still produces a valid altitude, a valid computed airspeed and a valid Mach number; what it loses is static air temperature and true airspeed. A TAT probe or its heater failing in flight therefore typically flags TAS, SAT and the wind vector fed to the FMS while the Mach display and the M_{MO} pointer keep working — recognising that pattern tells the engineer to look at the temperature probe rather than at the pitot-static system.

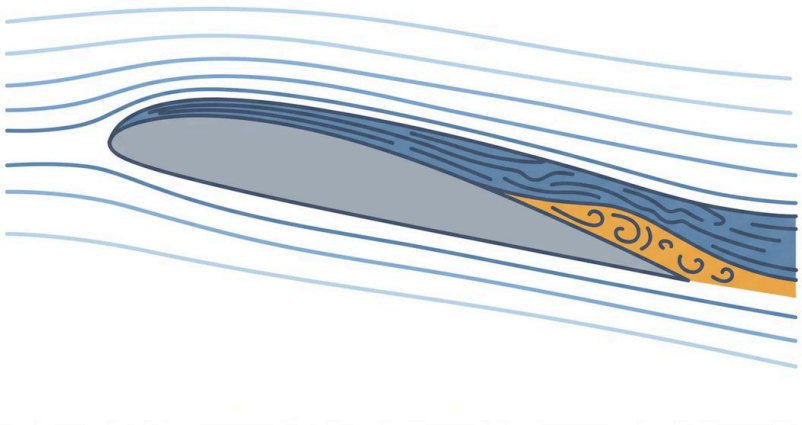
Boundary Layer

The boundary layer is the thin layer of air in direct contact with the wing surface where friction slows the airflow from the free-stream velocity to zero at the surface itself.

Air has viscosity, so the molecules immediately touching the skin do not move relative to it at all — the "no-slip" condition. Each layer above is dragged along by the one below through shear, and the velocity climbs from zero at the surface to essentially free-stream value within a layer that is typically only a few millimetres thick near the leading edge, growing to a few centimetres near the trailing edge of a large wing. Thin as it is, the boundary layer governs skin-friction drag, decides whether and where the wing stalls, and is the reason surface finish is an airworthiness matter rather than a cosmetic one.

Type	Characteristics	Skin Friction
Laminar	Smooth, orderly layers; found near the leading edge	Low
Turbulent	Irregular, mixed flow; better energy exchange with free-stream	Higher
Separated	Flow has left the surface; reversed flow and eddies beneath	Low friction but very high form drag

The point where the laminar boundary layer transitions to turbulent is called the **transition point**. A turbulent boundary layer resists separation better but produces more skin-friction drag. The point where the boundary layer separates from the surface is the **separation point** — when this moves forward significantly, the wing stalls.



Over the wing the boundary layer runs laminar, turns turbulent, then separates into reversed flow and eddies.

Why the Flow Separates

Over the front of the wing the pressure is falling in the direction of flow — a favourable gradient that accelerates the boundary layer and helps hold it on. Behind the point of minimum pressure the pressure rises again towards the trailing edge, and this **adverse pressure gradient** works against the flow. The air in the boundary layer has already lost energy to friction, so it is the least able to climb that pressure hill. Where its remaining momentum is exhausted the flow stops, reverses, and lifts off the surface: separation. Increasing

angle of attack steepens the adverse gradient and drags the separation point forward, which is the direct mechanism of the stall.

A turbulent boundary layer mixes high-energy free-stream air down towards the surface, so it carries more momentum near the skin and can fight further up the pressure hill before it gives up. That is the whole logic behind boundary layer control: designers frequently accept a small penalty in skin-friction drag in exchange for a much later separation. **Vortex generators** — the small rows of upright vanes seen ahead of the ailerons and on engine nacelles — stir high-energy air down into the boundary layer. **Slots** and slotted slats duct high-pressure air from beneath the wing onto the upper surface for the same reason: to reinforce and re-energise the boundary layer so that it stays attached to a higher angle of attack, raising the stalling angle.

✈ Maintenance Relevance

Surface condition changes where transition happens. Chipped paint, proud rivets, a mis-set access panel, insect residue on the leading edge, tape repairs and dried de-icing fluid all trip the laminar boundary layer early and move the transition point forward, increasing drag and cruise fuel burn. Missing or bent vortex generators are worse than a drag penalty — they were fitted to fix a specific separation problem, so removing them can degrade low-speed handling or roll control. Vortex generators are flight-critical fittings and their presence, count and alignment are inspection items, not cosmetic ones.

↑ Generation of Lift

Lift Equation

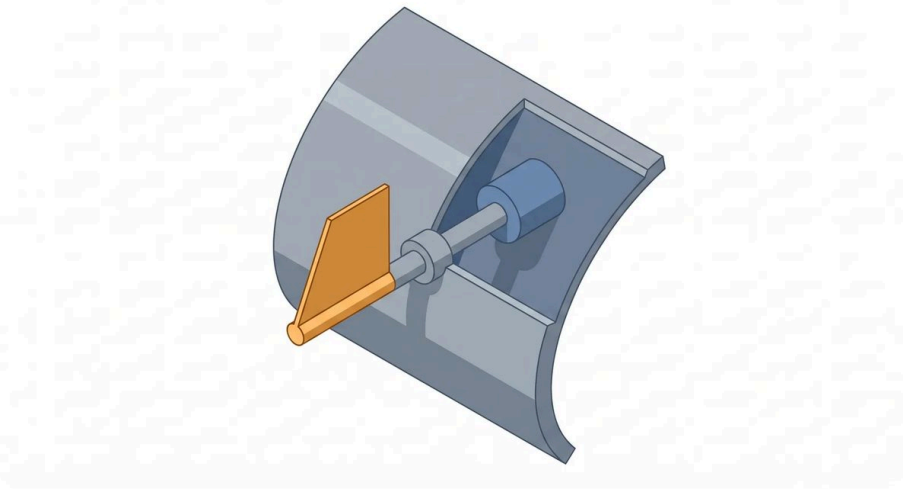
$$L = C_L \times \tfrac{1}{2}\rho V^2 \times S$$

Where C_L = lift coefficient (depends on angle of attack and aerofoil shape), ρ = air density, V = true airspeed, and S = wing area.

Angle of attack (AoA) is the angle between the chord line and the relative airflow. As AoA increases, C_L increases — up to a critical angle (typically 15–18°) beyond which the flow separates from the upper surface and lift drops abruptly: this is a **stall**.

Note carefully what angle of attack is measured *against*. It is the relative airflow — the direction of the air actually meeting the aerofoil — and not the horizon, not the aircraft's longitudinal axis, and not any fixed geometric reference. The angle between the chord line and the longitudinal axis is the fixed **angle of incidence** set at build; the angle between the longitudinal axis and the horizon is **pitch attitude**. All three can differ at once: an aeroplane descending steeply with a level attitude has a substantial positive angle of attack, and an aeroplane in a ballistic pushover can hold a nose-high attitude at zero or negative angle of attack. This

distinction is exactly why a stall warning system must sense angle of attack directly with a vane or probe rather than infer it from pitch attitude.



A stall warning system senses angle of attack directly with a vane, not from pitch attitude.

A rising gust makes the same point in reverse. If the air ahead of the leading edge is moving upwards, the relative airflow is tilted upwards even though the aeroplane has not moved, so the **angle of attack increases**. Combine that with a lowered trailing-edge flap — which already increases the effective camber and effective angle of attack of the section — and the wing can be pushed appreciably closer to the stall by turbulence alone. This is the mechanism behind gust-induced stall warnings during a flapped approach in gusty conditions, and it is why the manoeuvre and gust envelopes are drawn as they are.

Reading the Lift Equation

In steady level flight lift equals weight, so the equation is really a constraint: if ρ falls with altitude, or V falls in the approach, then C_L must rise to compensate, and the only way the pilot can raise C_L is by increasing angle of attack or changing configuration. That is why an approach is flown at a higher angle of attack than a cruise, and why a high-altitude cruise is flown closer to the buffet boundary than a low-level one. It also explains why the stall always occurs at the same *angle* but at very different *speeds*.

Wing loading is the aircraft's weight divided by its wing area, W/S. Because both quantities are properties of the aircraft itself, wing loading is **independent of altitude and of air density** — climbing does not change the weight carried per square metre of wing. Density changes how much lift a given speed produces, but it cannot change a ratio built from weight and area. High wing loading gives a smoother ride in turbulence and a higher stalling speed; low wing loading gives better field performance and a gustier ride.

Wing Loading and Aspect Ratio

$$\text{Wing loading} = \frac{W}{S} \qquad\qquad AR = \frac{b^2}{S}$$

Where W = weight, S = wing area and b = wing span. High aspect ratio (long, slender wings) reduces induced drag and improves glide performance; low aspect ratio favours roll rate, structural simplicity and high-speed flight.

Two more terms are needed to read an aerofoil drawing. The **chord line** is the straight line from leading edge to trailing edge; **camber** is the curvature of the mean line away from that chord. The **centre of pressure (CP)** is the point through which the resultant aerodynamic force acts. On a conventional cambered aerofoil the CP moves *forward* as angle of attack increases up to the stall, and then jumps sharply rearward at the stall — a rearward CP movement that produces the nose-down pitch many aircraft show at the break.

Aerofoil Types

Type	Features	Typical Use
Symmetrical	No camber; zero lift at zero AoA	Tail surfaces, aerobatic aircraft
Cambered (conventional)	Upper surface more curved; lift at zero AoA	General aviation, transport
Supercritical	Flat upper surface, reflex camber at trailing edge; delays shock wave	High-subsonic transport aircraft
Laminar flow	Maximum thickness far aft; low drag	High-performance gliders
Thin, sharp-nosed	Very low thickness/chord ratio and a sharp leading edge; minimises wave drag but stalls abruptly at low speed	Supersonic aircraft, missile fins

Drag

Total drag is the sum of **parasite drag** and **induced drag**.

Drag Equation

$$D = C_D \times \tfrac{1}{2}\rho V^2 \times S$$

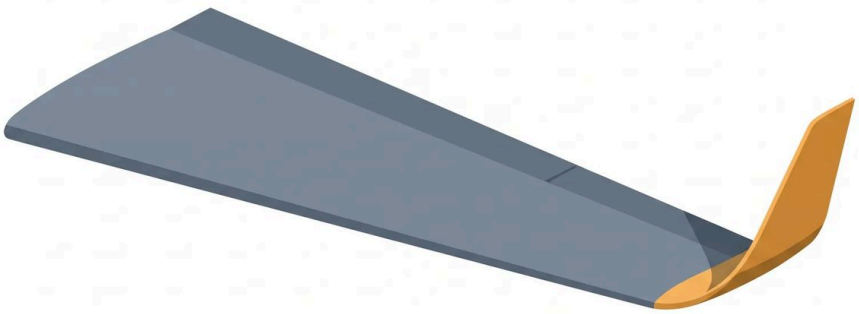
Drag Type	Cause	Varies With Speed
Parasite drag	Form (shape), skin friction, interference	Increases with V^2
Induced drag	By-product of lift; caused by wingtip vortices deflecting airflow downward	Decreases with V^2 (highest at low speed/high AoA)
Wave drag	Energy lost across shock waves once local flow reaches M 1.0	Zero below M_{crit}, then rises steeply with Mach number

The **lift/drag (L/D) ratio** peaks at the speed where parasite drag equals induced drag. This is the most aerodynamically efficient speed and determines best glide range.

The Three Components of Parasite Drag

- **Form (pressure) drag** — caused by the wake behind a bluff body, where the pressure never fully recovers. Streamlining and fairings attack this component directly; an unfaired landing gear leg or a missing gap seal can cost far more drag than its size suggests.
- **Skin friction drag** — the shear stress of the boundary layer acting over the whole wetted area. It depends on surface finish and on how much of the boundary layer is turbulent.
- **Interference drag** — extra drag generated where two flow fields meet, such as the wing-root, pylon and empennage junctions. This is why blended fillets exist and why a damaged or missing fillet fairing is not a purely cosmetic defect.

Induced drag is a direct and unavoidable consequence of producing lift with a finite wing. Higher pressure beneath the wing spills around the tip into the lower pressure above it, creating a trailing vortex at each tip. Those vortices tilt the local relative airflow downwards over the wing (downwash), which tilts the lift vector rearwards, and the rearward component of that vector *is* induced drag. Because it depends on how hard the wing is working, induced drag varies with C_L^2 and therefore falls rapidly as speed increases. It is inversely proportional to aspect ratio, which is why gliders have long slender wings, and it is what winglets and raked tips are fitted to reduce by weakening the tip vortex.



A winglet at the wingtip weakens the tip vortex, reducing the induced drag the wing produces.

Worked Example — Best Glide

Suppose an aeroplane has a maximum lift/drag ratio of 16 to 1, achieved at its minimum-drag speed. In a power-off glide the still-air glide ratio equals the L/D ratio, so from 6 000 ft above the terrain it can cover approximately $16 \times 6000 = 96\,000$ ft, or about 16 nm. Note what does *not* change this: weight. A heavier aeroplane glides the same distance but must fly faster to achieve the same L/D, and so arrives sooner. Wind does change it — a headwind shortens the ground distance achieved even though the air distance is unchanged. (The 16:1 figure is a stated assumption for this example, not a value for any particular type.)

Plotting the two components against speed gives the familiar U-shaped total drag curve. Its minimum is V_{MD}, the minimum-drag speed, which is also the maximum L/D speed. Flight slower than V_{MD} lies on the **back of the drag curve**, where reducing speed *increases* drag and therefore increases the thrust required — a speed-unstable region that requires deliberate power management on approach and is a recognised contributor to low-energy approach accidents.

Stalling

A stall occurs when the critical angle of attack is exceeded — the boundary layer separates from the upper surface and lift decreases sharply. The stall speed depends on weight, load factor, configuration (flaps/slats), and bank angle:

The single most important idea is that a wing stalls at an **angle**, not at a speed. For a given aerofoil and configuration the critical angle of attack is fixed, and an aeroplane can be stalled at any airspeed and any attitude if it is manoeuvred hard enough to reach that angle. The published stall speed is simply the speed at which, in one specific condition of weight, configuration and 1 g flight, the wing happens to arrive at that angle.

Change the weight, the load factor or the flap setting and the stall speed changes while the stall angle does not.

Stall Speed in a Turn

$$V_{s,\text{turn}} = V_{s,1\text{g}} \times \quad \not{\sqrt{}}$$

Where n = load factor. In a 60° bank, $n = 2$, so stall speed increases by a factor of $\sqrt{2} \approx 1.41$ (41% increase).

Worked Example — Bank and Weight

Assume a clean 1 g stall speed of 120 kt. In a 60° balanced turn the load factor is $n = 1/\cos 60^\circ = 2$, so the stall speed becomes $120 \times \sqrt{2} \approx 170$ kt. Weight follows the same square-root law: increase the weight by 21% and the stall speed increases by $\sqrt{1.21} = 1.1$, i.e. 10%, giving 132 kt. The 120 kt starting figure is an assumed value for the arithmetic, not a type-specific number.

Stall Warning

A stall warning must arrive early enough to be useful, so it is deliberately set to trigger **at a speed higher than the actual stall speed** — equivalently, at an angle of attack lower than the critical angle. A warning at, or below, the stall speed would be an announcement rather than a warning, arriving after the point at which recovery action could prevent the stall. Certification requires a clear and distinctive warning with an adequate margin in every configuration, and because the stalling angle changes with flap and slat position, the warning threshold must be scheduled against configuration. That scheduling is the reason a stall warning computer takes flap and slat position inputs at all — and the reason a mis-rigged flap position sensor can produce a spurious or absent stick shaker.

- **Angle of attack vane or probe** — a small external vane that aligns itself with the relative airflow and drives a resolver or synchro; the primary sensing element on transport aircraft.
- **Leading-edge lift transducer (reed or flapper switch)** — a spring-loaded flap in the leading edge that is sucked upwards as the stagnation point migrates down and aft at high angle of attack, closing a switch.
- **Stick shaker** — an eccentric motor on the control column giving an unmistakable tactile and audible warning, the standard transport-category warning device.
- **Stick pusher** — fitted where the natural stall characteristics are unacceptable (notably rear-engined T-tail aircraft prone to deep stall); it applies an automatic nose-down input rather than merely warning.
- **Aerodynamic pre-stall buffet** — turbulent wake from the separating inboard wing striking the tailplane. It is a natural warning, but it can be masked by ice or by a very clean modern wing, which is exactly why artificial warning is required.

Designing Where the Wing Stalls First

A wing that stalled at the tip first would lose the ailerons at the moment they are most needed, and the loss of lift outboard would generate a violent rolling moment. Designers therefore force the root to stall first, using **washout** (a built-in reduction of incidence towards the tip so the tip is at a lower angle of attack than the root), **stall strips** on the inboard leading edge to trip separation early, and root aerofoil sections with less gentle stalling behaviour than the tip sections. Sweptback wings fight this tendency, because spanwise flow thickens the boundary layer outboard, and so they commonly need **wing fences**, vortilons or saw-tooth leading edges as well.

⚠ Down-Going Aileron Near the Stall

Deflecting an aileron downwards adds camber and increases the effective angle of attack of that wing section. At low speed and an already high angle of attack this means the down-aileron wing **reaches the stall sooner — its local stall speed is increased**. Attempting to pick up a dropping wing with aileron near the stall can therefore stall the down-going wing further and aggravate the roll. Designers mitigate this with **differential ailerons** (the up-going aileron moves further than the down-going one), **Frise ailerons**, aileron up-float rigging and, on many transport aircraft, roll spoilers that take over roll control at high angle of attack.

Two special stalls are worth naming. A **deep stall** (superstall) affects T-tail aircraft when the stalled wing wake blankets the tailplane, removing elevator authority and locking the aircraft into the stall — the reason those types carry stick pushers. A **shock stall**, by contrast, **occurs at high speed**: the local airflow over the upper surface reaches M 1.0, a shock wave forms, and the boundary layer separates immediately behind it. Both are separations, but one is produced by exceeding the critical angle at low speed and the other by exceeding the critical Mach number at high speed. Confusing the two is a classic examination trap.

❄ Aerofoil Contamination

Ice, frost, or even rain on wing surfaces disrupts the boundary layer, reduces maximum C_L, increases drag, and lowers the stall angle of attack. Just 1–2 mm of frost can reduce lift by up to 30% and increase stall speed significantly. This is why the avionics stall-warning system calibration must account for clean-wing assumptions, and ice-detection systems must trigger appropriate crew alerts.

❄ Aerofoil Contamination — Ice, Frost, Snow and Rain

The callout above states the outcome; this section supplies the mechanism behind it and what the engineer is expected to do with that knowledge. Contamination is treated separately in the syllabus because it attacks the wing in a way no other condition does: it degrades performance and it defeats the warning system at the same

time. Every stall warning threshold, every published stall speed and every certified performance figure assumes a clean aerofoil. Contamination invalidates that assumption without producing any cockpit indication that the assumption has been invalidated. That is what makes the quoted lift loss so dangerous — not its size, but the fact that no instrument reports it and no protection compensates for it.

Mechanism

Roughness is the primary mechanism, and it does its damage where the flow is most delicate — on the upper surface just aft of the leading edge, where the boundary layer is thin, laminar and about to meet the adverse pressure gradient. Surface texture no coarser than medium sandpaper is enough to trip the boundary layer prematurely and rob it of the momentum it needs to stay attached. The separation point moves forward, so the wing reaches its maximum lift coefficient at a *lower* angle of attack than the clean wing does. The consequences follow directly:

- **The speed schedule stops being valid** — every V-speed on the crew's card was computed from clean-wing data, so a rotation or approach speed that carried a healthy margin can end up at or below the speed at which the contaminated wing now stalls. The margin is consumed silently, before the take-off roll begins.
- **The protection that assumes a clean wing is the protection that fails first** — the vane or reed switch still triggers at its fixed, clean-wing angle, but the wing now reaches maximum lift below that angle. The shaker therefore fires late or does not fire at all, and the artificial warning fitted precisely because natural buffet is unreliable is itself disabled by the same contamination.
- **The drag penalty arrives at the worst possible moment** — roughness drag and premature separation act together while the aircraft is heavy, slow and low, cutting the second-segment climb gradient in the one phase of flight where there is nothing in reserve to trade.
- **Handling degrades** — asymmetric contamination produces asymmetric stall onset and a roll at the break, and ice on the tailplane can produce a tailplane stall with an abrupt nose-down pitch, often on flap extension.
- **Weight increases and control surfaces can jam**, with ice in hinge gaps, tab linkages and balance-weight recesses restricting free movement of the surface.

Types of Contamination

Contaminant	How It Forms	Principal Effect
Hoar frost	Water vapour freezing directly onto a cold surface, typically overnight on the ground	Roughness with negligible weight or shape change; still severely reduces maximum lift
Cold-soaked fuel frost	Moisture freezing onto upper wing skin chilled by cold fuel after a long sector, even in above-freezing ambient	Same roughness effect, but forms in weather no one expects icing in

Contaminant	How It Forms	Principal Effect
Rime ice	Small supercooled droplets freezing on impact, trapping air	Opaque, brittle, follows the leading-edge shape; mainly a roughness and shape penalty
Clear (glaze) ice	Large supercooled droplets running back before freezing	Heavy, tenacious, can build behind the protected area as runback ice and form a lift-destroying horn
Snow and slush	Precipitation lying on the surface, or thrown up by the wheels on a contaminated runway	Bulk roughness and weight; slush also adds retardation drag during the take-off roll
Heavy rain	A water film with surface waves, plus droplet impact momentum	Effective surface roughening reduces maximum lift and adds drag while it persists

Avionics and Maintenance Consequences

For the B2 engineer the contamination case is where several systems intersect. Ice detectors — vibrating-probe types that sense the frequency shift caused by accreted ice, or optical types — provide the advisory or automatic anti-ice activation. Heated probes protect the pitot, static and total air temperature sensors and the angle of attack vanes; a failed vane heater in icing conditions is a serious defect precisely because a frozen vane can report a plausible but false angle of attack to the stall warning and, on fly-by-wire types, to the flight control laws. Blocked or partially blocked pitot and static ports produce erroneous airspeed and altitude, and any comparison monitor that then declares a disagreement is reporting a symptom, not the cause.

⚠ The Clean Aircraft Concept

No aircraft may be dispatched with frost, ice, snow or slush adhering to the wings, control surfaces, propellers, engine inlets or other critical surfaces. Thin, hard frost on the fuselage skin is generally acceptable within limits stated in the aircraft documentation, but the lifting and control surfaces are not negotiable. De-icing fluid removes existing contamination; anti-icing fluid provides a limited holdover time before it must be regarded as expired. Never assume a residual fluid film is harmless — dried fluid residue can rehydrate and refreeze in unpowered control-surface gaps, and the applicable procedures and holdover data are those in the approved documentation for the type and the fluid actually used.



Even thin frost on a wing's upper surface is enough roughness to cut maximum lift, so it must be removed.

⚖️ Stability

Static stability — the initial tendency to return to the original state after a disturbance. **Dynamic stability** — whether the oscillations following a disturbance decrease (stable), remain constant (neutral), or increase (unstable) over time.

The two are independent, and the distinction matters. An aeroplane can be statically stable but dynamically unstable: each disturbance produces a restoring moment, but the aircraft overshoots by more each time and the oscillation diverges. Static stability is a prerequisite for dynamic stability, but it is not a guarantee of it. Note also that stability and controllability pull in opposite directions — the more strongly an aeroplane resists being disturbed, the more control force is needed to manoeuvre it deliberately, which is why designers aim for adequate rather than maximum stability.

Axis	Stability Type	Primary Contributor
Longitudinal (pitch)	Most important	Horizontal stabiliser, CG position
Lateral (roll)	Dihedral effect	Wing dihedral, sweepback, high wing
Directional (yaw)	Weathercock stability	Vertical fin (area and moment arm)

Longitudinal Stability and the CG

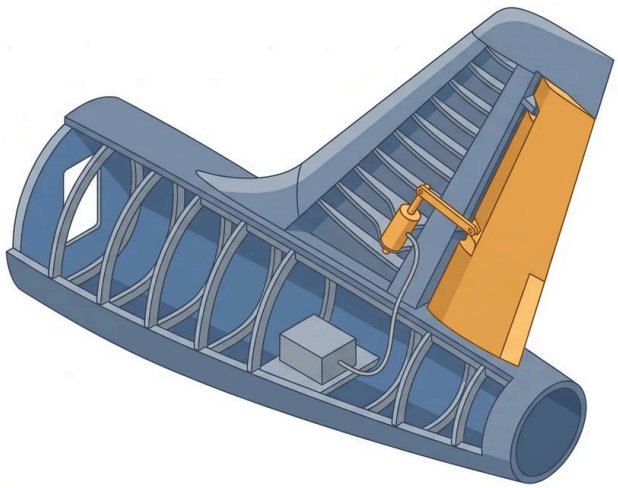
Longitudinal stability is dominated by the position of the centre of gravity relative to the **neutral point**, the CG position at which the aircraft would be neutrally stable. The distance between them, expressed as a fraction of the mean aerodynamic chord, is the **static margin**. A forward CG gives a large static margin: strongly stable, heavy in pitch, a higher stall speed because the tailplane carries more download, and a longer take-off run. An aft CG gives a small static margin: light and responsive in pitch, lower trim drag and better cruise economy,

but reduced stall recovery margin and, beyond the aft limit, no stability at all. This is the entire reason mass and balance limits exist, and why a loading error is an airworthiness matter rather than a performance nicety.

Dynamic Modes

- **Phugoid** — a long-period, lightly damped exchange of height and speed at roughly constant angle of attack. Slow enough that the pilot or autopilot corrects it without difficulty.
- **Short-period pitch oscillation** — a rapid oscillation in angle of attack, normally heavily damped. If it becomes lightly damped it is dangerous, because it is fast enough for pilot inputs to reinforce it.
- **Spiral mode** — a slow divergence in which a small bank is allowed to steepen. Common and usually acceptable, since it develops slowly enough to correct.
- **Dutch roll** — the coupled lateral-directional oscillation described below.

Dutch roll affects **yaw and roll simultaneously**, not pitch. It arises when the dihedral effect (lateral stability) is strong relative to the weathercock effect (directional stability) — precisely the combination produced by a swept, high-aspect-ratio wing on a modern jet. A small yaw makes the advancing wing produce more lift, so the aircraft rolls; the roll produces sideslip, which yaws the aircraft the other way; and the motion repeats as a wallowing, alternating yaw-and-roll. Its **frequency is low** — a period of several seconds — but because it is only lightly damped and worsens with altitude, it is uncomfortable and can be fatiguing to fly manually. The **yaw damper** senses the yaw rate with a rate gyro and applies small, rapid, opposing rudder inputs that are transparent to the pilot's own pedal inputs. Loss of the yaw damper is normally accompanied by an altitude or speed restriction for exactly this reason.



A yaw damper: a rate gyro senses yaw rate and drives small, rapid, opposing rudder inputs.

Avionics Relevance

Stability augmentation is increasingly handled electronically. Yaw dampers suppress Dutch roll (coupled yaw-roll oscillation), and fly-by-wire flight control computers provide artificial stability. The B2 engineer maintains the sensors (rate gyros, accelerometers) and computers that make these systems work.

Relaxed Static Stability

Fly-by-wire allows the CG to be scheduled aft of the position a mechanically controlled aircraft could tolerate, sometimes to the point of natural instability. The tailplane then carries less download, which reduces trim drag and fuel burn. The consequence for the engineer is that the flight control computers and their inertial and air data sources are no longer merely convenient — they are the only thing making the aircraft flyable, which is why those systems are multiply redundant and dissimilar, and why a triple air data disagreement forces a reversion to a degraded control law.

Flight Controls

Primary Controls

Control	Axis	Cockpit Input	Movement
Ailerons	Lateral (roll)	Control wheel / sidestick lateral	Differential — one up, one down
Elevator	Longitudinal (pitch)	Control column fore/aft	Up for nose up, down for nose down
Rudder	Directional (yaw)	Rudder pedals	Deflects to yaw aircraft
Elevons	Pitch and roll combined	Column and wheel, mixed mechanically or electronically	Together for pitch, differentially for roll
Ruddervators	Pitch and yaw combined	Column and pedals, mixed	Together for pitch, differentially for yaw
Stabilator (all-moving tailplane)	Longitudinal (pitch)	Control column fore/aft	Whole surface rotates; usually with an anti-balance tab

The two combined-function surfaces are frequently examined. **Elevons** are fitted to delta-wing and tailless aircraft, which have no separate tailplane; moved together they act as elevators and **control pitch**, moved differentially they act as ailerons and **control roll** — so a delta wing's pitch and roll both come from the elevons. **Ruddervators** are the two surfaces of a V-tail (butterfly tail); moved together they act as an elevator and **control pitch**, moved differentially they act as a rudder and **control yaw**. Learn the pairs by what the

configuration lacks: a delta has no tailplane, so its trailing-edge surfaces must do pitch and roll; a V-tail has no separate fin or tailplane, so its surfaces must do pitch and yaw.

What the Cockpit Input Actually Commands

Pulling the control column **back** raises the elevator, which produces a downward force on the tail, pitches the nose up and increases angle of attack. Moving the wheel or stick **left** rolls left, which requires the **left aileron up and the right aileron down** — the down-going aileron increases lift on the right wing, and the up-going aileron reduces it on the left. So a column pulled back and to the left gives **elevator up with the right aileron down**. Confirming this by hand during a control-and-rigging check, with the surfaces observed directly rather than assumed from an indicator, is a standard and non-negotiable duplicate-inspection item after any work on the flying control system.



In a rigging check the surfaces are watched directly as the column moves, not assumed from an indicator.

The rudder has a secondary rolling effect that follows from geometry. Pushing the **left rudder pedal** deflects the rudder and yaws the nose left; the outer (right) wing then travels faster through the air than the inner wing, generates more lift, and **rises**, so the aircraft rolls left as it yaws left. This roll-with-yaw coupling is the same effect that drives Dutch roll, and it is what makes rudder alone a usable, if crude, roll control after a lateral control failure.

⚠ Adverse Yaw

The down-going aileron produces more induced drag than the up-going one produces reduction, so the wing that is rising is also being dragged back. The nose therefore yaws *away* from the intended turn — adverse yaw, or aileron drag. Design fixes are differential aileron travel, Frise ailerons whose up-going leading edge protrudes into the airflow beneath the wing, roll spoilers, and aileron-rudder interconnects. A rigging error that removes the built-in differential will not stop the aircraft rolling, so it may pass a superficial functional check, yet it degrades handling in exactly the regime where handling matters most.

Secondary Controls

- **Trim tabs** — small surfaces on primary controls to relieve stick forces in steady flight
- **Flaps** — increase camber and wing area; lower stall speed for take-off and landing
- **Slats** — extend from the leading edge to energise boundary layer and increase stall angle
- **Spoilers** — rise from wing upper surface to destroy lift (in flight: speed brakes; on ground: lift dumpers)
- **Speed brakes** — increase drag without significantly changing pitch
- **Slots** — fixed or automatic gaps that duct high-pressure air from below onto the upper surface to reinforce the boundary layer and raise the stalling angle
- **Vortex generators** — small vanes that energise the boundary layer locally to delay separation over ailerons, nacelles and the tailplane

Tabs — Four Types, Four Behaviours

Tab	Moves	Purpose
Trim tab	Set independently by the pilot; then fixed relative to the surface	Holds the surface where the aerodynamic loads balance, so the pilot need hold no residual force — it eases control loading
Balance tab	In the **opposite** direction to the control surface, proportionally to its movement	Aerodynamic force on the tab assists the surface, reducing hinge moment and stick force
Anti-balance tab	In the **same** direction as the control surface, proportionally	Deliberately increases stick force to restore feel on an over-light surface, typically an all-moving stabilator
Servo tab	Driven directly by the pilot's control run; the main surface follows aerodynamically	Pilot moves only the small tab, which flies the large surface — used on large unpowered controls
Spring tab	Acts as a servo tab only once spring pre-load is overcome at higher airspeed	Gives assistance in proportion to airspeed, so stick force stays consistent across the speed range

The two examinable points are that a **trim tab eases control loading for the pilot** — it does not permit the centre of gravity to be flown outside its certified limits, and it does not make the controls finer or more precise — and that a **balance tab moves in the opposite direction to the surface it is attached to, proportionally to that surface's movement**. An anti-balance tab is its mirror image in both motion and purpose, so a tab linkage assembled one hole out, or reversed, can convert an assistance device into a resistance device or vice versa. Tab rigging and free-play checks exist for exactly this reason, and control-surface tabs are also a classic flutter source when their linkage develops wear.

Flaps and the Pitch Change

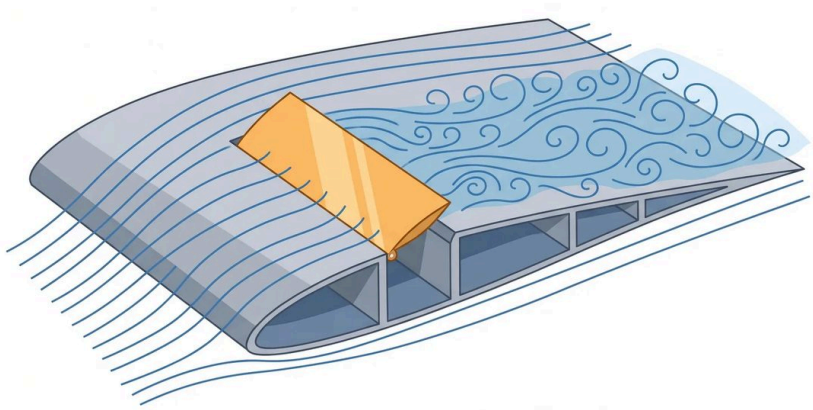
Flaps increase the camber of the rear of the wing, and the larger types (Fowler and slotted-Fowler) also translate rearwards to increase wing area. Both effects raise the maximum lift coefficient, so the wing can support the weight at a lower speed. Because a landing flap setting is a large deflection chosen for maximum lift and maximum drag, its principal benefit is a **lower landing speed** together with a steeper approach path and better speed control. A take-off flap setting is a smaller deflection, chosen to buy lift without the drag penalty that would spoil the climb gradient — so the full landing setting is not what reduces take-off speed, and selecting it for take-off would be actively harmful. Whatever the setting, lowering flap **increases both lift and drag together**: the same increase in camber and frontal area that generates the extra lift also generates the extra drag, and this is precisely what makes short-field landings possible.

Extending a trailing-edge flap loads the rear of the aerofoil more heavily, so the **centre of pressure moves rearward**. A rearward CP acting behind the centre of gravity produces a **nose-down pitching moment** on most configurations. Note two things the exam probes here: the movement is rearward, not forward, and the aircraft's **centre of gravity does not change** at all — moving a control surface redistributes aerodynamic pressure, it does not move mass in a way that shifts the CG. On a low-set tailplane the flap downwash may reach the tail and partly offset the pitch change; the net trim change is therefore type-specific, but the CP movement itself is not.

Flap Type	Action	Relative Lift / Drag
Plain	Rear portion of the aerofoil hinges down	Modest lift increase, large drag increase
Split	Lower surface only hinges down, upper surface unchanged	Similar lift to plain, more drag
Slotted	A gap ducts high-energy air over the flap upper surface	Better lift than plain for less drag, because separation is delayed
Fowler	Translates aft on tracks before deflecting, increasing area and camber	Greatest lift increase; drag stays low at small settings and rises steeply at large ones
Krueger (leading edge)	Hinges forward and down from the lower leading edge	Raises stalling angle; commonly fitted inboard alongside outboard slats

Spoilers, Speed Brakes and Lift Dumpers

Flaps and tabs add lift or trim; spoilers do the opposite, and they are the only aerodynamic device on the aeroplane whose purpose is to make the wing work *worse*. A spoiler is an upper-surface panel hinged at its leading edge that rises into the airflow. It does not simply add an obstruction: it forces the boundary layer to separate over the whole of the wing behind it, so the suction peak on the upper surface collapses. Lift is destroyed over that part of the span, and form drag rises as a consequence. That order matters for the exam — a spoiler is a **lift-destroying** device that also produces drag, whereas a true speed brake is a **drag-producing** device designed to change lift and pitch as little as possible.



A raised spoiler panel forces the flow behind it to separate, collapsing upper-surface suction and lift.

On most transport aircraft the same row of wing panels performs all three roles, and the flight control computers decide which one is active from the aircraft's state. The distinction the engineer must hold is functional, not physical.

Function	How the Panels Are Driven	What It Achieves
Speed brake (in flight)	Both wings together, partial travel, from the speed brake lever; travel is usually limited with flap out and blown down by airload at high speed	Increases drag so the aircraft can descend steeply or decelerate without changing thrust; the lift lost is made up by a slightly higher angle of attack
Roll spoiler (roll assist)	Asymmetrically, on the down-going wing only, mixed with the aileron command	Rolls the aircraft by spoiling lift on the inside wing; the extra drag is on the inside of the turn, so it yaws *into* the turn instead of producing adverse yaw
Lift dumper (ground spoiler)	All panels, full travel, automatically on touchdown once the arming, weight-on-wheels, wheel spin-up or thrust lever conditions are met	Kills residual lift so the weight transfers from the wings onto the wheels, which is what makes wheel braking and anti-skid effective; also adds a large drag increment

Two consequences follow that are easy to miss. First, the roll-spoiler function is why high-speed roll control can be handed to the spoilers: a spoiler cannot suffer the aileron reversal described later under high-speed flight, because it works by separating flow rather than by twisting the wing with a trailing-edge load. Second, the lift-dumper function is a *braking* system as much as an aerodynamic one — an aircraft that touches down with residual wing lift has little weight on its tyres, so the brakes and anti-skid have almost nothing to work with. This is the direct reason a failed weight-on-wheels switch, a wheel that does not spin up, or an unarmed speed brake lever produces a long landing roll that feels like a brake fault but is not one.

The dedicated **speed brake** of the original definition is the fuselage- or tail-mounted airbrake found on some types, and its "without significantly changing pitch" characteristic is a deliberate design result: the panels are placed and sized so that the drag line acts close to the centre of gravity and the wake misses the tailplane, so the aircraft slows without a trim change the pilot must fight. Wing-mounted spoilers used as speed brakes cannot achieve that as cleanly — their wake washes over the tail and their lift loss is well away from the CG, which is why deploying them typically produces buffet and a noticeable pitch and trim change that the type's flight manual describes.

Control System Protections

Because a full-travel surface deflection safe at approach speed would overstress the airframe at cruise speed, large aircraft limit authority with airspeed. Rudder travel is restricted by a rudder ratio changer or limiter scheduled on air data; ailerons may be locked out in favour of spoilers at high speed; and flap and slat systems carry load-relief logic that retracts a stage automatically if the placard speed is approached. Every one of those functions depends on air data and position feedback, so an out-of-tolerance flap position transmitter or a lagging air data input can produce control-law behaviour that looks like a hydraulic or mechanical fault but is neither. Mass balance weights and their attachments must also be treated as flight-critical: they exist to keep the surface's centre of gravity ahead of the hinge line and prevent flutter, and paint build-up, water ingress or an unapproved repair that changes surface mass distribution can reintroduce a flutter margin problem.

High-Speed Flight

As an aircraft approaches the speed of sound ($a = \sqrt{\gamma R T} \approx 340\,\mathrm{m/s}$ at sea level), compressibility effects become significant. The **Mach number** is the ratio of TAS to the local speed of sound:

Mach Number

$$M = \frac{V_{\mathrm{TAS}}}{a}$$

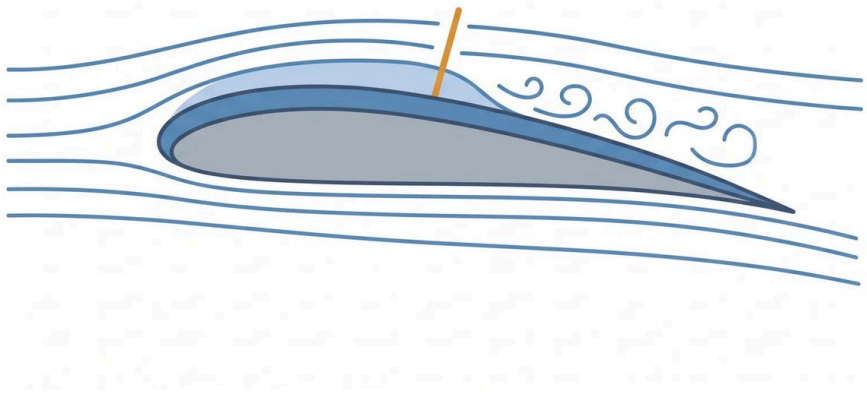
The formula $a = \sqrt{\gamma RT}$ contains one variable that matters operationally: temperature. The speed of sound depends on *temperature alone*, not on pressure or density, because sound is transmitted by molecular collisions and molecular speed is set by temperature. As an aircraft climbs into colder air, the local speed of sound falls — from about 340 m/s at ISA sea level to roughly 295 m/s at the tropopause. A constant true airspeed therefore represents a steadily increasing Mach number in the climb. This is why the climb profile changes over from a constant indicated airspeed to a constant Mach number at the crossover altitude, and why the airspeed indicator carries a moving barber-pole V_{MO}/M_{MO} pointer driven by the air data computer rather than a fixed red line.

Regime	Mach Range	Characteristics
Subsonic	M < 0.75	No shock waves
Transonic	0.75 – 1.2	Mixed sub/supersonic flow; shock waves form
Supersonic	1.2 – 5.0	Entire flow supersonic; bow and oblique shocks
Hypersonic	Above M 5.0	Severe kinetic heating; chemical changes in the airflow

Note the naming convention the exam uses: speeds above the speed of sound but not exceeding roughly four times it are **supersonic**, and only around M 5 and above does the term **hypersonic** apply. "Hyposonic" is not a recognised regime at all. Below the transonic band the flow is subsonic everywhere on the airframe; the transonic band is defined by the coexistence of subsonic and supersonic flow over the *same* aircraft, which is what makes it the most aerodynamically awkward region to fly in.

Mcrit is the free-stream Mach number at which local airflow first reaches M = 1.0 (typically on the upper wing surface). Beyond M_{crit}, shock waves produce **wave drag** and can cause shock-induced separation (Mach buffet), pitch changes (Mach tuck), and aileron reversal. Wing sweepback, thin aerofoils, and supercritical wing sections are all design features that raise M_{crit}.

The mechanism behind each of those effects is worth having straight. A shock wave is a near-instantaneous compression across which the flow decelerates from supersonic to subsonic, with an abrupt rise in static pressure and an unrecoverable loss of energy. That energy loss appears as **wave drag**, so as the aircraft accelerates past M_{crit} the **total drag increases** sharply — the drag-divergence rise — on top of the parasite and induced drag already present. Thrust is set by the engine and is not what changes here, and lift can still be maintained; the dominant aerodynamic change on approaching supersonic flow is that large increase in total drag.



Beyond Mcrit a shock forms on the upper surface, and the flow behind it separates: Mach buffet.

- **Shock-induced separation and Mach buffet** — the steep adverse pressure gradient behind the shock separates the boundary layer, giving high-speed buffet. This is the **shock stall**: a separation stall that occurs at high speed rather than high angle of attack.
- **Mach tuck** — the shock forms first near mid-chord and moves aft with increasing Mach number, shifting the centre of pressure rearward and producing a nose-down pitching moment that steepens the dive and increases Mach further. A Mach trim system senses Mach number and applies a compensating nose-up trim automatically.
- **Aileron buzz and control reversal** — a shock sitting over a control surface makes its hinge moment erratic, and on a flexible wing a down-aileron can twist the outer wing leading-edge-down enough to reverse the intended roll. High-speed roll control is therefore commonly given to inboard ailerons and roll spoilers.
- **Reduced control effectiveness** — pressure changes cannot propagate forward through a shock, so a trailing-edge surface behind a shock loses much of its influence on the wing ahead of it.

Sweepback is the principal design answer, because only the component of the airflow perpendicular to the wing leading edge determines the local Mach number over the section; sweeping the wing reduces that component and so raises M_{crit}. The costs are real — reduced maximum lift coefficient, a tip-stall tendency from spanwise flow, and a stronger dihedral effect that brings on Dutch roll. Thin aerofoil sections and **supercritical** sections attack the same problem differently: the supercritical section's flattened upper surface keeps local accelerations low, holds the inevitable shock weak and far aft, and recovers lift with reflex camber at the trailing edge.

⚠ Coffin Corner

As altitude increases, the low-speed buffet boundary (set by stalling angle of attack) rises in terms of true airspeed while the high-speed buffet boundary (set by M_{crit}) falls in terms of indicated airspeed. At the aerodynamic ceiling the two converge and the usable speed band shrinks towards nothing — the "coffin corner". A small speed excursion, a temperature shear, or a bank angle that raises the load factor can then trigger either a low-speed or a high-speed buffet from the same starting point. This is why air data accuracy at altitude is safety-critical rather than merely operational, and why an unreliable-airspeed event in the cruise is treated with such gravity.

Module 14 — Propulsion

14.1(a) — Turbine Engines

Introduction to Gas Turbine Engines

The **gas turbine engine** is the dominant powerplant for modern commercial, military, and many general aviation aircraft. It operates on the **Brayton cycle** (also called the gas turbine cycle or constant-pressure combustion cycle) and produces either **thrust** (turbojet/turbofan) or **shaft power** (turboprop/turboshaft) by accelerating a mass of air.

Fundamental Energy Concepts

Potential Energy: Energy stored due to position or state. In propulsion context: chemical potential energy in fuel, pressure energy in compressed air.

$$PE = mgh \quad (\text{gravitational})$$

Kinetic Energy: Energy of motion. This is what produces thrust — accelerating air rearward.

$$KE = \frac{1}{2}mv^2$$

A jet engine converts chemical energy (fuel) → heat energy (combustion) → kinetic energy (high-velocity exhaust).

The engine as an energy converter, not an energy source

The first law of thermodynamics says energy can be neither created nor destroyed, only changed in form. A gas turbine therefore creates nothing: it is a machine for taking the chemical energy locked in a hydrocarbon fuel and handing as much of it as possible to the surrounding air as useful momentum. Every stage of that hand-over is imperfect, and understanding *where* the energy is lost is what separates a Level 3 answer from a recited definition.

Aviation kerosene carries a lower calorific value of roughly 43 MJ per kilogram. That figure is worth committing to memory, because it lets you sanity-check almost any engine performance claim: multiply the fuel flow in kg/s by 43 MJ/kg and you have the total rate at which energy is being released inside the engine. Everything the engine does for the aircraft has to come out of that number, and in practice well under half of it does.

Worked example — how much of the fuel's energy actually moves the aircraft

A large turbofan in the cruise burns 1,150 kg of fuel per hour while producing 22 kN of net thrust at a true airspeed of 240 m/s.

Fuel energy released: $\dot{m}_f = 1150/3600 = 0.319\ \text{kg/s}$, so $0.319 \times 43 = 13.7\ \text{MW}$.

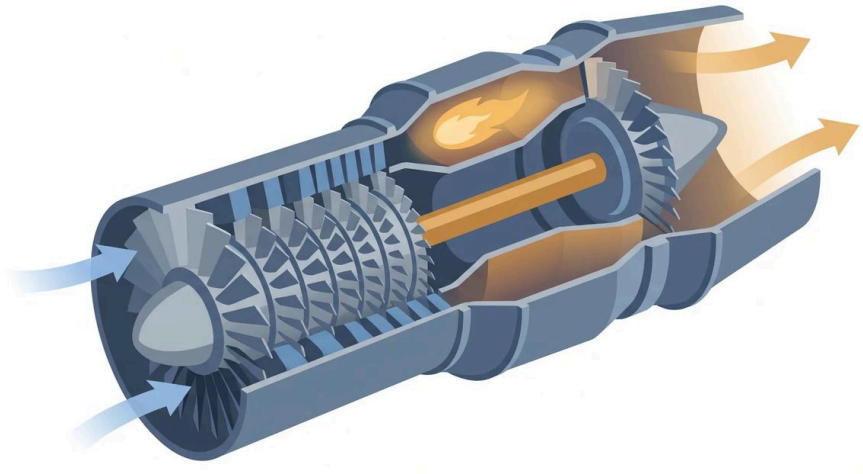
Useful propulsive power delivered: $P = F \times V = 22\,000 \times 240 = 5.28\ \text{MW}$.

Overall efficiency: $5.28/13.7 = 0.385$, i.e. about **38.5%**.

The other 61.5% leaves as hot exhaust, as kinetic energy still in the jet wake, and as friction, leakage, cooling-air and noise losses. An overall efficiency near 40% is excellent for a heat engine — but it means that for every tonne of fuel burned, roughly six hundred kilogrammes' worth of energy is thrown away as heat and turbulence.

Pressure energy: the reason the compressor exists

Pressure energy is the least intuitive of the stored forms, but it is the one the whole cycle is built around. Compressing air does not add any fuel to it, yet it makes the fuel that is added later far more valuable. The reason is that the work you can recover from a gas by expanding it depends on the *ratio* of the pressure you start from to the pressure you finish at. Burn fuel in air at atmospheric pressure and the hot gas has nowhere useful to expand to. Burn the same fuel in air that has already been squeezed to forty times atmospheric pressure and the gas can expand through a pressure ratio of forty on its way back out, doing work on a turbine the whole way down.



The compressor squeezes the air before combustion so the hot gas can expand back out through the turbine.

This is why the compressor absorbs an enormous share of the engine's internal power. On a turbojet the great majority of the work the turbines produce goes straight back into driving the compressor, and only what is left over leaves as jet thrust; on a high-bypass turbofan or a turboprop the turbines are sized to extract a large additional share of the gas power specifically to drive the fan or propeller, so that share is not a leftover at all. It is also why compressor damage is so serious: a compressor that has lost efficiency does not merely fail to compress, it consumes more turbine work to achieve less pressure, and the shortfall shows up as a higher gas temperature for the same thrust.

Why kinetic energy is expensive, and why that shapes every engine

Thrust and wasted energy do not scale the same way, and this single fact drives the entire evolution of the turbofan. Thrust is proportional to the change in velocity given to the air — it is a *linear* relationship. The kinetic energy left behind in the jet is proportional to the *square* of that velocity change. Doubling the velocity change doubles the thrust but quadruples the energy dumped into the wake.

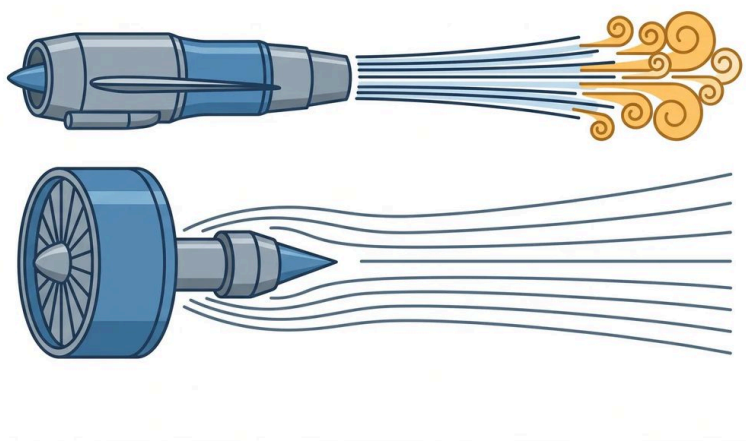
Thrust obtained (linear in velocity change):

$$F = \dot{m}\,\Delta V$$

Kinetic energy given to the air per second (quadratic in velocity change):

$$\dot{E}_K = \tfrac{1}{2}\,\dot{m}\,\Delta V^2$$

Because F rises with ΔV but $\dot{E}_K$ rises with ΔV^2, the cheapest way to buy a given thrust is always to take the **largest possible mass of air** and give it the **smallest possible velocity change**.



Same thrust two ways: a small mass of air sped up a lot, or a large mass of air sped up a little.

Worked example — same thrust, half the energy

Option A: accelerate 100 kg/s of air by 400 m/s.

Thrust $= 100 \times 400 = 40\,000\ \text{N} = 40\ \text{kN}$. Wake energy $= \tfrac{1}{2} \times 100 \times 400^2 = 8.0\ \text{MW}$.

Option B: accelerate 200 kg/s of air by 200 m/s.

Thrust $= 200 \times 200 = 40\,000\ \text{N} = 40\ \text{kN}$. Wake energy $= \tfrac{1}{2} \times 200 \times 200^2 = 4.0\ \text{MW}$.

Identical thrust, but Option B throws away exactly **half** the energy. It needs twice the airflow to do it, which is precisely why a modern engine has a fan two or three metres across bolted to the front of a comparatively small core.

Following the energy through a turbofan

It helps to think of the fuel's 43 MJ/kg as being spent in a fixed sequence, with a deduction at every step:

- **Combustion losses.** Modern combustion is very nearly complete, so only a small fraction of the fuel's chemical energy fails to be released as heat. This is the smallest of the deductions.
- **Turbine work taken back by the compressor.** The largest internal transfer in the engine. It is not a loss — the energy is stored back into the air as pressure — but it is unavailable for propulsion.
- **Turbine work taken by the fan or propeller.** On a high-bypass turbofan or a turboprop this is where most of the useful output goes.
- **Cooling and sealing air.** Air bled from the compressor to cool turbine blades and pressurise bearing chambers has been compressed at full cost but bypasses the combustor, so it contributes little or nothing to the cycle work. It is a deliberate, unavoidable penalty paid to allow a higher gas temperature.
- **Residual kinetic energy in the jet.** Any exhaust moving faster than the aircraft carries away kinetic energy that the aircraft never uses. This is the loss that propulsive efficiency measures.
- **Rejected heat in the exhaust.** By far the largest single deduction. It is inherent to the cycle rather than a defect: a heat engine must reject heat at the cold end, and in an open-cycle gas turbine the atmosphere is the cold end.
- **Friction, leakage, windage and noise.** Small individually, but they are the losses maintenance can influence — worn seals, opened tip clearances and eroded aerofoils all show up here.

Exam framing: a question that asks what "produces" thrust is testing whether you understand the engine as a converter. The engine does not manufacture a force out of nothing; it changes the momentum of a mass of air, and thrust is the reaction to that change. Every energy-related answer in this module should be traceable back to the chain of conversions above.

Bernoulli's Principle and Gas Laws

A gas turbine is, in engineering terms, a machine for trading three quantities against one another: pressure, velocity and temperature. Almost every component in the engine exists to make one of those trades in a controlled way. Two separate bodies of theory govern the trade. The **gas laws** describe the *state* of the air — how its pressure, volume and temperature are related when it is squeezed or heated. **Continuity and Bernoulli's principle** describe the *flow* — what happens to pressure and velocity when that air is pushed through a duct that changes shape. You need both to explain any part of the engine.

The gas laws

Law	Statement	Held constant	Where you meet it in the engine
Boyle's law	Pressure is inversely proportional to volume: $P \propto 1/V$	Temperature	The idealised squeeze of a gas — the starting point for understanding compression, though no compressor is isothermal in practice.

Charles's law	Volume is directly proportional to absolute temperature: $V \propto T$	Pressure	The combustion chamber. Heat is added at nominally constant pressure, so the gas must expand — and since it cannot expand sideways in a fixed duct, it expands by speeding up.
Pressure law (Gay-Lussac)	Pressure is directly proportional to absolute temperature: $P \propto T$	Volume	Any sealed volume that gets hot — a bearing chamber, a closed bleed duct, a tyre. Trap a fixed mass of air in a rigid vessel and heat it, and the pressure climbs in direct proportion to the absolute temperature, because no change in volume is available to relieve it. Apply it only where the volume genuinely is fixed: the temperature rise across a compressor is *not* this law, because there the volume is deliberately reduced and work is done on the gas.
Combined gas law	PV/T is constant for a fixed mass of gas	Mass of gas	Comparing conditions at two stations in the engine, e.g. compressor inlet versus compressor delivery.
Characteristic gas equation	$P = \rho R T$, with $R = 287\ \mathrm{J\,kg^{-1}K^{-1}}$ for dry air	—	The single most useful relation in engine performance: it converts a pressure and a temperature straight into a density, and density is what mass flow — and therefore thrust — depends on.

Density from pressure and temperature

$$\rho = \frac{P}{R\,T}$$

P in pascals, T in **kelvin** (never degrees Celsius), $R = 287\ \mathrm{J\,kg^{-1}K^{-1}}$ for dry air.

Absolute temperature: $T\ (\mathrm{K}) = t\ (^{\circ}\mathrm{C}) + 273$.

Worked example — why thrust falls on a hot day and at altitude

ISA sea level (15 °C, 101.3 kPa):

$\rho = 101\,300/(287 \times 288) = 1.225\ \mathrm{kg/m^3}$ — the standard sea-level density.

Same pressure, ambient temperature 40 °C (313 K):

$\rho = 101\,300/(287 \times 313) = 1.128\ \mathrm{kg/m^3}$ — about **8% less dense**.

ISA tropopause, 11,000 m (–56.5 °C = 216.7 K, 22.6 kPa):

$\rho = 22\,600/(287 \times 216.7) = 0.363\ \mathrm{kg/m^3}$ — about **30% of the sea-level value**.

The engine's compressor sweeps a fixed volume of air per revolution. If the density of that air falls by 8%, the mass flow at the same rotational speed falls by about 8% too, and since thrust is proportional to mass flow, so does the thrust. At the tropopause, the same engine at the same rpm handles only about a third of the mass flow it would swallow at sea level.

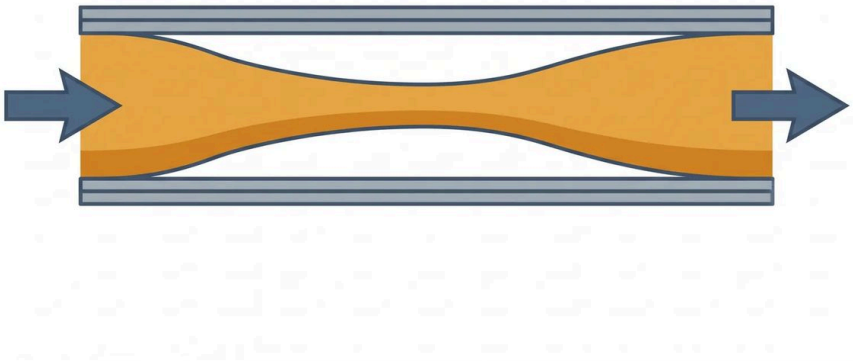
Continuity: the mass that goes in must come out

Continuity equation

$$\dot{m} = \rho A V$$

Mass flow = density × cross-sectional area × velocity. For steady flow through a duct, $\dot{m}$ is the same at every station.

Continuity is a constraint, not a choice: whatever mass of air enters the engine each second must leave it each second (plus the fuel added, minus any air bled off). That single constraint explains two pieces of engine geometry that otherwise look arbitrary. Through the compressor, pressure and temperature both climb, and pressure climbs faster than temperature, so density rises steadily. To keep the axial velocity roughly constant with a rising density, the area must shrink — which is why the compressor annulus tapers down from front to rear and the blades get progressively shorter. Through the turbine the reverse happens: the gas expands, its density falls, and the annulus must open out, which is why turbine blades get progressively longer towards the rear.



The gas path narrows through the compressor as density rises, then opens out through the turbine as it falls.

Bernoulli's principle and the shape of a duct

Bernoulli's principle (incompressible form)

$$P_s + \tfrac{1}{2}\rho V^2 = P_t = \text{constant}$$

Static pressure + dynamic pressure = total (pitot) pressure. In an ideal, loss-free duct the total pressure stays constant, so any gain in one term must be paid for out of the other.

The practical consequence is the rule every engine duct obeys. In **subsonic** flow, a duct whose area increases in the direction of flow slows the air down and its static pressure rises — that is a *diffuser*. A duct whose area decreases speeds the air up and its static pressure falls — that is a *nozzle*. Above Mach 1 the

behaviour reverses, because the density changes faster than the velocity and the density term in the continuity equation takes over.

Duct shape	Subsonic flow	Supersonic flow	Engine application
Convergent (area reducing)	Velocity **rises**, static pressure **falls**	Velocity falls, static pressure rises	The propelling nozzle of a subsonic engine; the passages between turbine nozzle guide vanes.
Divergent (area increasing)	Velocity **falls**, static pressure **rises**	Velocity rises, static pressure falls	The flight intake of a subsonic engine; compressor stator passages; the pre-diffuser ahead of the combustor; the exhaust unit downstream of the turbine.
Convergent-divergent	Accelerates the flow to Mach 1 at the throat	Continues to accelerate the flow supersonically in the divergent portion	Propelling nozzles of supersonic aircraft; supersonic intakes run the same geometry in reverse to decelerate the incoming flow.

Applying this to the intake of a subsonic turbofan gives the answer the exam expects. The intake duct is **divergent**: it slows the incoming air and converts the lost velocity into a rise in static pressure at the fan face. That free "ram" compression means the compressor starts its work from a higher pressure than ambient, which raises the effective overall pressure ratio and costs nothing in fuel. A convergent intake would do the opposite — accelerate the air and lower its static pressure — which is exactly what is not wanted ahead of a compressor.

The intake behaves differently on the ground. The divergent shape is designed for flight. With the aircraft stationary and the engine running, the air is not arriving with any velocity of its own — the engine has to draw it in, so the flow *accelerates* from still air towards the fan face and the static pressure there sits *below* ambient. Both statements are true; they describe different operating conditions. The low static pressure at the intake when static is why a running engine on the ramp can generate a ground vortex and lift debris into itself, and it is the reason for intake hazard areas, vortex dissipators and a thorough FOD walk before start.



On the ground, low static pressure at the intake can form a vortex that lifts debris into the engine.

Choking and the critical pressure ratio

A convergent duct cannot accelerate a gas past the local speed of sound at its exit. Once the pressure ratio across the duct reaches a value known as the **critical pressure ratio**, the exit velocity sticks at Mach 1 and the duct is said to be *choked*. For air at ordinary temperatures the critical pressure ratio is about 1.9:1; for hot exhaust gas, whose ratio of specific heats is a little lower, the critical value is slightly lower still, around 1.85:1. Modern engines operate well above this ratio at high power, so the propelling nozzle is normally choked in the take-off and climb regime.

Choking is not a fault, and it matters directly to the thrust equation. Once the nozzle is choked, any further increase in engine pressure ratio cannot show up as extra exit velocity — the exit velocity is pinned at the local speed of sound. Instead, the surplus appears as a static pressure at the nozzle exit that is *higher than ambient*. That excess pressure acts over the nozzle exit area and produces a genuine forward force in its own right. This is the origin of the pressure term in the full thrust equation, and it is why a propelling nozzle is correctly described as producing **both** momentum (velocity) thrust and pressure thrust rather than one alone. On a subsonic engine the pressure term is the smaller of the two, but it is not negligible at high power settings.

Compression is diffusion; expansion is acceleration

The gas laws and Bernoulli together explain what each bladed component is actually doing to the air, and getting the directions right here is the foundation for the whole of the rest of this note.

- **A compressor rotor** does work on the air. It raises the air's total energy, so both its velocity and its static pressure rise as it passes through the rotating row.
- **A compressor stator** does no work at all. It is a divergent passage, so it converts the extra velocity the rotor produced into a further rise in static pressure, and it straightens the flow ready for the next rotor. Through a compressor as a whole, **pressure and temperature both rise**.
- **A turbine nozzle guide vane** does no work either. It is a convergent passage, so the gas **accelerates and its static pressure falls** as it is turned onto the rotor blades at the correct angle.
- **A turbine rotor** extracts work from the gas. It removes energy, so through a turbine as a whole, **pressure and temperature both fall**.

The single most important asymmetry in the engine. A compressor is trying to push air from a low pressure to a high pressure — it works against an *adverse* pressure gradient, which is aerodynamically difficult and inherently unstable. That is why one axial compressor stage can only manage a pressure ratio of roughly 1.1:1 to 1.3:1, why a high overall pressure ratio needs a great many stages, and why compressors are vulnerable to stall and surge.

A turbine is doing the opposite: expanding gas from high pressure to low pressure, working *with* a favourable pressure gradient. That is aerodynamically easy, the flow stays attached, and a single turbine stage can extract enough work to drive many compressor stages. Count the stages on any engine cutaway and you will see the consequence — a dozen or more compressor stages driven by one or two turbine stages on the same shaft.

Force, Work, Power and Torque

These four quantities are routinely used loosely in conversation and precisely in examinations, and the difference between a jet engine and a shaft engine is at heart a difference in which of them the engine is designed to deliver. A turbojet or turbofan is rated in **force**. A turboprop or turboshaft is rated in **power**, which is force multiplied by speed. Getting the distinction straight explains why a jet engine's rating does not change with airspeed while its usefulness does, and why a propeller engine always has a gearbox.

Force and momentum

Newton's second law, in its two useful forms

$$F = m\,a \qquad \text{and} \qquad F = \frac{\Delta(mV)}{\Delta t} = \dot{m}\,\Delta V$$

The second form is the one that matters for propulsion: a force is a *rate of change of momentum*. The engine never accelerates a single fixed lump of air; it accelerates a continuous stream, so the mass term becomes a mass flow rate.

It is worth doing the dimensional check once, because it removes any doubt that $\dot{m}\Delta V$ really is a force. Mass flow is measured in kilogrammes per second and velocity change in metres per second, so the product has units of $\mathrm{kg}\cdot\mathrm{m/s^2}$ — and one newton is defined as exactly the force that accelerates one kilogramme at one metre per second squared. The units are identical. An engine handling 200 kg/s of air and adding 160 m/s to its velocity is therefore producing 32,000 N, or 32 kN, of thrust.

In non-SI units, one pound of thrust (lbf) is 4.448 N, so 1 kN is about 224.8 lbf. Engine data plates, thrust ratings and flight manuals still use both, and an exam question may quote either.

Work and power

Work is force multiplied by the distance moved *in the direction of the force*:

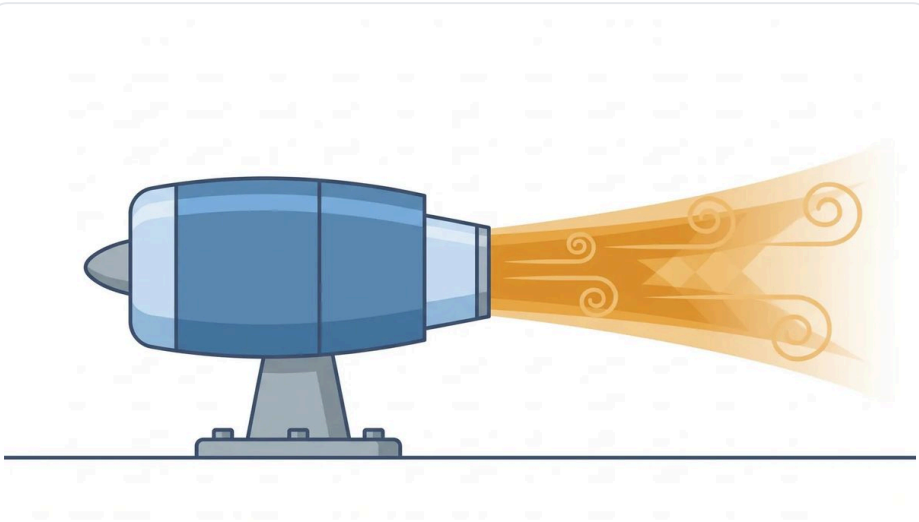
$$W = F \times d \qquad [\text{joules}]$$

Power is the rate of doing work — work divided by time, which for a steady force is force multiplied by velocity:

$$P = \frac{W}{t} = F \times V \qquad [\text{watts}]$$

1 horsepower = 745.7 W; 1 kW = 1.341 hp.

The definition of work carries a consequence that is easy to state and easy to forget: **if nothing moves, no work is done**. An engine held on a test bed at full take-off thrust is exerting an enormous force, but the distance moved is zero, so the work done on the airframe is zero and the propulsive power delivered is zero. Every joule of the fuel's energy is going into the jet wake and into heat. This is not a quirk of test benches — it is the reason propulsive efficiency at zero airspeed is exactly zero, and the reason a jet transport is at its least fuel-efficient during the take-off roll and at its most efficient in the cruise.



On a test bed at full thrust the engine pushes hard but moves nowhere, so it does no work on its mounting.

Worked example — the same thrust, different power

An engine holding a steady 22 kN of thrust:

At 240 m/s (about Mach 0.8 at cruise altitude): $P = 22\,000 \times 240 = 5.28\ \text{MW} \approx 7\,080\ \text{hp}$.

At 120 m/s (a climb speed): $P = 22\,000 \times 120 = 2.64\ \text{MW} \approx 3\,540\ \text{hp}$.

At 0 m/s (brakes on, engine at the same setting): $P = 22\,000 \times 0 = 0\ \text{W}$.

The engine's *rating* — 22 kN — has not changed at all. What has changed is how much useful work that force is doing per second. This is the clearest single illustration of why jet engines are rated in thrust and propeller engines in power: for a propeller the useful output is already a rate of work at the shaft, whereas a jet's useful output only becomes power once the aircraft is moving.

Torque and shaft power

Torque is a turning moment: a force applied at a perpendicular distance from an axis.

$$T = F \times r \qquad [\text{newton metres}]$$

Shaft power is torque multiplied by angular velocity:

$$P = T \times \omega = 2\pi N T$$

where ω is in radians per second and N is in revolutions per second.

Worked example — turboprop shaft power

A propeller shaft delivers 3,000 N·m of torque at 1,200 rev/min.

$\omega = 2\pi \times 1200/60 = 125.7 \text{ rad/s}$

$P = 3000 \times 125.7 = 377\,000 \text{ W} = 377 \text{ kW} \approx 506 \text{ shp}$

Notice that a torque reading alone tells you nothing about power. The same 3,000 N·m at half the propeller speed would be half the power. This is why a turboprop's torque gauge is only a valid power indication when the propeller is being governed at a constant speed — which is exactly how a constant-speed propeller is operated in flight.

Because power is the product of torque and speed, a reduction gearbox that reduces speed must increase torque in the same proportion, less the small loss in the gear train. A free power turbine turning at, say, 20,000 rev/min driving a helicopter rotor at 400 rev/min is a 50:1 reduction, so the torque at the rotor is roughly fifty times the torque at the power turbine. That is why the main rotor gearbox of a helicopter, or the propeller reduction gearbox of a turboprop, is such a large, heavy and highly stressed component — it is handling a torque an order of magnitude or two greater than anything inside the engine itself.

The reason the gearbox is unavoidable is a mismatch of optimum speeds. A turbine extracts energy most efficiently when its blade speed bears a particular relationship to the gas velocity passing through it, and that means running fast. A propeller, by contrast, is limited by the speed of its blade tips: once the helical tip speed approaches the speed of sound the blade suffers compressibility losses and generates a great deal of noise, so the propeller must run slowly. The two requirements cannot both be satisfied on one shaft, so a gearbox is placed between them. Exactly the same logic, applied to a very large fan rather than a propeller, is the argument for a geared turbofan.

How each engine type is rated

Engine type	Rated output	Typical units	Why
Turbojet / turbofan	Thrust	kN or lbf	The whole output is a force applied to the airframe; it is independent of airspeed, so quoting a force is meaningful at any condition.
Turboprop	Shaft horsepower (SHP) and equivalent shaft horsepower (ESHP)	kW or shp	Most of the output is shaft work delivered to the propeller, but a small residual jet thrust remains, and ESHP accounts for both.
Turboshaft	Shaft power	kW or shp	Essentially the entire output is shaft work; residual jet thrust is deliberately minimised and is not credited.

Equivalent shaft horsepower is the figure that expresses a turboprop's *total* power output. The engine delivers most of its power through the shaft, but the exhaust still leaves with enough velocity to contribute a modest jet thrust, and that thrust does useful work whenever the aircraft is moving. ESHP converts that residual thrust into an equivalent number of horsepower and adds it to the measured shaft horsepower. In flight the conversion depends on airspeed and propeller efficiency, since the jet thrust's contribution grows as

the aircraft goes faster. Under static sea-level conditions, where a speed-based conversion cannot be used, the long-established convention is to treat roughly 2.5 lb of residual jet thrust as equivalent to one shaft horsepower. Either way, SHP alone always understates a turboprop's true output; ESHP is the total.

Measuring torque, and what the limits mean

Shaft engines carry a torquemeter — on older engines a hydraulic device in which helical gear reaction is balanced against oil pressure, on modern engines usually an electronic sensor measuring the small angle of twist along a length of the output shaft. The cockpit gauge may read in newton metres, in pound-feet, or as a percentage of a reference torque. Because power is torque times speed, and the propeller or rotor is normally governed at a constant speed, a torque limit at that constant speed *is* a power limit, and that is how it should be read.

A turboshaft or turboprop typically has more than one limit that can bind first: torque, gas generator speed, and turbine gas temperature. Which one is reached first depends on the conditions. On a cold day at low altitude the engine will normally reach its torque (structural) limit before its temperature limit, because dense air lets it produce a great deal of shaft power at a modest gas temperature. On a hot, high day the temperature limit is reached first and the engine cannot reach its rated torque at all. The crew must respect whichever arrives first, and the airframe flight manual will say so.

Torque and energy share a unit but are not the same quantity. A newton metre of torque and a joule of work are dimensionally identical, and this catches people out. The difference is geometric: work is force multiplied by distance moved *along the line of the force*, while torque is force multiplied by the perpendicular distance *from an axis*, with nothing necessarily moving at all. A torque wrench held stationary at 50 N·m has done no work. Only when the fastener turns does the torque begin producing work.

The Brayton Cycle

The gas turbine cycle consists of four continuous processes (unlike the intermittent piston engine cycle):

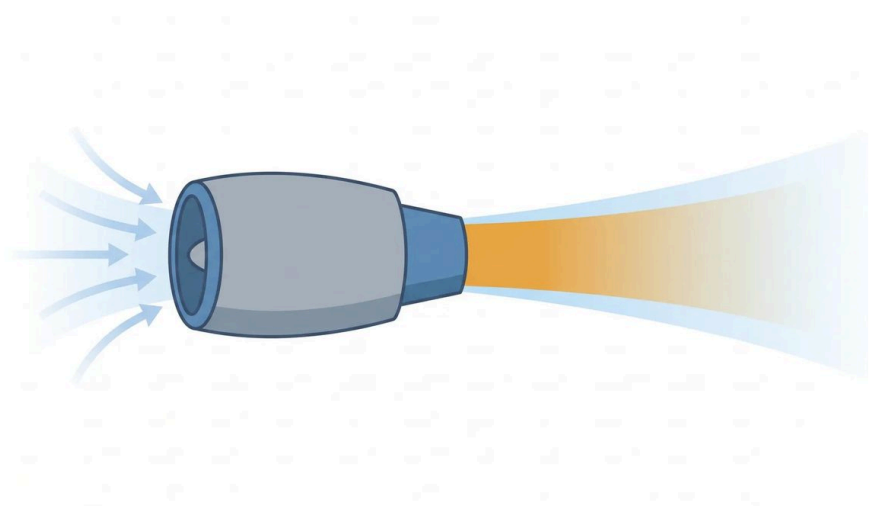
Process	Section	What Happens	Thermodynamic Change
1. Intake	Air intake / diffuser	Air is drawn in and directed into the compressor. At high speed, the intake acts as a diffuser (slows air, increases pressure).	Slight pressure rise, slight temperature rise
2. Compression	Compressor	Air is compressed to 10:1 – 50:1 pressure ratio (depending on engine type). Temperature rises significantly.	Large pressure rise, large temperature rise (adiabatic compression)
3. Combustion	Combustion chamber	Fuel is injected and burned at **constant pressure**. Temperature rises dramatically (up to ~2,000°C). Only ~25% of the air is used for combustion; the rest cools the flame and dilutes the exhaust to a temperature the turbine can withstand.	Constant pressure, very large temperature rise

4. Expansion	Turbine + exhaust nozzle	Hot, high-pressure gas expands through the turbine (which extracts energy to drive the compressor) and then through the exhaust nozzle (which converts remaining pressure into kinetic energy = thrust).	Pressure falls, temperature falls, velocity increases

The four points of the cycle

Thermodynamics texts number the cycle by the four states the air passes through rather than by the four processes, and it is worth being able to move between the two descriptions. Station 1 is ambient air at the intake. Station 2 is compressor delivery: pressure has risen by the overall pressure ratio and temperature has risen with it. Station 3 is combustor exit — turbine entry — where the fuel's heat has been added at nominally constant pressure, so the temperature has risen enormously while the pressure has not. Station 4 is the end of expansion, where the gas has given up its work to the turbine and the nozzle and is discharged to atmosphere at close to ambient pressure but still well above ambient temperature.

Because the engine breathes atmospheric air and discharges to atmosphere, this is an **open** cycle: the fourth process, heat rejection from state 4 back to state 1, does not happen inside any component. It happens in the sky behind the aircraft, where the exhaust plume mixes with the surrounding air and cools. That heat rejection is nonetheless a real and unavoidable part of the cycle, and it is by far the largest single energy loss in the engine.



The cycle is open: hot exhaust gives up its heat by mixing with the surrounding air behind the engine.

The working cycle as the examination states it

The four thermodynamic processes above describe what happens to the gas. The engine's **working cycle** is more often listed as five phases, because it names the discharge of the gas separately from the expansion that produced the work: **induction, compression, combustion, expansion, exhaust**. Both descriptions are of the same engine, and the order is the point being tested. Combustion must follow compression, and expansion must follow combustion — any sequence that places expansion before combustion, or combustion before compression, describes something that is not a gas turbine.

Continuous, not intermittent

The single most important structural difference between a gas turbine and a piston engine is that the gas turbine's cycle is **continuous**. It is not completed in one revolution of the shaft, nor in two. All five phases are happening at the same instant, simultaneously and permanently, in five different parts of the engine: air is being drawn into the intake at the same moment that other air is being compressed, other air is burning, other gas is expanding through the turbine and other gas is leaving the nozzle. Nothing waits its turn.

Feature	Piston engine (Otto cycle)	Gas turbine (Brayton cycle)
Heat addition	At approximately constant *volume*, in a closed cylinder	At approximately constant *pressure*, in a continuously open chamber
Timing of the phases	Sequential — each phase occupies the same cylinder in turn	Simultaneous — each phase occupies its own dedicated section of the engine, permanently
Cycle completion	One cycle per two crankshaft revolutions (four-stroke)	Continuous; the cycle has no beginning or end tied to shaft position
Power delivery	Intermittent power impulses, requiring a flywheel and producing torsional vibration	Smooth and continuous; no reciprocating masses to balance
Airflow handled	Limited by cylinder volume and breathing; modest mass flow	Very large mass flow for the frontal area, because the flow never stops
Component duty	Each part sees the full temperature swing of the cycle, but only briefly each cycle	Each part sees *one* condition, but sees it continuously — which is why turbine materials and cooling dominate the design

That last row is the price of continuity and it is the reason a gas turbine is a materials problem more than a mechanical one. A piston crown is exposed to peak combustion temperature for a few milliseconds per cycle and is cooled by the incoming charge for the rest of the time. A turbine nozzle guide vane sits in gas at turbine entry temperature every second the engine is running, with no relief.

Ideal thermal efficiency and why pressure ratio matters

Air-standard thermal efficiency of the Brayton cycle

$$\eta_{th} = 1 - \frac{1}{r^{(\gamma-1)/\gamma}}$$

where r is the overall pressure ratio and γ is the ratio of specific heats (about 1.4 for air).

The exponent $(\gamma - 1)/\gamma$ works out at approximately 0.286 for air.

Worked example — the payoff from a higher pressure ratio

An early engine with an overall pressure ratio of 10:1:

$10^{0.286} = 1.93$, so $\eta_{th} = 1 - 1/1.93 = 0.48$, i.e. about **48%**.

A modern engine with an overall pressure ratio of 40:1:

$40^{0.286} = 2.87$, so $\eta_{th} = 1 - 1/2.87 = 0.65$, i.e. about **65%**.

Quadrupling the pressure ratio lifts the ideal thermal efficiency by seventeen percentage points. This is the single largest reason overall pressure ratios have climbed steadily throughout the history of the gas turbine, and the reason so much design effort goes into squeezing more pressure out of fewer, better compressor stages.

These are *ideal* figures for a loss-free cycle in a perfect gas. A real engine falls well short of them, because compressor and turbine efficiencies are below 100%, because a few percent of pressure is lost in the ducts and combustor, and because a slice of compressed air is diverted to cool the turbine instead of passing through the combustor. And this is the efficiency of the *core* alone — the propulsive stage that turns core power into thrust applies its own efficiency on top, which is why an overall efficiency near 40% is a good result even from a core whose ideal figure is 65%.

Pressure ratio and turbine entry temperature pull in different directions

Two design parameters dominate gas turbine performance, and they do different jobs.

- **Overall pressure ratio (OPR)** governs *thermal efficiency* — how much of the fuel's heat becomes work. Raising it makes the engine burn less fuel for the same work.
- **Turbine entry temperature (TET)** governs *specific work* — how much work is obtained from each kilogramme of air. Raising it lets a smaller, lighter engine produce a given thrust.

For a fixed turbine entry temperature, increasing the pressure ratio raises the thermal efficiency but eventually begins to *reduce* the specific work, because the compressor delivery temperature climbs closer to the turbine entry temperature and the temperature rise available across the combustor shrinks. There is therefore a pressure ratio that maximises work per kilogramme of air and a different, higher one that maximises efficiency. Engines built for range are pushed towards the efficiency end; engines built for thrust in a small package are pushed towards the specific-work end. Both parameters are ultimately capped by the same thing — the temperature the turbine materials and their cooling can survive.

"Constant pressure" describes the ideal cycle. In a real combustion chamber a small total-pressure loss, typically of the order of a few percent, is unavoidable: the air has to be turned, slowed in the pre-diffuser, forced through the flame tube holes and mixed, and every one of those actions costs total pressure. Designers work hard to keep the loss low because every percent of pressure lost there is a percent of pressure the compressor produced at full cost in turbine work. The thermodynamic label "constant pressure" distinguishes this cycle from the constant-*volume* heat addition of a piston engine; it is not a claim that a real chamber loses nothing.

Where the turbine's work goes

An important consequence of the cycle is often missed: the turbine does not exist to produce thrust. It exists to drive the compressor, and in a turbofan or turboprop, to drive the fan or propeller as well. In a pure turbojet the turbine extracts only as much work as the compressor and the accessories need, and everything still left in the

gas after that is deliberately allowed to leave through the nozzle as velocity. In a high-bypass turbofan the low-pressure turbine has to drive a fan producing most of the aircraft's thrust, so a much larger share of the gas energy is taken out as shaft work and correspondingly less of it leaves as jet velocity. That is the same trade, expressed in cycle terms, that the bypass ratio section below expresses in propulsive terms.

Newton's Laws Applied to Jet Propulsion

Law	Application
1st Law (Inertia)	Air at rest stays at rest until acted upon by the compressor/combustor. The aircraft in flight continues at constant velocity unless a net force (thrust or drag) acts on it.
2nd Law (F = ma)	Thrust = mass flow rate × change in velocity. $F = \dot{m}(V_j - V_0)$ where $\dot{m}$ = air mass flow, V_j = jet velocity, V_0 = intake velocity.
3rd Law (Action/Reaction)	The engine accelerates air rearward (action); the reaction force pushes the engine (and aircraft) forward. This is the fundamental principle of jet propulsion.

Thrust equation:

$$F = \dot{m}(V_j - V_0) + A_e(P_e - P_0)$$

Where the second term accounts for any pressure imbalance at the nozzle exit. For most cases, this simplifies to:

$$F \approx \dot{m} \times \Delta V$$

To produce more thrust: increase mass flow ($\dot{m}$) and/or increase the velocity change (ΔV).

Gross thrust, ram drag and net thrust

The thrust equation above is written for the general case, but in service you will meet three separate thrust quantities and they must not be confused. The distinction only matters once the aircraft is moving — which is exactly why it is so often got wrong.

Gross (or momentum) thrust — the total forward force generated at the nozzle, ignoring how the air got in:

$$F_g = \dot{m}\,V_j + A_e(P_e - P_0)$$

Ram drag (intake momentum drag) — the rearward force needed to capture the incoming air and bring it to the engine's own speed:

$$D_{ram} = \dot{m}\,V_0$$

Net thrust — what the airframe actually feels:

$$F_n = F_g - D_{ram}$$

On the ground with the aircraft stationary, $V_0 = 0$, so the ram drag is zero and **net thrust equals gross thrust**. Static thrust is the highest net thrust the engine will produce at that power setting.

Worked example — what forward speed does to net thrust

Take an engine handling 200 kg/s with a jet velocity of 400 m/s. To isolate the effect of ram drag alone, hold the mass flow and jet velocity constant for a moment:

At $V_0 = 240$ *m/s:* gross $= 200 \times 400 = 80\ \text{kN}$; ram drag $= 200 \times 240 = 48\ \text{kN}$; **net = 32 kN**.

At $V_0 = 300$ *m/s:* gross $= 80\ \text{kN}$ still; ram drag $= 200 \times 300 = 60\ \text{kN}$; **net = 20 kN**.

Sixty metres per second of extra airspeed has removed more than a third of the net thrust, for no change in fuel flow or rpm. The direction is unambiguous: **at a constant engine speed, increasing forward speed reduces net thrust**, because ram drag grows in direct proportion to airspeed.

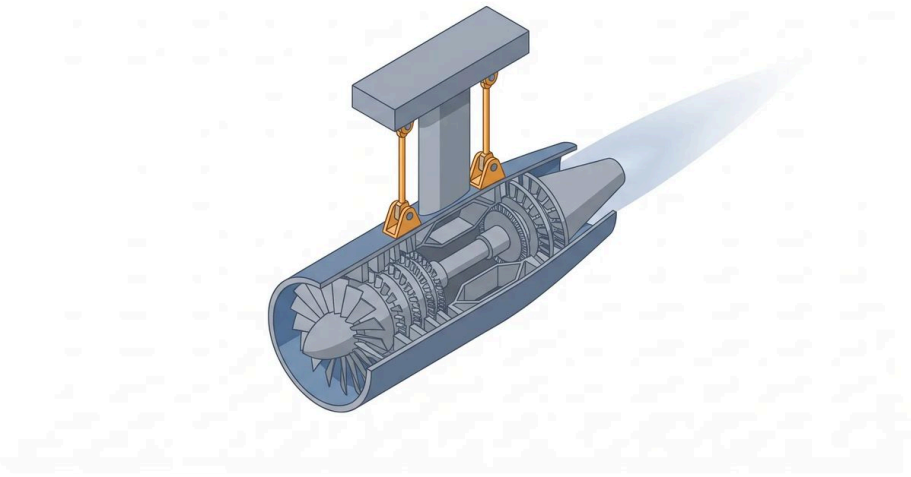
That worked case deliberately froze the mass flow, so it shows only one of the two competing effects. The second effect works the other way. As the aircraft speeds up, the intake recovers more of the free-stream dynamic pressure as static pressure at the compressor face — *ram compression*. That raises the pressure and density of the air entering the engine, which raises the mass flow the engine swallows and raises the jet velocity it produces. Gross thrust therefore does not stay fixed; it climbs.

Which effect wins depends on the speed. Through the low and middle subsonic range the ram drag term dominates and net thrust falls with increasing airspeed — a well-known characteristic of the jet engine, and the reason a take-off thrust figure is quoted as a static value. At higher Mach numbers the ram compression term grows rapidly (it depends on the square of the flight speed) and eventually overtakes the ram drag, so net thrust flattens out and then climbs again. This recovery is what makes sustained supersonic flight possible at all, and it is why an engine designed for very high speed is given a high jet velocity: only an engine whose jet is much faster than the aircraft still has a useful velocity difference left to work with when the aircraft itself is travelling at 800 mph.

Two thrust figures that are both correct. Altitude and airspeed act on thrust through different mechanisms. Increasing *altitude* reduces air density, so mass flow and therefore thrust fall. Increasing *airspeed* increases ram drag, which reduces net thrust, while simultaneously increasing ram compression, which increases gross thrust. A cruising aircraft is experiencing both at once. The reason it still works out well is that the aircraft needs far less thrust in the cruise than for take-off, and the specific fuel consumption is better in the cold, thin air of the upper atmosphere.

Where the thrust force actually acts

A very common misconception is that a jet engine pushes against the air behind it. It does not, and the third-law statement in the table above says why: the reaction to accelerating the gas rearwards acts on *the engine's own internal surfaces*, not on the atmosphere outside. Add up the pressure and momentum forces on every internal component and the result is a net forward force, which is transmitted to the airframe through the engine mounts and thrust links. In a turbojet the compressor casing, the diffuser and the combustor casing carry large forward loads, while the turbine and the exhaust system carry rearward loads; the algebraic sum of the two is the net thrust.



The forward force builds up inside the engine and reaches the airframe through the mounts and thrust links.

Do not answer that a jet works by "impinging on the surrounding air". A jet engine would work perfectly well in a vacuum if it carried its own oxidiser — which is exactly what a rocket is. The propulsive force is the reaction to the rearward-moving mass of gas, and it exists whether or not there is anything outside for the exhaust to strike. This is a favourite examination distractor precisely because the wrong answer sounds plausible.

Mass flow, and why it is a mass and not a volume

Mass flow is the quantity of air passing through the engine per unit time, expressed in kilogrammes per second: total mass divided by time. It is deliberately expressed as a mass rather than a volume, because the volume occupied by a given mass of air changes drastically with pressure and temperature. A cubic metre of air at sea level on a cold day contains far more molecules — and therefore far more oxygen to burn and far more momentum to accelerate — than a cubic metre of air at 11,000 m. Only the mass figure means the same thing at every station in the engine and at every point in the flight envelope.

Worked example — specific thrust

Specific thrust is the thrust obtained per unit of airflow, $F_n/\dot{m}$, and it is a compact way of describing what kind of engine you are looking at.

A high-bypass turbofan producing 100 kN while handling 350 kg/s: $100\,000/350 = 286\ \text{N per kg/s}$.

A turbojet producing the same 100 kN from 170 kg/s: $100\,000/170 = 588\ \text{N per kg/s}$.

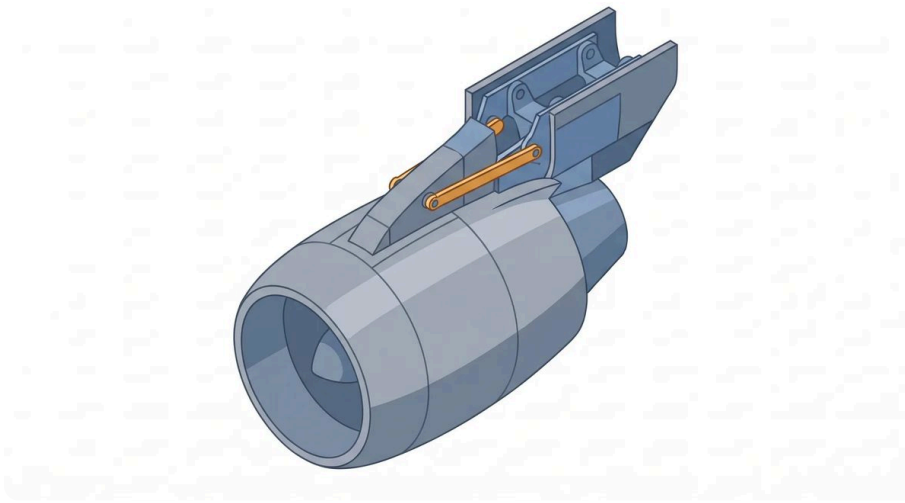
The turbojet does the same job with less than half the airflow, so it is physically much smaller in frontal area — but it does so by giving that smaller mass a far larger velocity change, which is expensive in energy. High specific thrust buys a compact engine at the cost of fuel consumption; low specific thrust buys fuel economy at the cost of size, weight and nacelle drag.

The first and third laws in the installation

Newton's first law is not only an abstraction about air at rest. The rotating assembly of a large engine has enormous inertia, and it takes several seconds for it to accelerate from idle to take-off power. That spool-up

time is a direct consequence of the first law, and it is why a stabilised approach requires the thrust to be set early, and why go-around performance is calculated from a defined approach thrust setting rather than from idle. Those same rotating masses also behave as large gyroscopes: they resist any change to their plane of rotation, and the resulting moments are among the loads the engine mounts and pylon are designed to carry.

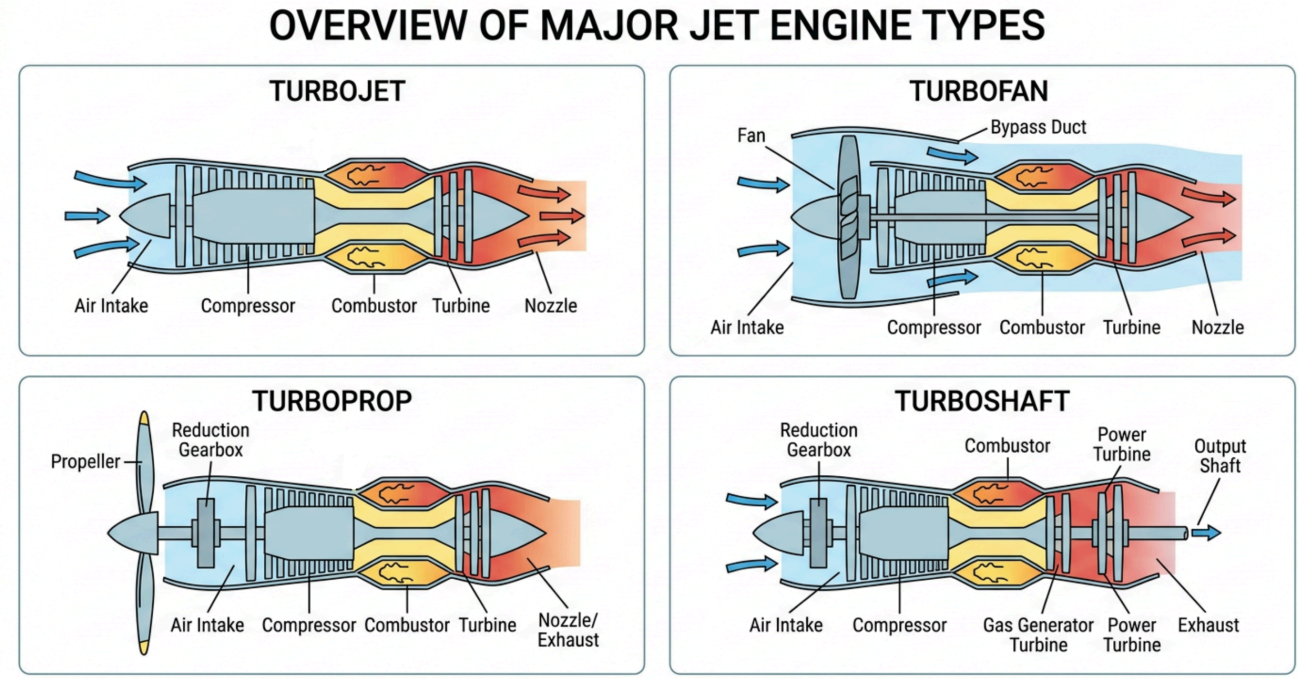
The third law governs the installation as much as the gas path. The reaction that drives the aircraft forward has to be carried into the structure somewhere, which is the job of the thrust links or thrust struts in the mount system. Their condition is an airworthiness matter, and the mount arrangement also has to accommodate the engine's thermal growth without transmitting distorting loads into the casings — an engine casing pulled out of round by its mounts opens turbine tip clearances and costs performance.



Thrust links carry the engine's forward reaction force from the casing into the pylon and airframe structure.

Engine Types

All four families of gas turbine share one thing: an identical **gas generator** at the heart of the engine, made up of a compressor, a combustion chamber and a turbine to drive that compressor. What distinguishes a turbojet from a turbofan, a turboprop or a turboshaft is nothing to do with how the gas is produced. It is entirely a question of what is done with the energy that remains in the gas *after* the compressor has been driven. The diagram below shows the arrangements side by side; the sections that follow explain each one and, more importantly, why each suits the flight regime it does.



The common core: the gas generator

The gas generator — compressor, combustor and the turbine stage or stages that drive that compressor — is a machine whose only product is a stream of hot, high-pressure gas. On its own it produces no useful output at all; every joule the turbine extracts is fed straight back into the compressor. Everything downstream of the gas generator is the *power section*, and it exists to convert what is left in the gas into whatever form the aircraft needs.

Type	What happens to the leftover gas energy	Propulsive device	Typical application	Rated in
Turbojet	All of it is expanded through the propelling nozzle as velocity	The jet itself	High-speed and supersonic aircraft; missiles; legacy transports	Thrust
Turbofan	Most is extracted by extra turbine stages to drive a ducted fan; the remainder leaves as core jet velocity	Ducted fan, plus a residual core jet	Commercial transports; business jets; modern combat aircraft (low bypass)	Thrust
Turboprop	Nearly all is extracted as shaft work through a reduction gearbox; a small residual jet remains	Propeller	Regional airliners, utility and military transports, trainers	SHP and ESHP
Turboshaft	Essentially all of it is extracted as shaft work; residual jet thrust is deliberately minimised	Rotor, or any external driven load	Helicopters; auxiliary power units; ground and marine power	Shaft power

Turbojet

In a turbojet every kilogramme of air taken in passes through the core, and the entire propulsive output leaves through the propelling nozzle. Since the whole airflow is being given a large velocity increase, the jet velocity

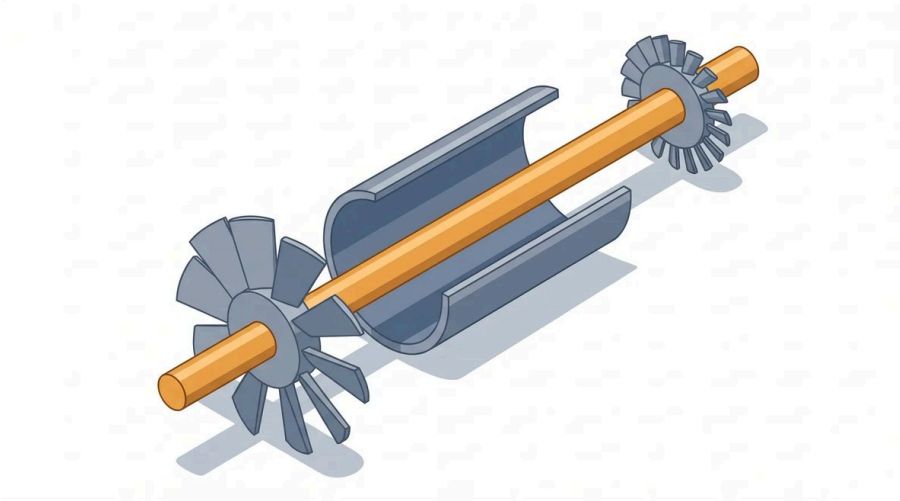
is very high and the specific thrust — thrust per unit of airflow — is the highest of the four types. That makes the engine compact in frontal area for the thrust it produces, which matters greatly on an aircraft designed for high speed, where nacelle drag is punishing.

The high jet velocity is also the turbojet's weakness at low speed and its strength at high speed. At subsonic cruise the exhaust is travelling far faster than the aircraft, so a great deal of kinetic energy is thrown away and the specific fuel consumption is poor. It also makes the engine extremely noisy: jet mixing noise rises with roughly the eighth power of jet velocity, so a modest reduction in exhaust speed produces a dramatic reduction in noise. But at very high flight speeds — approaching and beyond 800 mph — the aircraft's own velocity is approaching the exhaust velocity of a low-jet-speed engine, and only an engine whose jet is genuinely fast still has a useful velocity difference to exploit. That is why the turbojet remains the answer for the highest speed regimes, usually with reheat, while it has been displaced everywhere else.

Turbojets may be single-spool (one compressor, one turbine, one shaft) or twin-spool. Adding reheat — burning additional fuel in the jet pipe downstream of the turbine, where there is still plenty of unused oxygen — gives a large increase in thrust for a very large increase in fuel consumption, and requires a variable-area propelling nozzle so the extra volume of gas can be passed without changing the working conditions of the turbine.

Turbofan

A turbofan adds a large-diameter fan ahead of the core. Air leaving the fan divides: part enters the core compressor, and the rest — the bypass air — passes through the fan duct and straight out. The fan is mounted on the **low-pressure shaft** and is driven by the **low-pressure turbine** at the rear of the engine, through a shaft running concentrically inside the high-pressure shaft. This is a mechanical drive through metal, not an aerodynamic effect: the fan is not turned by air passing over it, and it is not turned by combustion gases striking it, and it is not driven by the accessory gearbox. The gases expand through the low-pressure turbine, that turbine turns the shaft, and the shaft turns the fan.



The fan's shaft runs inside the hollow high-pressure shaft, linking the low-pressure turbine to the fan.

On a modern high-bypass engine the fan is doing most of the propulsive work — commonly something like three-quarters of the total thrust comes from the bypass stream, with the core jet providing the remainder. The design point of the whole engine follows from that: the low-pressure turbine has to be large enough and have enough stages to absorb the power the fan needs.

Two exhaust arrangements are in common use. In a **separate-jet** (short duct) installation the bypass air discharges from its own nozzle around the outside of the core, and the two streams mix in the open air. In a **mixed-exhaust** (long duct) installation the two streams are brought together inside the nacelle, usually through a lobed forced mixer, before leaving through a common nozzle; that recovers a small performance benefit and reduces noise, at the cost of a heavier duct. A **geared** turbofan places a reduction gearbox between the low-pressure turbine and the fan, so that the fan can turn slowly — as a large fan must, to keep its tip speed acceptable — while the low-pressure turbine turns several times faster, where it is efficient and needs fewer stages.

Turboprop

A turboprop uses the propeller as its propulsive device and the gas turbine purely as a source of shaft power. Because a propeller must turn far more slowly than any turbine, a reduction gearbox is always fitted, typically reducing the speed by a large factor and multiplying the torque correspondingly.

The energy split is heavily weighted towards the shaft: as a general figure, on the order of nine-tenths of the available gas energy is extracted by the turbines and delivered to the propeller, leaving roughly a tenth to leave the exhaust as a modest jet reaction. That residual jet is small but real, it does useful work whenever the aircraft is moving, and it is the reason a turboprop's total output is quoted as equivalent shaft horsepower rather than shaft horsepower alone. An answer that reverses this split — that the majority goes to jet reaction with the residue driving the propeller — describes no turboprop that has ever been built.

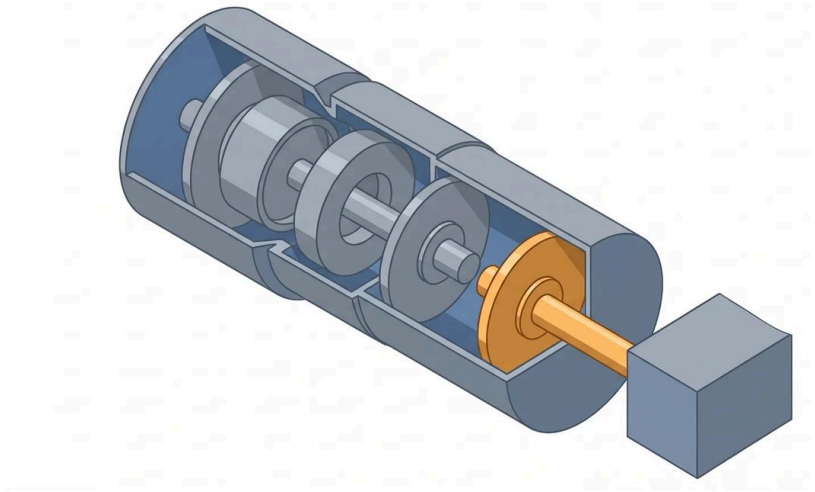
Two shaft arrangements exist. In a **direct-coupled** (fixed-shaft) turboprop the propeller reduction gearbox is driven from the same shaft as the compressor, so propeller and compressor speeds are permanently locked together. In a **free-turbine** turboprop a separate power turbine, mechanically independent of the gas generator, drives the gearbox.

The turboprop's ceiling is set by the propeller, not by the engine. As flight speed rises, the helical speed of the propeller blade tips — the vector sum of rotational and forward speed — approaches the speed of sound, and compressibility losses cut propeller efficiency sharply. That is what limits the type to the lower end of the speed range, and it is why the aircraft that cross that boundary switch to a ducted fan, where the duct and the fan's own pressure ratio hold the relative flow at the blades under control.

Turboshaft

A turboshaft is a gas turbine arranged to deliver essentially its entire output as shaft power to an external load — most commonly a helicopter transmission, but also an auxiliary power unit, a ground power set or a marine drive. The exhaust is deliberately arranged to leave at low velocity, since any residual jet thrust on a helicopter would be a nuisance rather than a benefit, and on a military helicopter the exhaust is often turned and mixed with cool air to reduce its infra-red signature.

Almost all helicopter turboshafts use a **free turbine**. The reason is a matter of operating requirements rather than efficiency. A helicopter rotor must be held at very nearly constant speed at all times, while the gas generator has to be free to accelerate and decelerate over a wide speed range as the pilot changes collective pitch. Locking the two together would make that impossible. With a free turbine, the gas generator runs at whatever speed the power demand requires and the power turbine turns with the rotor at its own governed speed, the two connected only by the stream of gas passing between them.



In a free-turbine turboshaft the power turbine sits on its own shaft, linked to the gas generator only by gas.

Free turbine and direct-coupled arrangements compared

Aspect	Direct-coupled (fixed shaft)	Free turbine
Mechanical arrangement	Output drive taken from the same shaft that carries the compressor and its turbine	A separate power turbine with no mechanical connection to the gas generator shaft; the only link is the gas stream
Starting	The starter must turn the compressor *and* the load through the gearbox, so a higher starting torque is needed	The starter turns only the gas generator, so starting torque is much lower
Load at start	The propeller or rotor turns as soon as the engine turns	The propeller or rotor can be held stationary, for example by a rotor brake, while the core is running
Speed matching	Compressor and output speeds are locked in a fixed ratio; the compressor must run near its design speed whenever the load is being driven	Each rotating assembly finds its own best speed independently
Engine braking	Available — the load is mechanically tied to the compressor, so a windmilling propeller drives the core	None — there is no mechanical path from the load back to the gas generator

Overspeed risk	Loss of load cannot instantly overspeed the assembly, because the compressor remains as a load	Sudden loss of load can allow the power turbine to accelerate very rapidly, so an independent power-turbine overspeed protection system is essential
Typical use	Some turboprop installations	Virtually all helicopter turboshafts and many turboprops

Whichever arrangement is used, the energy accounting in a free-turbine engine is worth stating plainly, because it is a favourite examination point: the gas generator's job is only to *make* the high-energy gas stream, and the bulk of the energy in that stream is then extracted by the **free (power) turbine** to drive the output shaft. It is not mostly consumed driving the gas generator, and it is not mostly expelled through the exhaust — the whole design intent is to convert as much of it as possible into shaft power.

Free-turbine overspeed. Because there is no mechanical connection between the power turbine and the gas generator, anything that suddenly removes the load — a sheared drive shaft, a decoupled gearbox, a propeller or rotor departing its mounting — leaves the power turbine with the full gas stream still passing through it and nothing to absorb the work. Acceleration can be extremely rapid, and a burst turbine disc is not containable. This is why free-turbine engines carry a dedicated overspeed protection system, usually acting independently of the normal fuel governing, and why its periodic functional test is not an optional item.

Matching the type to the flight regime

Speed band (typical)	Best-suited type	Reason
Low subsonic, up to roughly Mach 0.5–0.6	Turboprop	A propeller moves an enormous mass of air for a small velocity increase, giving the highest propulsive efficiency of all at low speed. Above this band, tip compressibility ends the advantage.
High subsonic, roughly Mach 0.75–0.9	High-bypass turbofan	The bypass stream keeps the jet velocity close to the flight speed, which is where propulsive efficiency is best, while the duct keeps the fan blades working in controlled conditions.
Transonic and low supersonic	Low-bypass turbofan, usually with reheat	A higher jet velocity is needed to keep a useful velocity difference at these speeds, and reheat provides large thrust increases for short periods.
High supersonic	Turbojet	Only a very high jet velocity still exceeds the flight speed by a useful margin; ram compression in the intake also does an increasing share of the compression.

How the speed question is usually asked. If an examination question gives you a flight speed of around 800 mph and asks which engine type would typically be used, it is testing the velocity-difference argument, not aircraft-specific knowledge. At that speed a high-bypass fan, whose exhaust leaves at only a little above the aircraft's own speed at cruise, would have almost no velocity difference left to produce net thrust. The engine that still works is the one with the high jet velocity: the turbojet.

Bypass Ratio

Bypass Ratio (BPR) = mass flow of bypass air ÷ mass flow through the core

$$BPR = \frac{\dot{m}_{bypass}}{\dot{m}_{core}}$$

Type	Typical BPR	Characteristics
Turbojet	0 (no bypass)	All air through core. High jet velocity. Noisy. Efficient at very high speeds.
Low-bypass turbofan	0.3 – 2:1	Military fighters. Afterburner capable. Balance of speed and efficiency.
High-bypass turbofan	5:1 – 12:1	Modern airliners. Large fan moves enormous mass of air at moderate velocity. Very fuel-efficient, quiet.
Ultra-high-bypass	12:1+	Latest generation engines (geared turbofans). Maximum propulsive efficiency.

Why high bypass is efficient: Propulsive efficiency is maximised when the jet exhaust velocity is only slightly faster than the aircraft speed. A high-bypass fan accelerates a *large mass* of air by a *small amount* — same thrust, less wasted kinetic energy, less noise.

Calculating and reading a bypass ratio

Worked example

An engine takes in a total of 350 kg/s at its design point. Of that, 50 kg/s passes through the core and the rest goes down the bypass duct.

Bypass flow $= 350 - 50 = 300\ \text{kg/s}$

$BPR = 300/50 = 6$, quoted as **6:1**.

Read that back carefully: a bypass ratio of 6:1 does not mean six-tenths of the air bypasses the core. It means six kilogrammes bypass for every one that goes through, so **six-sevenths** — about 86% — of the total intake flow never sees the combustor at all.

A bypass ratio is quoted for a stated operating condition, usually cruise or sea-level static, because the split between the two streams shifts somewhat with flight condition and power setting. When comparing two engines, make sure the conditions match.

Propulsive efficiency, put into numbers

Propulsive (Froude) efficiency — the fraction of the energy given to the air that ends up as useful propulsive work:

$$\eta_p = \frac{2V_0}{V_0 + V_j} = \frac{2}{1 + V_j/V_0}$$

where V_0 is the flight speed and V_j the jet velocity. The expression is exactly 1 when $V_j = V_0$ — but that produces no thrust at all — and exactly 0 when $V_0 = 0$, which is why an engine on the brakes does no useful work no matter how much fuel it burns.

Worked example — the same thrust from two very different engines

Both aircraft are cruising at $V_0 = 240$ m/s.

Turbojet, jet velocity 600 m/s: $\eta_p = 2 \times 240/(240 + 600) = 480/840 = 0.57$, i.e. **57%**. Velocity difference available for thrust: $600 - 240 = 360$ m/s per kg/s of airflow.

High-bypass turbofan, effective jet velocity 320 m/s: $\eta_p = 2 \times 240/(240 + 320) = 480/560 = 0.86$, i.e. **86%**. Velocity difference available: $320 - 240 = 80$ m/s per kg/s.

The turbofan converts a far higher share of the energy it puts into the air into useful work — but it gets only 80 N of thrust from each kg/s of airflow, against the turbojet's 360 N. To match the turbojet's thrust it must handle $360/80 = 4.5$ times the mass flow. That factor of four and a half, in one number, is why the fan is so large.

What a large fan costs

Raising the bypass ratio is not free, and a complete answer has to give both sides of the ledger. The penalties are:

- **Diameter, weight and drag.** The nacelle grows with the fan. A larger nacelle has more wetted area and more external drag, and the structure to carry it is heavier. Beyond a point the installed drag and weight penalty begins to eat the fuel saving the cycle produced.
- **Ground clearance and undercarriage.** A larger fan on a low-wing aircraft either needs a longer, heavier undercarriage or a flattened nacelle and a pylon reposition — airframe consequences that are expensive on a re-engined design.
- **Shaft speed mismatch.** A big fan must turn slowly, because its tip speed is limited. The low-pressure turbine driving it is then forced to run slowly too, and a slow turbine is an inefficient turbine, so more low-pressure turbine stages are needed to absorb the same power. Adding a reduction gearbox solves the mismatch but adds a large, heavily loaded, oil-cooled component of its own.
- **Containment and reverser mass.** Fan blade-off containment structure and the thrust reverser both scale with fan diameter, and both are pure weight.

- **Thrust lapse with forward speed.** This is the aerodynamically important one. Because a high-bypass engine's jet velocity is only a little above the flight speed, the velocity difference — and therefore the net thrust — is eroded quickly as the aircraft accelerates. The worked example above shows the mechanism: an engine with 80 m/s of velocity difference has far less margin to lose than one with 360 m/s.



A bigger fan on a low-wing aircraft cuts ground clearance, needing a longer, heavier undercarriage.

Which side dominates depends entirely on the flight speed. At the cruise speeds flown by commercial transports, the propulsive efficiency gain is large and the penalties are manageable, so the fuel saving wins decisively — which is why bypass ratios have climbed steadily for more than fifty years and continue to climb. Above roughly Mach 1.5, the argument inverts: a high-bypass engine's jet is simply not fast enough to produce useful net thrust against the ram drag at those speeds, so a low bypass ratio, or none at all, wins. There is no single "best" bypass ratio, only a best one for a stated design speed.

Bypass ratio, fan pressure ratio and specific thrust

Three design parameters move together, and recognising the pattern makes engine data far easier to interpret. As the bypass ratio rises, the fan pressure ratio falls: a bigger fan moving more air does not need to raise the pressure of that air by as much to produce the same thrust. Lower fan pressure ratio means a lower bypass jet velocity, which means lower specific thrust — less thrust per kilogramme per second of airflow — which is exactly what forces the engine to be physically larger. So high bypass ratio, low fan pressure ratio, low specific thrust, large diameter and low specific fuel consumption are all descriptions of the same design choice seen from different angles.

Cold stream and hot stream

The bypass air is universally called the **cold stream** and the core exhaust the **hot stream**. The terminology matters practically as well as descriptively: because on a high-bypass engine the great majority of the thrust is carried by the cold stream, a thrust reverser that acts on the cold stream alone recovers most of the available reverse effect while keeping the reverser structure away from turbine-exit gas temperatures. That is why the cascade-type cold-stream reverser described below is the standard fit on high-bypass engines, while hot-stream designs belong to engines with little or no bypass flow.

Bypass ratio is a ratio of mass flows. It is not a ratio of duct areas, not a ratio of thrusts, and not a percentage. An engine with a bypass ratio of 9:1 does not produce nine times more thrust from the fan than from the core — the thrust split is different from the mass split, because the two streams leave at different velocities. Read the definition in the formula box above literally and the arithmetic will come out right.

Engine Constructional Arrangement

Section	Components	Function
Air intake	Nacelle inlet lip, intake duct	Delivers clean, uniform airflow to the compressor. Subsonic intakes are a simple divergent duct. Supersonic intakes use variable geometry (ramps/cones).
Compressor	Axial (rows of rotating blades + stator vanes) or centrifugal (impeller + diffuser)	Raises air pressure 10–50×. Axial: higher efficiency, many stages. Centrifugal: simpler, more robust, used in smaller engines.
Combustion section	Can-type, cannular (can-annular), or annular combustors, fuel nozzles, igniters	Burns fuel at constant pressure. Must mix fuel and air, sustain a stable flame at high airflow, and cool the liner to survive extreme temperatures.
Turbine section	Rows of stator (nozzle guide vanes) + rotor blades. HP turbine drives HP compressor; LP turbine drives LP compressor/fan.	Extracts energy from hot gas to drive the compressor (and fan in turbofans). Made of nickel superalloys, often single-crystal. Air-cooled blades.
Exhaust section	Exhaust cone, jet pipe, nozzle (convergent or convergent-divergent)	Accelerates remaining gas energy into thrust. Convergent nozzle for subsonic jets; C-D nozzle for supersonic applications.

Air intake

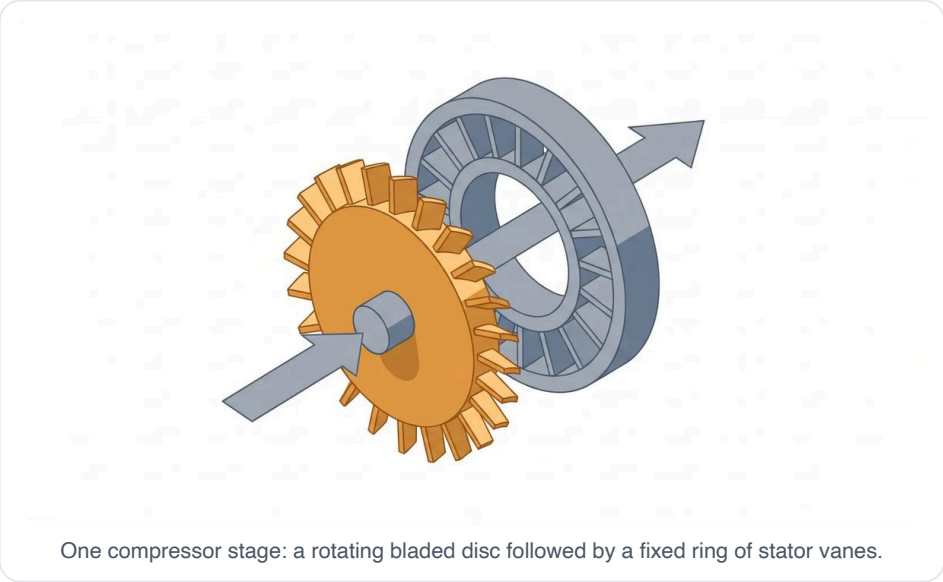
The intake is the only part of the engine with no moving parts and no fuel burnt in it, and it is nonetheless one of the most performance-critical. Its job is to deliver air to the compressor face with the highest possible total pressure and with as little distortion across the face as it can manage. Total pressure lost in the intake can never be recovered — the compressor can only multiply the pressure it is given — so a percentage of total pressure lost in the duct translates almost directly into a loss of thrust for the same fuel flow. That is why intake ducts are kept short, smooth, free of steps and gaps, and why damage to an intake lip or acoustic lining is treated as a performance matter and not merely a cosmetic one.

Distortion is the second concern. If part of the compressor face receives air at a different pressure or angle from the rest — because of a crosswind on the ground, a high angle of attack, sideslip, or a damaged lip — each compressor blade passes alternately through good and poor air once per revolution. The blades in the distorted sector run at a higher incidence than the rest, and it is those blades that stall first. Intake distortion therefore **reduces** surge margin, and an engine that surges only in a crosswind or at high angle of attack is telling you where to look.

Other intake features an engineer meets in practice: hot-air **anti-icing** supplied from the compressor to the lip, because ice accreting on the lip both distorts the flow and can shed into the compressor as a solid slab; **acoustic linings** in the duct wall; drain holes; and the vortex dissipation arrangements fitted to some installations. Supersonic intakes are a different discipline altogether: they use variable ramps or a translating centre-body to position a system of shock waves so that the flow arrives at the compressor subsonically and with good pressure recovery, and at high Mach numbers the intake performs a large share of the total compression before the compressor sees the air at all.

Compressors

A compressor **stage** is one row of rotating blades followed by one row of fixed blades: **one rotor plus one stator**. That pairing is the unit of compression, and a multi-stage compressor is simply that unit repeated. The rotor does work on the air, raising both its velocity and its static pressure; the stator, which is a divergent passage and does no work, converts the added velocity into a further static pressure rise and straightens the flow so the next rotor row meets it at the right angle. Inlet guide vanes, where fitted, sit ahead of the first rotor to set the initial swirl and are not themselves part of a stage.



One compressor stage: a rotating bladed disc followed by a fixed ring of stator vanes.

The pressure rise available from a single axial stage is small — of the order of **1.1:1 to 1.3:1** — because the stage is working against an adverse pressure gradient and the flow will separate if pushed harder. A single **centrifugal** stage, by contrast, can achieve something on the order of **5:1**, because the impeller raises the pressure by centrifugal action as well as by diffusion, and centrifugal effects do not suffer the same separation limit.

Worked example — stage pressure ratios multiply, they do not add

A compressor of ten axial stages, each achieving 1.30:1:

$1.30^{10} = 13.8$, so the overall pressure ratio is about **13.8:1** — not 13:1, and certainly not 10 × 1.3.

Add a further four stages at 1.25:1 apiece and the ratio becomes $13.8 \times 1.25^{4} = 13.8 \times 2.44 = 33.7$, i.e. about **34:1**. This multiplication is why an overall pressure ratio of 40:1 or more needs a great many stages, and why every fraction of a percent of stage efficiency is worth chasing.

Compressor pressure ratio is defined as delivery pressure divided by inlet pressure:

$$r = \frac{P_{outlet}}{P_{inlet}}$$

It is an overall figure for the compressor as a whole, taken across it from front to back. It is not a measure of airflow, and it is not the ratio achieved by any one rotor row in isolation.

Feature	Axial compressor	Centrifugal compressor
Pressure ratio per stage	Roughly 1.1:1 to 1.3:1	Of the order of 5:1
Stages needed for a high overall ratio	Many — ten to twenty is common	One or two
Why the efficiencies differ	Each stage turns and diffuses the flow only slightly, and the flow stays axial from front to back, so the losses per stage are small	The air is turned through a right angle, thrown outward at very high velocity and then decelerated through a large area change in the diffuser; each of those actions costs total pressure
Frontal area for a given mass flow	Small — the flow stays axial	Large — the flow is thrown radially outward
Robustness and tolerance of damage	Lower; thin aerofoils are vulnerable to FOD and erosion	Higher; the impeller is a thick, stiff component
Typical application	All large engines; the whole compression system of a transport engine	Small turboprops, turboshafts and APUs; also as the final stage of an axi-centrifugal compressor

Several construction details follow from the physics. Axial blades are **twisted** from root to tip because blade speed increases with radius while the axial air velocity does not, so the angle at which the air meets the blade would otherwise change along the span. The annulus **tapers** from front to rear so that the rising density does not slow the axial velocity. Materials change through the compressor as the temperature climbs — titanium alloys in the cooler front stages, nickel-based alloys and steels at the rear where delivery temperatures can exceed 600 °C on a high-pressure-ratio engine. Modern front stages are frequently made as **blisks**, with blades and disc machined from a single forging, which removes the root fixings and their weight but means an individual blade can no longer simply be replaced.

Compressor stall and surge

Because the compressor is working against a rising pressure, its blades behave exactly like small aerofoils and will stall if the angle of attack becomes too great. Two related phenomena result. **Rotating stall** is a localised cell of stalled flow that propagates around the annulus at less than rotor speed; it may be sustained and may not be obvious except as vibration and a loss of efficiency. **Surge** is a complete breakdown of the flow, in which the pressure gradient briefly wins and the air in the compressor reverses direction. Surge

announces itself with a loud bang, a rapid rise in exhaust gas temperature, a drop or hesitation in rotor speed, vibration, and sometimes visible flame from the intake or the tailpipe.

The causes are all variations on the same theme — the compressor is being asked for more pressure rise than it can deliver at that speed:

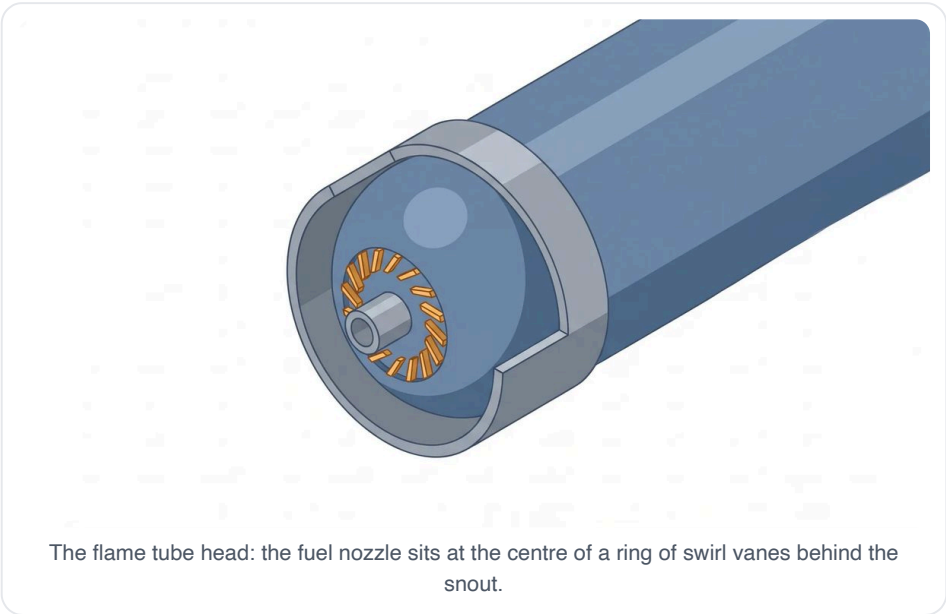
- Too rapid an increase in fuel flow, raising the turbine entry temperature and the back-pressure faster than the rotor can accelerate.
- Intake distortion from crosswind, sideslip, high angle of attack or damage.
- Deterioration: eroded or fouled aerofoils, opened tip clearances, worn seals. Deterioration moves the working line closer to the surge line, so an old engine has less margin than a new one.
- Faults in the handling system: a bleed valve that has failed closed at low speed, or variable stator vanes out of rig.
- Blade damage, contamination or ice accretion.
- Excessive customer bleed extraction from the high-pressure compressor, which reduces the flow available to the turbine and moves the working line up towards the surge line.

The countermeasures are all about matching the front and rear of the compressor at off-design speeds. At low rotational speed the overall pressure ratio is low, so the air arriving at the rear stages is much *less* dense than it is at the design point. A stage can only pass so much volume flow, and a given mass of low-density air occupies far more volume than the rear stages were sized to handle, so they cannot swallow the volume flow the front stages are delivering. Get the direction right, because it is a standing examination trap: high density *raises* a stage's capacity to swallow mass flow, and it is the low density at low speed that limits what the rear of the compressor can pass. The front stages back up against that restriction and are driven towards stall, so surplus air has to be spilled. That is the job of the **variable bleed valves** (also called handling bleed valves), and it fixes their operating logic: they are **open** during starting and at low power, spilling air from an intermediate stage into the bypass duct or overboard, and they **close** progressively as the engine accelerates and the compressor becomes matched. A worked check makes the direction unmistakable: at 20% N2 during a start, the valves are fully open; at take-off power they are shut.

Variable stator vanes attack the same problem from the other direction. At low speed they are held in a partly closed position, adding swirl in the direction of rotor rotation so that the following rotor blades see an acceptable angle of attack; as speed rises they are progressively opened towards their high-speed setting. **Multiple spools** are the third and most fundamental answer: splitting the compressor into two or three independent assemblies lets each run at the speed that suits it rather than forcing all stages onto one shaft.

Combustion section

The combustion chamber has to do something that at first sight looks impossible. Air leaves the compressor at a velocity far higher than the flame speed of kerosene, so a flame placed directly in that stream would simply be blown away. The chamber solves this by creating a sheltered, low-velocity, recirculating region. Air entering the flame tube through the snout and through swirl vanes around the fuel nozzle is given a strong rotational component, which sets up a toroidal vortex in the **primary zone** that continuously carries hot burning gas back upstream towards the incoming fuel spray. That recirculation is the flame holder, and everything else about the chamber is arranged around it.



The flame tube head: the fuel nozzle sits at the centre of a ring of swirl vanes behind the snout.

The airflow through the chamber is divided deliberately, and the division is what makes the turbine survivable:

Zone	What happens there	Why it exists
Primary zone	Fuel is atomised into recirculating air at close to the chemically correct (stoichiometric) mixture strength and burns. Local flame temperature here can approach 2,000 °C.	Complete, stable combustion requires something near stoichiometric conditions; nothing else will hold a flame reliably.
Secondary (intermediate) zone	Further air is admitted to complete the burning of any fuel and carbon monoxide leaving the primary zone.	The primary zone alone leaves unburnt products; the intermediate zone finishes the job before the gas is quenched.
Dilution zone	The air not consumed by combustion — around 60% of the chamber flow — is shared between this zone and the liner cooling film in the row below. The larger part of it enters through the dilution holes to bring the gas down to a temperature the turbine can withstand and to produce an acceptable temperature profile across the annulus.	Turbine entry temperature is set here, and so is the radial temperature distribution that determines blade and vane life.
Wall cooling film	The balance of that same 60% is admitted through rows of small holes along the flame tube wall to lay a protective film of relatively cool air between the metal and the flame.	The flame tube cannot survive direct radiation and contact from a 2,000 °C flame without being shielded.

Two figures for "combustion air", and what each one counts. You will meet more than one number for the proportion of the chamber's airflow that takes part in combustion, and they are not competing claims — they draw the boundary in different places. The narrower figure counts only the air admitted to the *primary* zone, the air actually mixed with fuel at close to the chemically correct ratio, which is a fifth to a quarter of the chamber flow. The wider figure counts everything admitted before the dilution holes, that is the primary and secondary zones together, which comes to approximately **40%** of the chamber flow, leaving about 60% for dilution and cooling. Both describe the same chamber. When an examination question asks what percentage of the air passing through the combustion section is burned, the figure it is looking for is **40%**.

All of the air surrounding and cooling the flame tube is **high-pressure compressor delivery air**. It has to be: the air must be at a higher pressure than the gas inside the flame tube in order to flow inwards through the liner holes at all, and only compressor delivery air is. Low-pressure compressor air or fan air would simply not have the pressure to penetrate, so it cannot be used to cool a combustion chamber.

Three chamber layouts are in service:

Type	Construction	Advantages	Disadvantages
Can (multiple)	Several complete flame tubes, each inside its own separate air casing, arranged around the engine. Interconnectors link them so the flame propagates at start and the pressures equalise.	Individual cans can be removed and replaced without disturbing the rest; easy to develop and test one can at a time.	Heavy and bulky; higher pressure loss; poor outlet temperature uniformity.
Cannular (can-annular)	Individual flame tubes, but all housed inside one common annular air casing.	A compromise: retains individual flame tubes while shortening and lightening the assembly.	Still more complex than an annular chamber; outlet profile less even.
Annular	A single continuous chamber. An **inner and an outer flame tube skin** form the annular combustion space, and these sit within inner and outer casings. The flame burns all the way round the annulus.	Shortest and lightest for a given output; lowest pressure loss; the most even outlet temperature distribution; the best surface-area-to-volume ratio for cooling.	Cannot be inspected or replaced piecemeal — access normally requires the engine to be split.

Ignition is needed only to start the fire, not to keep it burning. Two high-energy igniter plugs, usually positioned in the primary zone, fire during the start sequence and during a relight; once the flame is established it is self-sustaining and the igniters are switched off. Continuous ignition is selected as a precaution in heavy rain, icing conditions, severe turbulence and during take-off and landing on some types, because a momentary interruption to the flame would otherwise become a flame-out.

The failure modes worth recognising are cracking and distortion of the flame tube from thermal cycling, burning and local hot spots where the cooling film has been disrupted, and carbon build-up on the fuel nozzles. That last one is deceptively serious: a carboned nozzle produces a distorted, streaky spray pattern instead of a fine cone, which puts a hot streak straight onto the nozzle guide vanes and turbine blades

downstream and burns them locally. Borescope inspection of the chamber and the first-stage vanes is the standard means of finding all of this before it becomes a removal.

Turbine section

A turbine **stage** consists of a row of stationary **nozzle guide vanes followed by a row of rotating turbine blades** — in that order, NGV first. This is worth stating explicitly because it is the reverse of the compressor convention, where the rotor comes first and the stator follows. The reason is functional. In a compressor the rotor must do work on the air before there is any surplus velocity for a stator to diffuse. In a turbine the gas must first be accelerated and turned onto the correct angle before a rotor can extract work from it, and that is precisely what the nozzle guide vanes do. Inlet guide vanes belong to a compressor; a turbine stage never begins with them.

The nozzle guide vanes form a convergent passage, so static pressure falls and velocity rises through them. In normal operation the NGV throat is choked, which means the high-pressure NGV throat area effectively meters the mass flow through the core. It is one of the most critical dimensions in the engine, which is why NGV throat area is a controlled parameter at overhaul and why replacing vanes with the wrong classification changes the engine's whole running line.

Through the rotating blades the gas gives up work, so **total pressure and total temperature both fall** across a turbine — the exact opposite of what happens across a compressor. Turbine blades are twisted from root to tip for the same blade-speed reason as compressor blades, and the twist is usually accompanied by a change in the type of blading: closer to impulse at the root, where blade speed is low, and closer to reaction at the tip, where it is high.

The spool arrangement determines which turbine drives what:

- **Single spool:** one compressor and one turbine on one common shaft. They are mechanically joined, so the compressor and turbine speeds are always **equal** — there is no gearing and no separate assembly to make them differ.
- **Twin spool:** two independent rotating assemblies, that is **two turbines on two shafts**, one running concentrically inside the other. The **high-pressure turbine** — the first turbine stage or stages, immediately behind the combustor where the gas is hottest and at its highest energy — drives the **high-pressure compressor**. The **low-pressure turbine**, the later stages further downstream, drives the **low-pressure compressor and the fan**. Each spool is free to find its own speed.
- **Triple spool:** the same idea taken one step further, with an intermediate-pressure compressor and intermediate-pressure turbine forming a third independent assembly between the two.

Getting this the right way round matters. An arrangement in which the high-pressure turbine drove the low-pressure compressor, or in which one turbine drove both compressors, is not a twin-spool engine. And note that the term for an engine with two independent rotating systems is precisely **twin-spool** — not "compound", which describes something else entirely, and not "complex".

Optimum turbine speed is defined as the rotational speed at which the turbine works most efficiently: the speed at which the ratio of blade speed to gas velocity gives the best compromise between the work extracted and the losses incurred. It is a point of best efficiency, not a fixed percentage of maximum engine rpm. This definition is the root of several design decisions elsewhere in the engine, because the turbine's optimum

speed is usually much higher than the optimum speed of the fan or propeller it has to drive — hence multiple spools, hence reduction gearboxes.

Materials and cooling. Turbine blades and vanes are made from nickel-based superalloys, cast rather than forged, and increasingly cast as *directionally solidified* or *single crystal* components. The purpose of removing grain boundaries is specific: grain boundaries transverse to the load are the preferred path for creep at high temperature, so a blade with no transverse boundaries — or none at all — survives a higher temperature for longer. On top of the alloy goes a ceramic thermal barrier coating, and inside the blade is a network of cooling passages fed with high-pressure compressor air, using internal convection, impingement onto the hottest internal surfaces, and film cooling through rows of small holes that lay a protective layer of cooler air over the outer surface. Taken together, these allow the gas entering the turbine to be hotter than the melting point of the blade alloy. The cooling air is a real cycle penalty — it was compressed at full cost and then largely bypassed the combustor — and it is spent anyway because raising turbine entry temperature is worth more.

Blades are attached to the disc by a **fir-tree** root, deliberately a slightly loose fit when cold so that the blade can be assembled and so that differential expansion is accommodated; centrifugal load and thermal growth take up the clearance in operation. Blade tips may be shrouded, which reduces tip leakage and provides damping through the interlocking shrouds, or unshrouded to save weight at the highest temperatures. Tip clearance itself is a first-order efficiency item — gas leaking over the tip does no work — and many engines control it actively by spraying fan or compressor air onto the turbine casing at cruise: **cooling the casing shrinks it, which closes the tip clearance and raises efficiency**.

The characteristic turbine failure modes are creep, which is the slow permanent extension of a blade under centrifugal load at high temperature; thermal fatigue cracking from repeated start-stop cycles; oxidation and sulphidation attack of the surface; tip rubs; and blocked cooling passages, which cause rapid local burning. **Creep is cumulative and irreversible** — a blade that has crept does not recover on cooling — which is why turbine blade growth is a measured, life-limited parameter and why exceeding a temperature limit even briefly is recorded against the engine. Turbine and compressor discs are separately life-limited on cycles rather than hours, because their limiting mechanism is low-cycle fatigue driven by the number of times the engine has been started and taken to power.

Exhaust section

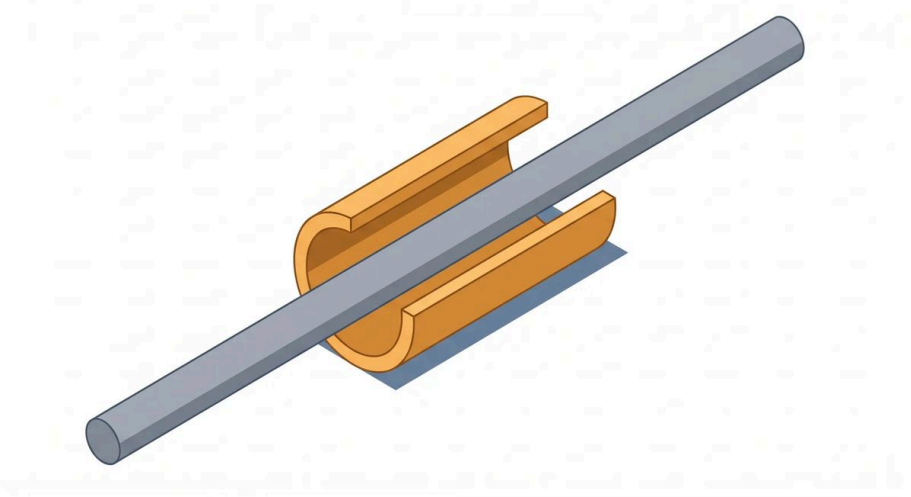
Gas leaves the turbine with a good deal of swirl and at a velocity that is too high to be useful. The exhaust unit therefore begins with a **divergent** passage, formed between the outer jet pipe and the exhaust cone, with radial struts that also serve to straighten the swirl. Slowing the gas here recovers static pressure and reduces friction losses before the propelling nozzle does the final acceleration. The exhaust cone also prevents the gas from flowing across the flat rear face of the turbine disc, which would cause severe turbulence.

The **propelling nozzle** is the smallest cross-sectional area in the exhaust system, and it is the component that sets the back-pressure on the turbine. Its area is therefore a critical matching dimension: reducing the nozzle area at a given fuel flow raises the pressure and temperature in the jet pipe and alters the running line of the whole engine. On subsonic engines the nozzle is a fixed convergent duct. Supersonic engines, and any engine with reheat, use a convergent-divergent nozzle with variable geometry, so that the nozzle can be opened to pass the greatly increased volume of gas when reheat is selected without disturbing the turbine.

As explained above, once the nozzle pressure ratio exceeds the critical value the nozzle chokes and produces pressure thrust in addition to momentum thrust. Noise treatment lives here too: the serrated trailing edges (chevrons) seen on modern nacelles, and the lobed forced mixers used inside long-duct nacelles, both work by promoting a gentler, more gradual mixing of the exhaust with the surrounding air, which reduces the intensity of the shear that generates jet noise.

Shafts, spools and how the sections are held together

The low-pressure shaft runs the full length of the engine inside the hollow high-pressure shaft, which is short and stiff. This concentric arrangement is what allows the two assemblies to run at completely different speeds while occupying the same axis. Some designs run their spools in opposite directions, which reduces the net gyroscopic moment on the aircraft and can improve the aerodynamics at the turbine interduct. Each spool is supported on its own bearings, which are dealt with in the final section of this note.



The long low-pressure shaft runs inside the short, stiff, hollow high-pressure shaft on the same axis.

Modular construction and maintainability

Modular construction means that the engine is designed and built as a set of self-contained major assemblies — typically a fan module, compressor modules, a combustor and high-pressure turbine module, a low-pressure turbine module and an accessory drive module — each individually balanced and each interchangeable with another of the same standard. The purpose is that a defective major assembly can be removed and replaced without stripping the whole engine, so an engine can be repaired in a matter of shifts rather than weeks, often without returning it to the manufacturer.

Two common misreadings are worth heading off. Modular construction does *not* mean that all engines share a specific component layout, and it does *not* refer to any particular technique of assembly. It is a statement about how the engine is divided for maintenance, nothing more. Alongside it sit the other maintainability features an engineer meets daily: borescope ports at defined stations so the gas path can be inspected without disassembly; on-condition maintenance driven by trend monitoring rather than fixed overhaul intervals; and the quick engine change (QEC) build-up, which keeps as much of the installation hardware as possible with the airframe.

Direction of change through the engine — the summary table

If you learn one table in this section, learn this one. Almost every incorrect answer about gas turbine behaviour is a correct mechanism with one of these arrows reversed.

Section	Static pressure	Temperature	Gas velocity
Intake (in flight, subsonic)	Rises	Rises slightly	Falls — the duct is a divergent diffuser
Compressor	**Rises** — substantially, by the overall pressure ratio	**Rises** — compression heats the air, so both go the same way	Axial velocity held roughly constant by the tapering annulus
Combustion chamber	Nominally constant; a small loss in practice	Rises very sharply	Rises — the gas expands at constant pressure, so it must speed up
Turbine	**Falls** — work is being extracted	**Falls** — again both go the same way, but downwards	Rises through the nozzle guide vanes, then falls across the rotor as work is taken out
Exhaust unit / jet pipe	Rises slightly — a divergent passage	Static temperature rises slightly as the flow slows; total temperature is unchanged	Falls
Propelling nozzle (convergent, subsonic)	**Falls**	Falls	**Rises** — this is where the remaining pressure energy becomes thrust

General Engine Requirements and Performance Parameters

Everything described so far explains how a gas turbine works. This section deals with what it is required to achieve, and with the small set of numbers by which its achievement is measured. These are the parameters that appear on a data plate, in a flight manual, on a trend-monitoring plot and in an examination question, and an engineer needs to be able to say not only what each one is but which way it moves when a condition changes.

What a certificated engine has to deliver

An aircraft engine is type-certificated against a dedicated code — in the European system, the Certification Specifications for Engines, CS-E — and the airframe installation is certificated separately under the relevant aircraft code. The requirements the engine has to meet fall into a small number of families:

- **Thrust or power for its weight and size.** Everything installed on an aircraft is weighed, and an engine's thrust-to-weight ratio and frontal area feed directly into aircraft performance and drag.
- **Fuel economy.** Over the life of a transport aircraft, fuel dominates direct operating cost, so specific fuel consumption is usually the single most heavily traded parameter in the design.

- **Reliability and dispatchability.** An engine is judged by in-flight shutdown rate and by how rarely it prevents an aircraft leaving the gate. On extended-range twin operations these figures are the licence to operate.
- **Life on wing.** How long the engine stays installed between shop visits, which in modern practice is governed by deterioration and life-limited parts rather than by a fixed overhaul interval.
- **Structural integrity and containment.** A failed blade must be contained within the casing; rotor discs must not burst, which is why they are life-limited on cycles; overspeed and overtemperature cases must be survivable or protected against.
- **Handling and surge margin.** The engine must accelerate and decelerate on demand across the whole envelope without surging or flaming out, and it must do so with distorted intake flow and with bleed air being taken off.
- **Starting and relight.** It must start reliably on the ground over a wide temperature range, and it must relight in flight within a defined altitude and airspeed envelope.
- **Ingestion tolerance.** Birds, hail, heavy rain, ice shed from the intake or airframe, sand and volcanic dust are all certification cases, and the engine must either continue running or fail safely.
- **Noise and emissions.** Both are subject to international standards, and both have become primary design drivers rather than afterthoughts.
- **Maintainability.** Modular build, borescope access, line-replaceable accessories, on-condition monitoring — discussed in the previous section — are certification and commercial requirements in their own right.

Thrust and its ratings

An engine does not have one thrust; it has a set of ratings, each with its own limits and its own permitted duration:

- **Take-off thrust** — the highest rating, time-limited (commonly five minutes, extended in the event of an engine failure on a twin).
- **Maximum continuous thrust** — the highest rating with no time limit, reserved for abnormal situations such as engine-out operation.
- **Maximum climb** and **maximum cruise** — the normal in-service ratings, set well below take-off thrust.

Derated and reduced-thrust take-offs matter to the engineer as much as to the crew. Where runway length and obstacle clearance allow, the crew may take off at less than full rated thrust, either by selecting a lower published derate or by using an assumed (flexible) temperature. The purpose is engine life: reducing take-off thrust lowers the turbine entry temperature, and because the rate of hot-section deterioration climbs very steeply with temperature, even a modest reduction produces a disproportionate extension of on-wing life. The direction is worth stating plainly — *less* take-off thrust means a *lower* operating temperature and therefore a *longer* life, which is why operators use it whenever performance permits.

Flat rating is the other rating concept an engineer meets constantly. An engine is capable of far more thrust on a cold day than on a hot one, because the air is denser. Rather than let the crew have that extra thrust — which the airframe and the engine structure would have to be designed to absorb — the engine is *flat rated*: its control system holds the rated thrust constant as ambient temperature varies, up to a stated flat-rating temperature, commonly in the region of ISA + 15 °C. Below that temperature the engine is deliberately

restrained and produces exactly its rated thrust. Above it, a temperature limit binds first, and **available thrust falls as ambient temperature rises further**. This is why take-off performance charts show a flat region followed by a declining one, and why a hot-day departure can become performance-limited quite suddenly.

Thrust cannot be measured directly in flight, so it is set by a parameter that varies with it. Depending on the engine, the crew sets thrust either on **engine pressure ratio (EPR)** — the ratio of turbine discharge total pressure to engine inlet total pressure — or on **fan speed (N1)**. Both are valid thrust-setting parameters; which one a given engine uses is a design decision, and the flight manual will say.

What changes the thrust

Change	Effect on thrust	Mechanism — and the direction of it
Ambient temperature rises	**Thrust falls** (above the flat-rating temperature)	Hotter air is less dense, so at the same rotational speed the engine ingests less mass per second. Thrust is proportional to mass flow, so it falls with it.
Altitude increases	**Thrust falls**	Density falls with altitude, so mass flow and thrust fall. Up to the tropopause the falling temperature partly offsets the falling pressure; above the tropopause the temperature stops falling, so density — and thrust — drop away more steeply with height.
Forward speed increases	**Net thrust falls** through the low and middle subsonic range, then recovers at high Mach	Two competing effects: ram drag grows directly with airspeed and reduces net thrust, while ram compression raises intake pressure, mass flow and jet velocity and increases gross thrust. The first dominates at moderate speed; the second, which grows with the square of speed, eventually takes over.
Humidity increases	Negligible effect	Unlike a piston engine, where water vapour displaces oxygen in a fixed cylinder volume and measurably reduces power, a gas turbine's continuous flow and large excess of air make the effect small enough to ignore in practice.
Rotational speed increases	Thrust rises, and rises disproportionately fast at the top of the range	Raising rpm increases both the mass flow and the pressure ratio, and the two effects compound. Most of the engine's thrust is produced in the upper part of its speed range, which is why a small percentage change in N1 near take-off power is a large change in thrust.
Bleed air or shaft power is extracted	**Thrust falls**	Air taken from the compressor for air conditioning, anti-icing or pressurisation has been compressed at full cost and then removed before it can do work in the turbine; shaft power taken through the accessory gearbox is likewise unavailable for propulsion.

Specific fuel consumption

Thrust specific fuel consumption (TSFC) for a thrust-producing engine:

$$TSFC = \frac{\dot{m}_f}{F_n}$$

Fuel flow divided by net thrust. Common units are kg per kN per hour, or lb of fuel per lb of thrust per hour.

For a shaft engine the equivalent is **specific fuel consumption based on power**: fuel flow divided by shaft power, in kg per kW per hour or lb per shp per hour.

SFC is a cost, not a score — lower is better. It measures how much fuel you must burn to buy one unit of thrust for one hour, so the *smaller* the number, the better the engine. Every improvement in thermal or propulsive efficiency **lowers** SFC. Raising the bypass ratio at subsonic cruise lowers SFC. Raising the overall pressure ratio lowers SFC. If you find yourself writing that an engine has become more efficient and its SFC has therefore increased, the sign is the wrong way round.

Worked example — what a few percent of SFC is worth

An engine holding 22 kN of cruise thrust on a fuel flow of 1,150 kg/h:

$TSFC = 1150/22 = 52.3 \text{ kg per kN per hour.}$

A design improvement lowers the SFC by 5%, to 49.7 kg/kN/h. At the same 22 kN of thrust the fuel flow becomes $49.7 \times 22 = 1\,093 \text{ kg/h}$, a saving of 57 kg per hour per engine.

Over a ten-hour sector on a twin, that is $57 \times 10 \times 2 = 1\,140 \text{ kg}$ of fuel — and, because the aircraft no longer has to carry that fuel, a further small saving on top. A five percent SFC improvement is a very large commercial result, which is the reason engine manufacturers will accept significant weight and complexity penalties to obtain one.

The three efficiencies, and how they connect to SFC

Thermal efficiency — how much of the fuel's heat the core turns into mechanical work. Raised chiefly by increasing overall pressure ratio.

Propulsive efficiency — how much of that work ends up moving the aircraft rather than stirring up the wake:

$$\eta_p = \frac{2}{1 + V_j/V_0}$$

Overall efficiency is the product of the two:

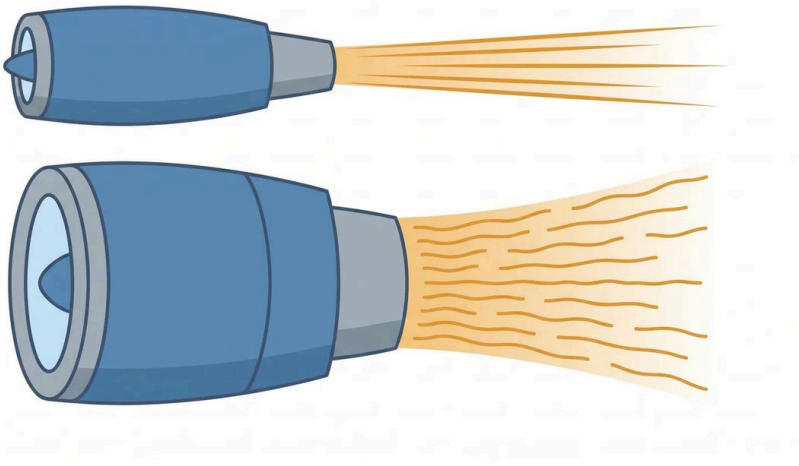
$$\eta_o = \eta_{th} \times \eta_p$$

and specific fuel consumption follows directly from it:

$$\frac{\dot{m}_f}{F_n} = \frac{V_0}{\eta_o \times LHV}$$

where LHV is the fuel's lower calorific value. At a given flight speed, SFC is **inversely proportional** to overall efficiency — the formal statement of "lower is better".

The two ways of improving an engine are therefore quite distinct, and they are pursued by different parts of the design. Thermal efficiency is improved by working the core harder: higher pressure ratio, higher turbine entry temperature, better component efficiencies, less cooling air for the same temperature. Propulsive efficiency is improved by moving more air more slowly: higher bypass ratio, lower fan pressure ratio. Modern engines pursue both at once, which is why they have both a very high overall pressure ratio and a very large fan.



More air moved more slowly: a large fan's broad, gentle stream compared with a narrow, fast jet.

Limits, indications and condition monitoring

The parameters the engineer works with day to day are the ones that keep the engine inside its certificated envelope:

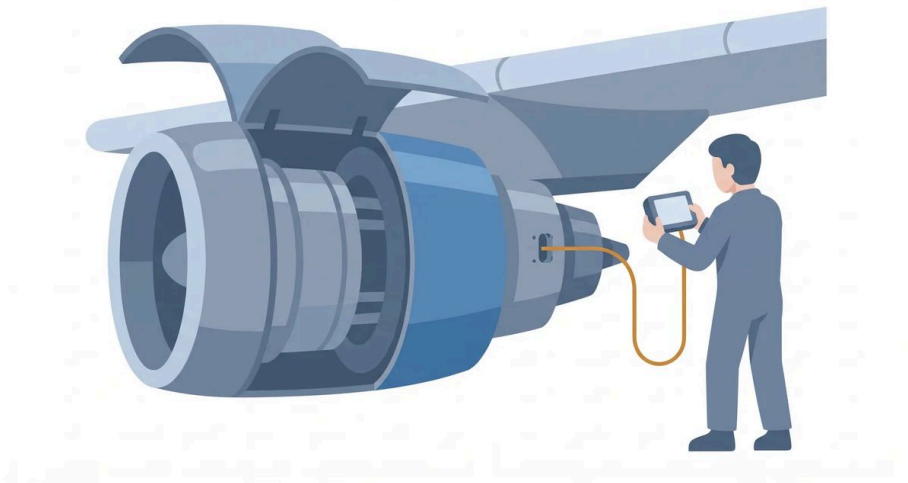
- **Rotor speeds** (N1, N2 and where fitted N3) — each with a maximum, and each an indication of the health of its own spool.
- **Exhaust gas temperature** (EGT, or ITT/TGT depending on where it is measured) — the most closely watched limit on any gas turbine, because it is the proxy for turbine entry temperature, which cannot be measured directly.
- **Fuel flow** — the controlled variable, and a sensitive indicator of deterioration when compared against thrust.
- **Oil pressure, oil temperature, oil quantity** and **vibration** — the mechanical health group.

Power on a gas turbine is adjusted by one thing and one thing only: **fuel flow**. The thrust lever does not open an air valve. Increasing fuel flow raises the gas temperature and energy through the turbine, the turbine produces more work, the spool accelerates, and the compressor consequently draws in more air. The increase in airflow is therefore a *result* of the change in fuel flow, not something the crew commands separately.

EGT margin is the single most useful health parameter on a modern engine. It is the difference between the EGT the engine actually reaches at a reference take-off condition and the certificated limit. A new engine has a comfortable margin. As it deteriorates — tip clearances open, aerofoils erode and become fouled, seals wear, cooling holes partially block — it becomes less efficient, so it must run **hotter** to produce the same thrust, and the margin **falls**. When the margin approaches zero the engine can no longer achieve its rated thrust on a hot day, and that, rather than any single failure, is what most often drives a shop visit. Trend monitoring of EGT

margin, fuel flow and rotor speeds against a reference model is what turns this into a planned removal rather than a delay.

An overtemperature is an event, not a reading. Turbine limits are set by the creep and oxidation behaviour of the hot-section materials, and creep damage is cumulative and permanent. An exceedance therefore has to be recorded against the engine with its magnitude and duration, and the maintenance manual will prescribe the inspection — typically a hot-section borescope, sometimes a module removal — that follows. It cannot be dismissed on the grounds that the engine subsequently ran normally, because a crept blade runs perfectly normally right up until it does not.



After an overtemperature, the manual typically prescribes a hot-section borescope inspection of the engine.

Thrust Reversers

Redirect engine exhaust forward to **decelerate the aircraft on the ground** after landing. Types:

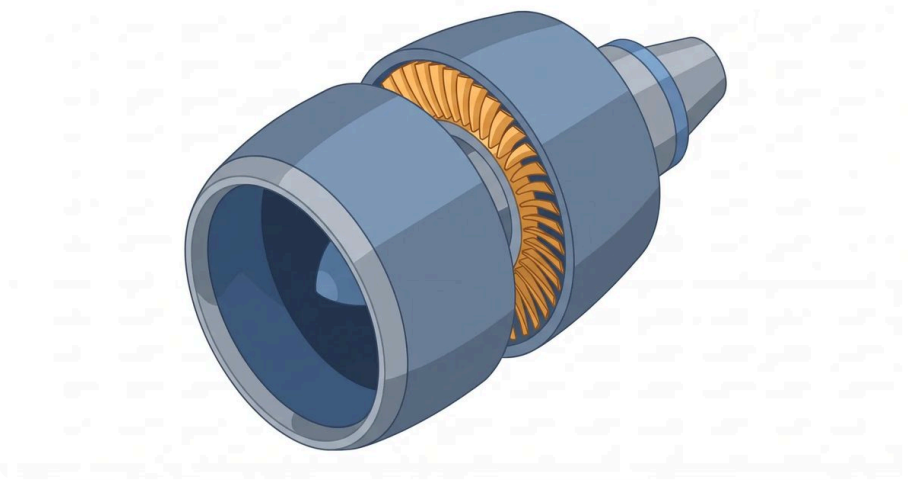
Type	Mechanism	Typical Use
Cascade (cold stream)	Translating cowl slides aft, exposing cascade vanes that redirect bypass air forward	High-bypass turbofans (most common)
Clamshell/bucket	Two doors (buckets) swing into the exhaust stream, deflecting it forward and outward	Low-bypass engines, older designs
Blocker door	Doors block the fan duct and deflect air through openings in the nacelle	Some high-bypass turbofans

What a reverser actually does to the flow

The name is misleading and the misunderstanding it causes is worth clearing up first. A thrust reverser does **not** reverse the direction of the gas flowing through the engine core. Air still enters at the front, is still compressed, burned and expanded in the normal direction, and the core still runs exactly as it does in forward thrust. What the reverser does is intercept the efflux *after* it has been produced — on a high-bypass engine,

overwhelmingly the bypass efflux — and redirect it so that it leaves with a large forward-pointing component. The reaction to that forward-moving mass of air is a rearward force on the aircraft, and that is what decelerates it.

Note "component", not "direction". The redirected air is turned forward and outward at an angle, never straight ahead: it has to clear the wing, the fuselage and the ground, and it must not be re-ingested by its own engine. Because the turning is incomplete and the redirected flow is aerodynamically untidy, the rearward force obtained is only a fraction of the forward thrust the same power setting would produce — commonly quoted as being in the region of 40 to 50%.



With the rear sleeve slid aft, a ring of vanes turns the bypass air forward and outward, never straight ahead.

Cold stream and hot stream reversers

On a high-bypass turbofan the great majority of the thrust is carried by the cold bypass stream. Reversing that stream alone therefore recovers most of the available reverse effect, and it does so with a mechanism that never sees turbine-exit gas temperatures — a large saving in weight, cost and complexity. That is why the cascade type is the standard fit on modern transports and why the core exhaust on such an installation is usually left to continue rearward, unreversed, throughout the reverse selection.

Engines with little or no bypass flow have no cold stream worth reversing, so their reversers must act on the hot exhaust. Clamshell or bucket doors, swinging into the efflux behind the propelling nozzle, are the classic solution, and they have to be built from materials and with clearances appropriate to exhaust gas temperature.

How a cascade reverser operates

The mechanism of the common cold-stream reverser repays understanding, because most reverser defects are mechanical:

- The rear portion of the nacelle is a **translating sleeve** running on tracks. In the stowed position it forms the smooth outer line of the nacelle and covers the cascade vanes completely.
- On selection, actuators — pneumatic, hydraulic or electric depending on the type — drive the sleeve **aft**, uncovering a ring of fixed **cascade vanes** that are shaped to turn the air forward and outward.

- **Blocker doors**, hinged inside the sleeve and connected to the fixed inner structure by drag links, are pulled across the bypass duct as the sleeve translates. With the duct blocked and the cascades uncovered, the fan air has nowhere to go but out through the cascades.
- Actuators are usually tied together by flexible synchronising shafts so the sleeve cannot translate unevenly and jam.

Selection, interlocks and indication

Reverse is selected in a defined sequence that is protected by interlocks, and the first of those interlocks is the position of the thrust lever. **Reverse thrust can only be selected with the thrust lever at the idle stop** — that is, closed. Only then can the reverse levers be raised to command deployment, and only once deployment is confirmed can reverse power be applied. The purpose is obvious once stated: deploying a reverser with the engine at high forward power would subject the structure to loads it is not designed for and would produce a violent and unpredictable change in the aircraft's behaviour. Additional interlocks commonly include air/ground logic and radio altimeter or wheel-speed inputs, so that deployment is physically inhibited in flight.

The crew's confirmation that the reversers have actually deployed is a **sequence of lights**, supported by messages on the engine display system. A typical arrangement gives one indication as the reverser unlocks, another while it is in transit, and a third when it reaches the fully deployed position, so the crew can see each step of the sequence. What the crew does *not* rely on is the physical sensation of deceleration or an audible warning — deceleration is felt from the wheel brakes as well, and a reverser can be unlocked without being deployed. Positive position indication is the only acceptable confirmation.

> **Inadvertent deployment is a catastrophic failure case.** A reverser deploying in flight, particularly asymmetrically, produces an immediate and severe loss of thrust on one side combined with a large disturbance to the wing's airflow, and it has caused hull losses. That is why certification requires multiple independent locks — the actuators' own locks plus at least one further, independent locking system — why auto-restow logic is fitted, and why reverser lock rigging, functional checks and deactivation procedures are among the most tightly controlled tasks in the maintenance manual. A reverser deactivated under the minimum equipment list must be secured *stowed and locked* in accordance with the approved procedure, not merely selected off.

Operational and maintenance considerations

- **Stowing speed.** Reversers are stowed before the aircraft slows below a speed specified for the type, typically somewhere in the region of 60 to 80 kt. Below that speed the reversed efflux can be re-ingested by the engine, which raises intake temperature and can cause surge, and it can also lift debris from the runway into the intakes.
- **Where reverse actually earns its keep.** On a long, dry runway the wheel brakes do most of the stopping and reverse contributes comparatively little. Its value rises sharply on a wet or contaminated runway, where the tyre-to-surface friction available to the brakes is poor but the reverser's effectiveness is unaffected by the surface, and in reducing brake energy absorption and brake wear on every landing.
- **Reverse idle and maximum reverse.** Most operators use reverse idle as the normal case and reserve maximum reverse for short, contaminated or otherwise demanding landings, partly for noise and partly

because high reverse power at low speed aggravates re-ingestion.

- **Inspection points.** Cascade vane erosion and cracking; blocker door hinges, seals and drag link wear; translating sleeve tracks, rollers and seals; actuator and synchronising shaft condition; lock function; and any evidence of hydraulic or pneumatic leakage in the actuation supply. Reverser rigging is a controlled task, because an out-of-rig reverser can produce asymmetric drag or fail to reach a locked position.

Lubricants, Fuels, Lubrication Systems and Fuel Systems

Two fluids circulate through every gas turbine, and neither of them is doing only the job its name suggests. The oil lubricates, but its dominant duty is to carry heat away from bearings and gears that no airflow can reach. The fuel provides the energy, but it is also the engine's principal heat sink and the lubricant for its own high-pressure pump. Understanding the second job of each fluid explains most of what looks arbitrary about the two systems.

What a turbine lubricant has to do

- **Lubricate** — reduce friction and wear at bearings, gear teeth, splines and seals.
- **Cool** — this is the dominant duty. A bearing chamber sits inside a hot engine, receives heat conducted along the shafts and generated by the bearing itself, and has no cooling airflow of its own. The oil carries that heat away to an oil cooler, and the quantity of oil circulated is set by the cooling requirement, not by the lubrication requirement.
- **Clean** — carry wear debris away from the contact surfaces to the filters and magnetic chip detectors, where it can be found.
- **Protect** — prevent corrosion of parked or stored components.
- **Damp** — provide the squeeze-film damping that controls shaft vibration at bearing supports.

Turbine oils are **synthetic ester-based** products, not mineral oils. Mineral oils cannot survive the bulk and local temperatures in a turbine bearing chamber: they oxidise, break down and form carbon deposits (coking) in the hottest passages, and coke blocks oil jets and vent lines. Two broad classes are in use, usually described by their viscosity: a Type I class of about 3 centistokes and the more widely used Type II class of about 5 centistokes, covered respectively by the long-standing specifications MIL-PRF-7808 and MIL-PRF-23699, with higher-thermal-stability variants of the latter also available. Which one an engine takes is stated by the manufacturer, and it is not a matter of preference.

Two practical points that catch people out.

Do not mix oils. Different types, and in some cases different brands of the same type, may not be intermixed without approval. The maintenance manual states which products are approved and what the mixing rules are; a mixture outside those rules can produce sludge, seal swell or reduced thermal stability.

Turbine oils are a health hazard. The ester base and the additive package can be absorbed through the skin and are irritant to eyes and airways, and oil mist is a respiratory hazard. Wear the gloves specified in the safety data sheet, avoid skin contact, and treat contaminated clothing accordingly.

Check the level in the specified window. Oil migrates from the tank into the bearing chambers and gearbox after shutdown. The manual specifies a period after shutdown within which the oil level check is valid; a check made outside that window can read low when the tank is full or full when it is low, and either error can put an aircraft into service in the wrong condition.

The lubrication system is three systems

A gas turbine lubrication system is conventionally described as **three** sub-systems, and questions on this point expect that number:

Sub-system	What it does	Principal components
Pressure (feed) system	Delivers oil from the tank to every bearing chamber and to the gearboxes at a controlled pressure and flow.	Oil tank with deaerator, pressure pump, pressure filter with an impending-bypass indicator and a bypass valve, pressure relief or regulating valve, oil cooler, distribution pipework, oil jets and last-chance filters.
Scavenge system	Collects the oil that has done its work in each chamber and returns it to the tank.	One scavenge pump per bearing chamber and one for each gearbox, scavenge filters, magnetic chip detectors, scavenge return line and deaerator.
Breather (vent) system	Controls the internal air pressures, venting the air that leaks into the chambers past the seals so the chambers do not pressurise.	Vent pipework from each chamber, a centrifugal breather (rotating deoiler) in the gearbox, and an overboard or exhaust-duct vent.

Pressure system detail, and the two system types

The oil tank is deliberately larger than the oil it holds, both to allow for expansion and to give the returning aerated oil somewhere to separate. A deaerator at the scavenge return spins the froth so the air can be released, because oil full of bubbles neither lubricates nor cools properly and would give a false quantity indication.

Two arrangements govern the delivered pressure, and the distinction is a standard examination point:

- **Pressure relief valve system.** A relief valve set to a chosen pressure spills surplus oil back to the tank or to the pump inlet, so the delivered pressure is held at approximately that constant value once the engine is above the speed at which the pump can supply more than the engine needs.
- **Full-flow system.** There is no relief valve holding a set pressure in normal running; the whole of the pump's output is delivered to the engine. Because a positive-displacement pump delivers in proportion to

its speed, the resulting oil pressure **varies with engine speed and therefore with the throttle setting** rather than sitting at a single fixed value. A full-flow system's oil pressure indication is expected to rise and fall as the power is changed, and it is read against a schedule rather than against one number.

Filtration is arranged in layers. A main pressure filter takes out the bulk of the contamination and carries a differential-pressure indicator that pops when the element is approaching bypass, so a clogging filter is found before it starts passing unfiltered oil. Downstream of it, small **last-chance filters** sit immediately upstream of the oil jets, whose orifices are small enough to be blocked by a single particle.

Scavenge system detail

Oil is not returned by gravity — an engine has no consistent "down" and the chambers are pressurised — so each sump has its own scavenge pump, all driven from the accessory gearbox. The total capacity of the scavenge pumps is deliberately made **greater** than the delivery of the pressure pump. The reason is that scavenged oil has been thrown about at high speed inside the chamber and is heavily aerated, so it occupies considerably more volume than the same mass of oil did on the way in. A scavenge system sized to the pressure pump's volume would flood the chambers.

Magnetic chip detectors in the scavenge lines are the primary early warning of internal mechanical distress, and their position matters: a detector in one chamber's scavenge line implicates that chamber's bearings, which is a great deal more useful than a single detector in the common return.

Breather system detail

Bearing chambers are sealed with air, not with oil-tight seals, and that air has to go somewhere. Compressor air at a higher pressure than the chamber leaks inwards past the labyrinth seals, which is exactly what stops oil leaking outwards. That inward leakage would steadily pressurise the chamber, and a pressurised chamber would force oil past its own seals into the gas path. The breather system vents the chambers to keep their internal pressure below the sealing air pressure, and a centrifugal breather in the gearbox — a rotating drum that flings the oil droplets outward and lets clean air pass through the middle — recovers the oil from the vent air before that air is discharged overboard or into the exhaust.

Cooling the oil

Most engines cool the oil with the fuel, in a **fuel-cooled oil cooler**, and the direction of the exchange is worth stating explicitly: the oil *gives* heat *to* the fuel. The oil is cooled, which is the primary purpose, and the fuel is warmed, which is a valuable secondary benefit because warm fuel is far less likely to precipitate ice crystals in the fuel filter. Some installations add an air-cooled oil cooler in the bypass duct, particularly where the fuel flow at low power is not sufficient to absorb the heat load.

Indications and condition monitoring

- **Oil pressure** and a separate **low oil pressure warning**; **oil temperature**; **oil quantity**.
- **Filter impending-bypass** indication — a rising differential pressure across the filter means the element is loading up with contamination.
- **Chip detector** indication, either as a warning in the flight deck or as an item examined at maintenance.

- **Oil consumption trend** — recorded uplift per flight hour, one of the most reliable long-term indicators of seal condition.
- **Spectrometric oil analysis**, where used, which identifies the metals present in the oil in very small quantities and can point to which component is wearing.

Turbine fuels

Grade	Type	Freeze point (max)	Flash point (min)	Notes
Jet A-1	Kerosene	–47 °C	38 °C	The international civil standard, available worldwide.
Jet A	Kerosene	–40 °C	38 °C	Used chiefly in the United States domestic market.
Jet B	Wide-cut (kerosene/gasoline blend)	–50 °C	Much lower — the fuel is volatile	Cold-climate and some military use. Its volatility makes it easier to start and relight in extreme cold but significantly increases the fire and handling risk, so it is not used in normal civil operations.
JP-8	Kerosene	As Jet A-1	As Jet A-1	Military equivalent of Jet A-1, distinguished by its military additive package.
JP-5	High-flash kerosene	–46 °C	60 °C	The deliberately high flash point is for shipboard storage, where a fuel that gives off flammable vapour at ordinary temperatures would be unacceptable.

The properties that matter operationally are:

- **Calorific value** — approximately 43 MJ/kg for all the kerosene grades. This is what sets fuel flow for a given power.
- **Density** — roughly 0.78 to 0.84 kg per litre at 15 °C and falling as temperature rises, which is why an uplift measured in litres must be converted with the actual density, and why fuel is loaded and planned by mass rather than volume.
- **Freeze point** — not the temperature at which the fuel solidifies, but the temperature at which wax crystals begin to appear and can block filters. On a long, high-altitude sector the fuel in the tanks cold-soaks towards ambient, and fuel temperature is a monitored parameter for that reason.
- **Volatility and flash point** — a safety property. A kerosene with a flash point of 38 °C does not produce an ignitable vapour at ordinary ambient temperatures; a wide-cut fuel does.
- **Viscosity at low temperature** — thick fuel atomises badly, which affects starting and relight.
- **Thermal stability** — the fuel is used as a heat sink, so it must not break down and lay down deposits in heat exchangers, filters and burner passages.
- **Lubricity** — the high-pressure pump and the metering unit are lubricated by the fuel passing through them, so a fuel with poor lubricity causes pump wear.
- **Electrical conductivity** — kerosene flowing through pipes and filters generates static, so conductivity is controlled to let the charge relax safely.

The additive package normally includes a **fuel system icing inhibitor** (di-ethylene glycol monomethyl ether, commonly abbreviated to DiEGME), a **static dissipator** to raise conductivity, an **antioxidant**, a **metal deactivator**, a **corrosion inhibitor** which doubles as a lubricity improver, and, where authorised for a specific treatment, a **biocide**.

Water is the enemy of a fuel system. Water enters fuel dissolved and as free droplets, and it comes out of solution as the fuel cools. Free water settles to the low points of the tanks, where three things happen: it freezes into ice crystals that block filters and can restrict flow at exactly the moment the engine needs it, it corrodes tank structure, and it forms an interface at which micro-organisms grow. Microbiological contamination produces a dark slime that blocks filters and probes and causes severe pitting corrosion under the deposit. This is why tank water drains are checked as a routine task, why fuel is checked for water and contamination before uplift, and why biocide treatment is a controlled procedure rather than a precaution to be applied at will.



Free water settles to the tank low points, so the tank water drains are checked as a routine task.

The engine fuel system

Tracing the fuel from the aircraft tank to the flame gives the whole system in order:

- **Aircraft boost pumps** deliver fuel to the engine at a low positive pressure.
- **Low-pressure (LP) engine fuel pump** — usually a centrifugal pump driven from the accessory gearbox.
- **LP filter**, with an impending-bypass indication and a bypass valve.
- **Fuel-cooled oil cooler**, where the fuel picks up heat from the engine oil.
- **High-pressure (HP) fuel pump** — a positive-displacement pump, most commonly a gear pump, delivering considerably more fuel than the engine can ever need.
- **Fuel metering unit** (also called the fuel control unit or hydromechanical unit), which meters the required flow to the engine and spills the surplus back to the pump inlet.
- **Fuel flow transmitter**, providing the flight-deck indication.
- **Pressurising and dump valve**.

- **Fuel manifold and spray nozzles**, delivering an atomised spray into the primary zone of the combustion chamber.

The **purpose of the LP pump** is the point most often tested, and it is not simply to move fuel. Its function is to raise the pressure at the *inlet of the HP pump* comfortably above the vapour pressure of the fuel, so that the HP pump cannot **cavitate**. Without it, at high altitude or with warm fuel, the pressure at the HP pump inlet would fall low enough for vapour bubbles to form; those bubbles collapse violently as the pressure rises inside the pump, eroding the pump elements and disrupting the accuracy of the fuel metering that depends on a solid column of liquid. Aircraft tank boost pumps deliver fuel to the engine, and the metering unit governs the flow, but neither of those is the LP pump's defining purpose.

The **pressurising and dump valve** does two jobs. In running, it holds a minimum pressure in the fuel system so that the metering elements and any pressure-operated servos work correctly even at the very low flows demanded at idle and during a start. On shutdown, it dumps the fuel remaining in the manifold, so that residual fuel is not left to be baked in the hot burners and turned to carbon — coke in a spray nozzle ruins its spray pattern and burns turbine components downstream.

Spray nozzles exist in several forms. A *simplex* nozzle has a single swirl chamber and orifice, which gives a good spray only over a narrow flow range. A *duplex* nozzle adds a second, larger passage that opens above a certain pressure, so the primary passage handles starting and idle flows and the main passage takes over at higher power. An *airspray* (air-blast) nozzle uses compressor air to shear the fuel film into fine droplets, which gives good atomisation across the whole range at low fuel pressure and produces markedly less carbon. *Vaporising* systems take a different approach again, passing the fuel through a heated tube inside the flame so that it enters the primary zone already vaporised.

Fuel control and the limits it enforces

On a modern engine the fuel metering unit is commanded by a **FADEC**: an electronic control unit that reads thrust lever position, rotor speeds, temperatures and pressures, and computes the fuel flow required, driving a torque motor or servo valve in the metering unit. It is powered in flight by its own permanent-magnet alternator on the accessory gearbox, so that the engine's control is independent of the aircraft electrical system.

The control's most important job, beyond setting steady-state power, is limiting how fast the fuel flow may change. Two schedules do this and the directions are opposite:

- **Acceleration (maximum fuel) schedule.** If fuel is added faster than the rotor can accelerate, the turbine entry temperature and the back-pressure on the compressor rise ahead of the compressor's rotational speed. The compressor is then asked for more pressure rise than it can deliver at that speed, and it surges — and the same excess fuel simultaneously threatens an overtemperature. The schedule caps the rate of increase.
- **Deceleration (minimum fuel) schedule.** If fuel is cut back too quickly, the mixture in the primary zone becomes too weak to sustain combustion and the engine flames out. The schedule sets a floor below which the fuel flow may not be reduced.

This is also the mechanism behind a statement made earlier: on a gas turbine, **power is adjusted by varying the fuel flow**, and nothing else. The airflow increase that accompanies a power increase is the consequence of the spool accelerating, not a separate control input.

The fuel system has a second job. As well as feeding the engine, the fuel is the main heat sink for the engine oil and, on many aircraft, for the integrated drive generator oil as well. That coupling explains several things that otherwise look unconnected: why fuel temperature is an indicated parameter, why an oil cooler problem can show up as a fuel temperature indication, and why fuel filter icing is a design concern at all — the heat exchanger exists partly to prevent it. It is also why fuel thermal stability is a specification property rather than a laboratory curiosity.

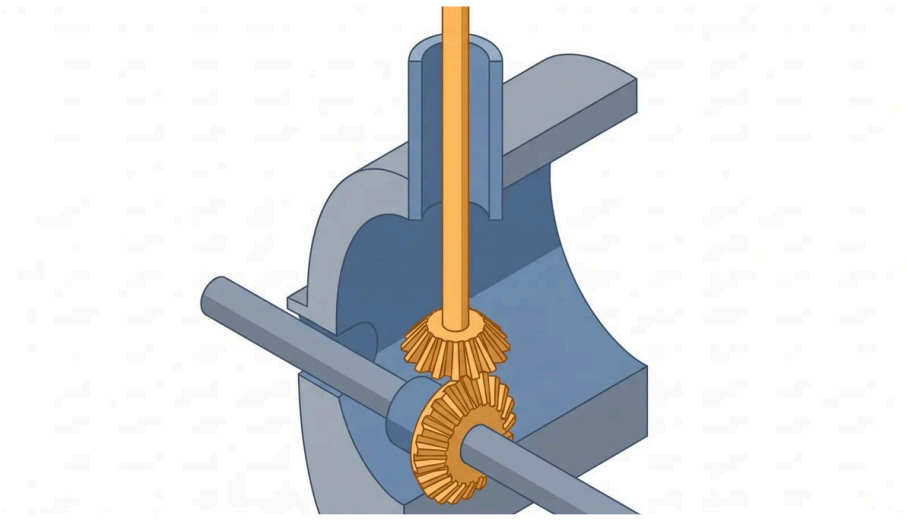
Accessory Gearbox, Bearings, and Seals

- **Accessory gearbox (AGB):** Driven by the HP shaft via a tower shaft/bevel gears. Provides mounting and drive for: fuel pump, oil pump, hydraulic pump, electrical generator, starter, speed sensors.
- **Bearings:** Ball bearings (thrust loads) and roller bearings (radial loads) support the shafts. Fed by the oil system.
- **Seals:** Labyrinth seals (non-contact, uses close-tolerance fins) and carbon seals prevent oil from leaking into the gas path and hot gas from entering bearing chambers.

Why the accessory gearbox is driven from the high-pressure shaft

Taking the accessory drive from the HP spool rather than the LP spool is a deliberate choice for three reasons, and the first of them is the one most often overlooked. The **starter drives the engine through the accessory gearbox**, and the shaft that has to be turned to start a gas turbine is the high-pressure spool — it is the HP compressor that must reach a minimum speed before fuel and ignition can be introduced. The drive train therefore has to connect the starter to that spool, and once it does, it is the natural place to hang everything else. Second, the HP spool's speed varies over a much narrower range between idle and maximum than the LP spool's does, which suits accessories that need a reasonably consistent drive speed. Third, motoring the engine for a dry or wet motoring run, or for cooling, is done on the same shaft through the same path.

The mechanical arrangement is standard across most large engines. An **internal gearbox** sits around the HP shaft inside the compressor casing, driven by a bevel gear on the shaft. A **radial or tower drive shaft** takes the drive outward through a hollow compressor casing strut to an intermediate or angle gearbox, and from there a horizontal shaft carries it to the **external accessory gearbox** mounted on the fan case or the underside of the core.



A bevel gear on the HP shaft turns a radial tower shaft that runs outward through a hollow casing strut.

Beyond the accessories already listed, several items on a modern gearbox are worth knowing about specifically:

- A **permanent magnet alternator** that supplies the FADEC with its own electrical power, so engine control does not depend on the aircraft's electrical system.
- The **centrifugal breather (deoiler)**, which recovers oil from the vent air before it is discharged.
- The **starter drive**, usually through a pad that also carries the starter's own clutch or shaft, so the starter is disconnected or freewheels once the engine is self-sustaining.
- **Shear sections** or torque-limiting necks machined into individual accessory drive shafts. If an accessory seizes, its own drive shaft shears at a designed weak point and the failure is confined to that accessory instead of destroying the gearbox and disabling the fuel and oil pumps with it.
- **Quick attach-detach (QAD)** mountings and V-band clamps on many accessories, so a line replacement can be made without disturbing rigging.

The gearbox is lubricated by the engine oil system and has its own scavenge pump; it is not a separately serviced unit on most engines.

Bearing arrangement: one bearing locates, the others let the shaft grow

Each spool is supported on at least two bearings, and their roles are deliberately different. One bearing on each shaft is the **location bearing**: a ball bearing, whose inner and outer races both capture the balls axially, so it fixes the shaft's axial position and reacts the net axial gas load acting on the compressor and turbine. The remaining bearings are **roller bearings**, whose rollers can slide along a plain race, so they carry radial load only and permit the shaft to grow and shrink axially with temperature.

That division is not a refinement, it is a necessity. A gas turbine shaft several metres long changes length measurably between a cold start and stabilised take-off power. If both ends of the shaft were located axially, differential expansion between the shaft and the surrounding casings would put a very large axial preload into the bearings and destroy them. Fixing one end and letting the other float removes the problem completely.

The choice of anti-friction bearings — ball and roller — over plain or sintered types follows from the duty. Shafts run at very high rotational speeds under substantial radial and axial loads, and plain bearings would generate unacceptable friction and heat and would depend entirely on a hydrodynamic film that could not be maintained at start-up and shutdown. Ball and roller bearings have low friction, are tolerant of the wide speed range, and are cooled by the same oil that lubricates them.

Other features an engineer meets: bearings are numbered by station from the front of the engine, which is how a chip detector or a vibration signature is traced to a particular location; **squeeze-film damping**, an oil film deliberately maintained around the bearing outer race in its housing, absorbs shaft vibration and accommodates small amounts of shaft movement; and some engines use an **inter-shaft bearing**, running directly between the LP and HP shafts rather than between a shaft and the casing, which saves a bearing chamber at the cost of a bearing whose relative speed is the difference between two rotor speeds.

Bearing chambers, oil and sealing air — and which way the leakage goes

A bearing chamber is a small sealed compartment containing the bearing, fed by oil jets, drained by its own scavenge pump, sealed by air and vented by the breather system. The sealing arrangement contains a direction that must be stated correctly, because reversing it makes nonsense of the whole design.

The air outside the chamber is at a higher pressure than the air inside it. Air bled from the compressor is delivered to the outside of the seals, and the breather system holds the chamber's internal pressure below that. The pressure difference therefore drives a small, continuous leakage of air *inwards*, through the seals, into the chamber. That inward flow is exactly what prevents oil escaping outwards into the gas path — oil cannot travel against it. This is why a breather system blockage is serious: if the chamber cannot vent, its pressure rises towards the sealing air pressure, the pressure difference across the seal collapses, and oil begins to leak out.

Seal types

Seal type	Construction	Contact?	Leakage	Where used
Labyrinth	A series of fine fins on the rotating member running at a close, controlled clearance against a static, often abradable, land.	Non-contact	A small, deliberate and controlled leakage — the seal works *because* air flows through it	The standard bearing chamber and gas path seal throughout the engine; tolerant of high speed and temperature.
Carbon (face or ring)	A carbon element spring-loaded against a polished rotating face or shaft.	Contact	Very low, close to zero	Where oil loss must be minimised, typically at gearbox and accessory drives; limited by rubbing speed and by the heat generated at the face.
Brush	A dense pack of fine wire bristles bearing lightly on the shaft.	Light contact	Much lower than a labyrinth of the same size	Used in place of, or alongside, labyrinth seals where a tighter seal is wanted without the temperature limitations of carbon.

Hydraulic	A rotating annulus of oil trapped by centrifugal force in a groove, sealing against a static fin dipping into it.	Non-contact	None while the seal is spinning; the seal is lost when it stops	Found on older engines; effective only while the shaft is turning.

Failure modes and what the engineer actually sees

- **Rising oil consumption.** Usually the first symptom of seal deterioration. The trend matters more than any single figure, which is why uplift per flight hour is recorded.
- **Oil in the bypass duct, or oil mist from the exhaust or the drains mast.** Evidence that oil is escaping from a chamber into the gas path, which points at a seal or at a chamber vent problem.
- **Oil smell in the cabin air.** Bleed air is taken from the compressor, so a bearing seal that is passing oil into the compressor can put oil decomposition products into the air conditioning supply. It is an airworthiness matter and it is also a crew and passenger health complaint; it must be investigated, not written off as a nuisance odour.
- **Chip detector captures.** The maintenance manual classifies the debris found: fine "fuzz" or a light smear may be normal running wear, flakes suggest bearing race or gear tooth spalling, and larger fragments indicate a failure in progress. Follow the classification and the associated inspection; do not judge a capture by eye and clear it.
- **Vibration signature change.** A change in the vibration level at a particular rotor's frequency localises the problem to that spool, and a shift in phase can indicate a shifting or damaged component.
- **Scavenge oil temperature rise at one sump.** An unusually hot return from one chamber points directly at that chamber's bearing.
- **Bearing chamber fire.** The chamber contains an oil mist in a hot environment; a loss of oil supply or a bearing failure can generate enough heat to ignite it. This is why chamber temperatures, oil pressures and chip warnings are treated as urgent rather than advisory.

A chip detector indication is not a nuisance item. The temptation to wipe a detector, refit it and see whether it comes back is understandable and wrong. Anti-friction bearings and gear teeth do not shed metal in significant quantity until something has begun to fail, and the interval between the first flakes and a bearing collapse can be short. The maintenance manual sets out how the debris is to be examined, classified and reported, and what inspection follows; that procedure is the only acceptable response.

Module 15 — Gas Turbine Engine

15.1 — Fundamentals

The gas turbine engine is the powerplant of virtually every modern commercial and military aircraft. Understanding its fundamental principles — the conversion of chemical energy in fuel into useful thrust or shaft power — is the cornerstone of all subsequent engine study. This section covers the physics that make gas turbine engines work, and the four main engine configurations you will encounter in aviation maintenance.

Everything that follows rests on one idea: a gas turbine is an **energy converter**, and the substance it works on is **air**. Air is the *working fluid* — the medium that is drawn in, compressed, heated, expanded and thrown out again. Kerosene is the *fuel*: it is the source of the energy, but it is burned *in* the working fluid rather than being the working fluid itself. The distinction is not a quibble, because it explains the proportions inside the engine. Air enters a large civil engine at hundreds of kilograms per second while fuel enters at a few kilograms per second, so the fuel adds less than about two per cent to the mass passing down the gas path. Almost everything the engine handles, accelerates, heats and finally ejects is air.

The conversion runs in a fixed order, and that order is itself examined. **Chemical energy** is locked in the molecular bonds of the fuel. Burning releases it as **heat energy**, which raises the temperature — and therefore the internal energy — of the compressed air passing through the combustor. That hot, high-pressure gas is then expanded, which converts part of its heat and pressure energy into **kinetic energy** of fast-moving gas and part into **mechanical work** at a turbine shaft. Finally the kinetic energy of the departing jet, or the mechanical work delivered to a propeller or a rotor, appears as **propulsive thrust**. The chain runs chemical to heat to kinetic to mechanical and thrust, and it never runs the other way: nothing in the engine turns kinetic energy of the intake air back into chemical energy, and nothing drives the compressor from an external electrical supply once the engine is running.

No step in that chain creates energy. The principle of conservation of energy holds throughout: the energy entering the engine as fuel, plus the energy the incoming air already carries, exactly equals the energy leaving it. What leaves is a mixture of useful output — thrust or shaft power — and unavoidable losses: hot exhaust gas carrying heat away, residual kinetic energy in the jet that is never used, noise, and heat rejected into the oil system and the bypass air. A gas turbine is not a perfect machine and cannot be, so the question an engineer actually asks is never whether energy is lost but *where* it goes, how much of it can be recovered as useful work, and what a change in that balance is telling us about the engine's health.

Energy, Force and Newton's Laws

Potential Energy

Potential energy is the energy a body possesses because of its position or state. In gas turbine theory, we are mainly concerned with two forms:

- **Gravitational potential energy** — energy due to height above a reference datum. An aircraft at altitude has potential energy $E_p = mgh$, where m is mass (kg), g is gravitational acceleration (9.81 m/s²), and h is height (m).
- **Chemical potential energy** — energy stored in the molecular bonds of aviation fuel. When kerosene (Jet A-1) burns, each kilogram releases approximately **43 MJ** of heat energy. This is the primary energy input to every gas turbine engine.

A third form matters inside the engine itself, and it is easy to overlook because nothing is being lifted: **pressure energy**. Air that has been squeezed by the compressor is in a raised *state* rather than a raised *position*, but it holds energy in exactly the same sense — it will do work on anything that lets it expand. Potential energy is therefore best defined as energy stored by virtue of **position or condition**, a definition wide enough to cover height, chemical state and pressure alike. Compressor delivery air on a large engine sits at forty or more times atmospheric pressure and several hundred degrees Celsius, and it is that stored pressure energy, topped up by the heat of combustion, that the turbine converts back into work a short distance further downstream.

A fourth form appears constantly on the flight line without ever being called potential energy: **elastic strain energy**, stored in a component that has been deliberately deformed. A torqued bolt is a stretched spring, and the clamping force holding a flanged joint together is the bolt trying to return to its original length. That is why a fastener tightened by an angle-of-turn or bolt-stretch method is controlling stored strain energy directly rather than relying on the friction relationship that a torque wrench measures, and why a bolt that has been stretched past its yield point can look perfectly serviceable while storing far less energy than the joint needs.

Worked Example — How much energy is in the fuel flow?

A large high bypass turbofan at take-off burns of the order of **3.5 kg of fuel per second**. Each kilogram of kerosene releases approximately 43 MJ, so the rate at which chemical energy is being released is:

$3.5 \times 43 = 150.5$ MJ every second, that is **150.5 MW**.

For scale, one engine at take-off power is releasing heat at roughly the rate of a medium generating unit in a power station, and the aircraft has at least two of them. Only a fraction of that ends up as useful propulsion; the majority leaves as hot gas. This is the arithmetic behind two familiar procedures — the cooling period at idle before shutdown, which lets the hot section give up its stored heat gradually instead of soaking it into the bearing housings, and the size of the exhaust hazard area behind a running engine.

Gravitational potential energy is easy to dismiss as classroom material, but it is the reason an aircraft can trade height for distance, and putting a number on it is instructive. An aircraft of mass 70,000 kg cruising at 11,000 m holds $E_p = mgh = 70{,}000 \times 9.81 \times 11{,}000$, which is approximately **7,554 MJ** relative to sea level. Divide that by the 150.5 MW calculated above and the entire height of the aircraft is worth about **fifty seconds** of one engine at take-off power. Far more energy sits in the fuel tanks than in the aeroplane's altitude, which is why a drift-down after an engine failure buys range but never buys very much of it.

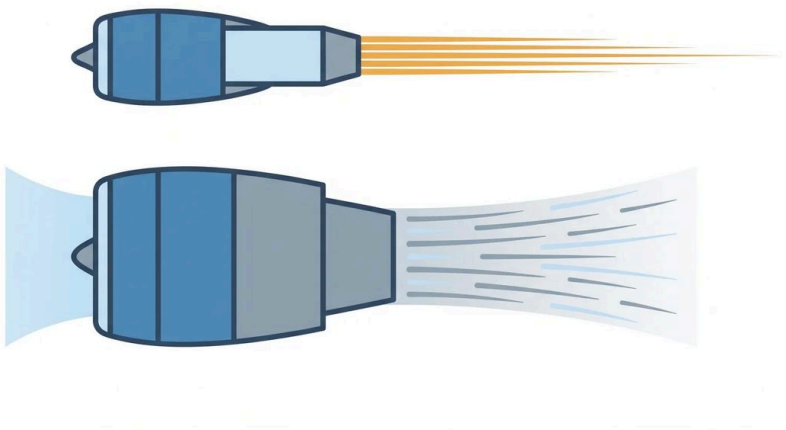
Chemical potential energy deserves one more remark, because its sheer density is the reason aviation looks the way it does. At roughly 43 MJ per kilogram, kerosene stores something like fifty times as much energy per kilogram as a typical rechargeable battery cell. An engine that can convert that energy at high power-to-weight is therefore not merely convenient; it is what makes long-range flight possible at all, and it is the obstacle every electric and hybrid propulsion concept has to confront.

Kinetic Energy

Kinetic energy is the energy a body possesses because of its motion. For a mass m moving at velocity v:

Kinetic Energy: $E_k = \frac{1}{2}mv^2$

In a gas turbine engine, kinetic energy appears in the high-velocity exhaust gases leaving the nozzle. A turbojet produces thrust almost entirely by accelerating a relatively small mass of air to a very high velocity. A turbofan, by contrast, accelerates a much larger mass of air to a lower velocity — this is more efficient because kinetic energy increases with the *square* of velocity, meaning doubling the mass at the same velocity gives twice the momentum (thrust) but only twice the energy, while doubling the velocity at the same mass gives twice the momentum but *four times* the energy cost.



A turbojet throws a small mass of air out very fast; a turbofan moves a much larger mass more slowly.

Kinetic energy is not confined to the jet pipe, because the gas is moving everywhere inside the engine. Air arrives at the compressor face already travelling at well over a hundred metres per second; it is accelerated by every rotor row and decelerated by every stator row; it is slowed sharply in the pre-combustor diffuser so that a flame can be held in it; and it is accelerated hard again through the turbine nozzle guide vanes. What the engine does, over and over, is trade kinetic energy against pressure energy. Useful output only appears at the

very end, when the last of those trades leaves the gas travelling rearwards faster than it was travelling when it arrived.

Worked Example — Same thrust, very different energy bill

Two engines each produce **30 kN** of thrust on a static test bed, where the incoming air velocity is effectively zero. Thrust is mass flow multiplied by the velocity change given to that flow:

- **Engine A** passes $\dot{m} = 50$ kg/s and accelerates it to 600 m/s. Thrust $= 50 \times 600 = 30{,}000$ N.
- **Engine B** passes $\dot{m} = 300$ kg/s and accelerates it to 100 m/s. Thrust $= 300 \times 100 = 30{,}000$ N.

Now compare the kinetic energy each one pours into its jet every second, $\frac{1}{2}\dot{m}v^2$:

- Engine A: $\frac{1}{2} \times 50 \times 600^2 = 9{,}000{,}000$ W, that is **9 MW**
- Engine B: $\frac{1}{2} \times 300 \times 100^2 = 1{,}500{,}000$ W, that is **1.5 MW**

Identical thrust, but Engine A spends six times as much energy per second buying it. That factor of six is exactly the ratio of the two jet velocities, 600 divided by 100 — which is no coincidence, because for the same thrust the energy rate is proportional to the velocity change. The whole commercial case for the high bypass turbofan is contained in this one comparison.

The energy Engine A spends extra is genuinely wasted, and it is worth being precise about where it goes. Kinetic energy left in the departing jet does no work on the aircraft; it is simply handed to the atmosphere and stirred away as turbulence and heat. The standard measure of how much of the energy actually reaches the aircraft is **propulsive efficiency**, which for a simple jet is:

Propulsive Efficiency: $\eta_p = \dfrac{2V_0}{V_0 + V_j}$

Where V_0 is the aircraft's forward speed and V_j the jet velocity, both in the same units.

Put the two engines above on an aircraft flying at 250 m/s. Engine A adds 600 m/s to the flow, so $V_j = 850$ m/s and $\eta_p = 500/1100 = 0.45$, about **45%**. Engine B adds only 100 m/s, so $V_j = 350$ m/s and $\eta_p = 500/600 = 0.83$, about **83%**. Propulsive efficiency therefore *rises* as the jet velocity is brought *closer* to the aircraft's own speed. It would reach 100% only if the jet were left at exactly aircraft speed — which would also mean no velocity change and therefore no thrust at all. Every real engine is a compromise between those two ends, and where a designer sits on that compromise is decided by the speed the aircraft is meant to fly at.

Two distinct quantities are being traded against one another here, and naming them makes the trade easier to hold on to. The useful output is a change of **momentum**, mv; the bill is paid in **kinetic energy**, $\frac{1}{2}mv^2$. One of them is linear in velocity and the other quadratic, and that mismatch is the whole of the argument for moving more air gently rather than less air hard. Every change in engine architecture since the 1960s — larger fans, higher bypass ratios, geared fan drives, open rotors — is an attempt to move further along that curve, and the thing that stops it going further each time is not the thermodynamics but the weight, diameter and drag of the fan needed to do it.

⚠ **Why Impact Damage Rises So Steeply With Speed:** Because kinetic energy varies with the *square* of velocity, closure speed dominates impact damage far more than mass does. A 1.8 kg bird striking at 90 m/s arrives with $\frac{1}{2} \times 1.8 \times 90^2 = 7{,}290$ J. The same bird at 150 m/s arrives with $\frac{1}{2} \times 1.8 \times 150^2 = 20{,}250$ J — the speed has risen by two thirds but the energy has almost tripled. This is why fan containment casings are so heavy, why bird ingestion certification testing is carried out at high fan speeds rather than at idle, and why a strike reported at climb speed justifies a far more searching inspection than the same species reported during taxi.

Newton's Laws of Motion

All three of Newton's Laws are fundamental to understanding gas turbine thrust production:

Law	Statement	Engine Application
1st Law (Inertia)	A body remains at rest or in uniform motion unless acted upon by an external force.	Air entering the intake is at rest relative to the surrounding atmosphere; the compressor must do work (apply force) to accelerate it rearward.
2nd Law (F = ma)	Force equals the rate of change of momentum: $F = \dot{m} \times \Delta v$	Thrust is the product of **mass flow rate** ($\dot{m}$, kg/s) and the **change in velocity** of the air/gas through the engine.
3rd Law (Action–Reaction)	For every action, there is an equal and opposite reaction.	Hot gases are expelled rearward (action); the engine — and therefore the aircraft — is pushed forward (reaction). This is the most intuitive explanation of jet thrust.

Reading the Three Laws Inside a Real Engine

The **First Law** is usually taught as a statement about air, but on the flight line it bites hardest as a statement about *rotating mass*. A large fan and its shaft represent a very substantial rotational inertia, and inertia resists a change of motion in both directions equally. That is why a big turbofan takes several seconds to accelerate from flight idle to go-around thrust: the fuel schedule cannot simply be opened, because the gas generator has to be accelerated against its own inertia while keeping enough compressor surge margin to survive doing it. It is why the same engine takes minutes rather than seconds to run down after shutdown. And it is why free rotation of the compressor is confirmed before a start is attempted — a rotor that will not turn freely is telling you something inside the engine has changed, and the starter has no way of telling you gently.

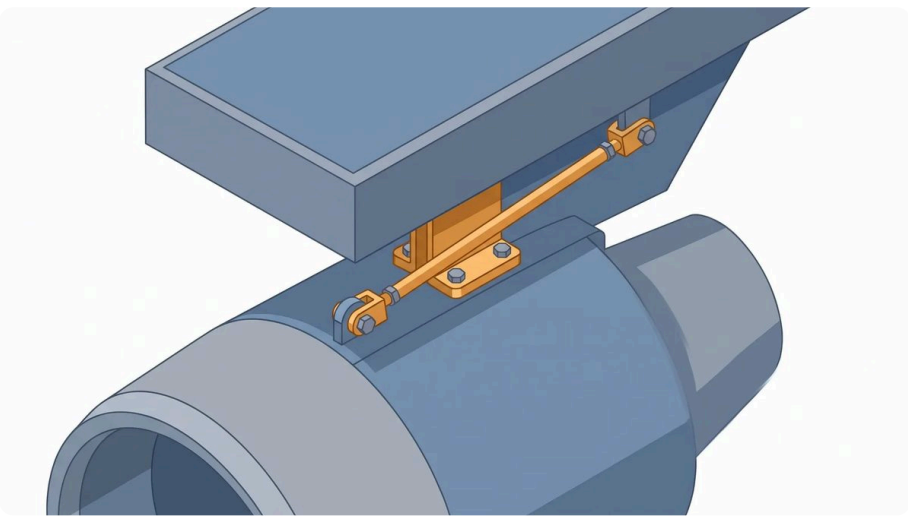
The **Second Law** is the one that produces the numbers. In its schoolroom form it is $F = ma$: a force applied to a mass produces an acceleration. An engine, however, does not accelerate one fixed lump of air; it accelerates a continuous stream of it. Written for a steady flow it takes the form already given in the table above, $F = \dot{m}\,\Delta v$ — mass flow rate multiplied by the change in velocity. The two statements are not rivals and neither is an approximation of the other: $\dot{m}$ is mass per second and Δv is the velocity change that mass receives, so their product is still mass multiplied by acceleration, merely accounted for per second rather than

per lump. Both phrasings appear in examination questions as the correct description of thrust, and a candidate who has understood that they are the same statement will recognise either.

Only two levers therefore exist for producing thrust: raise the **mass flow**, or raise the **velocity change**. Every configuration described later in this note is a different answer to the question of how to divide the work between those two. Raising mass flow means a bigger engine or a bigger fan, which costs diameter, weight and nacelle drag. Raising velocity change means burning more fuel per newton, for the reasons the worked example above set out. There is no third lever, and no engine escapes the trade.

The **Third Law** is the one that repays careful reading. Thrust is the reaction to accelerating a mass of gas rearwards. It is *not* the exhaust pushing against the outside air, and it is not the exhaust pushing against the ground: a rocket works perfectly in a vacuum, and an engine on a test stand produces its rated thrust with nothing behind it but open air. The reaction force appears on the engine's own internal surfaces — on the fan and compressor blades, on the combustor casing, on the turbine nozzle guide vanes, on the internal walls of the intake — as the pressure differences that accelerate the gas act back on the hardware creating them. Add every one of those internal loads together and the resultant is the forward force we call thrust. The same correction applies to a propeller: a turboprop's thrust is the reaction to the prop-wash being accelerated rearwards, not the prop-wash striking the air behind it. In examination wording, *impingement* on the surrounding air is always the wrong answer.

Because that resultant is distributed over the engine's internal structure, it has to be collected and delivered to the airframe. It travels through the engine casings into the **thrust links** or thrust-carrying mount, and from there into the pylon and the wing or fuselage structure. That is the practical reason engine mount hardware is inspected for cracking, torqued to specification and, on many types, life-limited: the mounts are the only path by which the reaction force reaches the aircraft. A turboprop or turboshaft installation adds a second reaction to restrain — the **torque reaction** from driving the propeller or rotor, which tries to rotate the engine and its gearbox in the direction *opposite* to the driven shaft, and which the mounts and torque struts have to hold.



Thrust links and the mount join the engine casing to the pylon: the only path thrust has to the aircraft.

Rotating mass produces one further reaction that has to be designed for, and it is a direct consequence of the First Law applied to rotation. A spinning rotor resists any attempt to change the *direction* of its axis, and if it is forced to change direction it responds with a couple at right angles to the applied one. Every engine rotor is therefore a gyroscope, and when the aircraft pitches or yaws, the mounts have to react gyroscopic loads on top of thrust and weight. This is one of the reasons an engine mounting system is a more complicated structure than it looks, why manoeuvre loads appear in mount fatigue calculations, and why the certification case for an engine includes the loads produced by a windmilling rotor that has become unbalanced after a blade release, when both the out-of-balance force and the gyroscopic couple are at their worst.

✈ **Aviation Context:** All three laws are visible in a single ground run. Nothing turns until the starter applies a torque, and the rotor's inertia decides how long the first part of the start takes (First Law). The rate at which the engine then accelerates depends on the surplus torque available above what the compressor is absorbing (Second Law). And the aircraft strains against its chocks and brakes because the reaction to gas leaving the jet pipe is pushing the airframe forwards (Third Law). If the aircraft creeps during a high-power run, the physics is behaving exactly as it should and the fault lies with the restraint.

Basic Thrust Equation: $F_N = \dot{m}(V_j - V_0) + A_e(P_e - P_0)$
Where: F_N = net thrust, $\dot{m}$ = mass flow rate, V_j = jet exit velocity, V_0 = aircraft forward velocity (intake velocity), A_e = nozzle exit area, P_e = exit pressure, P_0 = ambient pressure. For a fully expanded nozzle (one whose exit pressure has fallen to ambient), the pressure term is approximately zero and thrust simplifies to $F_N = \dot{m}(V_j - V_0)$.

Reading the Thrust Equation Term by Term

That equation repays being taken apart slowly, because almost everything in engine performance is a consequence of one of its four quantities. The first part of the first term, $\dot{m}V_j$, is the **momentum thrust**: the rate at which momentum leaves the nozzle. The subtracted part, $\dot{m}V_0$, is the **momentum drag**: the air already possessed momentum before it reached the engine, and the engine only earns credit for the momentum it *adds*. The final term, $A_e(P_e - P_0)$, is the **pressure thrust**: if the gas is still above ambient pressure as it crosses the exit plane, that excess pressure acting over the exit area contributes force as well.

Momentum drag is the whole reason thrust and airspeed are connected. On a static ground run V_0 is zero, none of the momentum thrust is given back, and the reading on the test-bed load cell is the largest number the engine will ever produce at that setting. Push the same engine forward through the air and V_0 grows, the subtraction grows with it, and net thrust falls even though nothing inside the engine has changed.

Worked Example — The same engine, static and in flight

Take an engine passing $\dot{m} = 200$ kg/s and leaving a fully expanded nozzle at $V_j = 400$ m/s, so the pressure term is zero.

- **Static, $V_0 = 0$:** $F_N = 200 \times (400 - 0) = 80{,}000$ N, that is **80 kN**.
- **In flight at $V_0 = 250$ m/s:** $F_N = 200 \times (400 - 250) = 30{,}000$ N, that is **30 kN**.

The momentum drag term has taken away $200 \times 250 = 50$ kN. In practice the loss is not quite this severe, because forward speed also rams air into the intake and raises the mass flow the engine can swallow, which pushes thrust back up. The two effects compete, and which of them wins depends on the flight regime. At the low and moderate subsonic speeds a transport aircraft actually uses, the momentum drag term dominates and net thrust falls as airspeed rises. At high Mach numbers ram compression becomes so effective that it first arrests and then reverses the decline — which is precisely why a turbojet remains a sensible choice for supersonic flight while a high bypass turbofan does not.

The pressure term needs a word of explanation because it is often quietly ignored. A simple convergent nozzle can only accelerate gas up to the local speed of sound at its throat; once that happens the nozzle is **choked** and no further increase in velocity is possible there, however much pressure is available upstream. Surplus pressure therefore has nowhere to go and simply arrives at the exit plane still above ambient, where it acts over the exit area and contributes force. Most civil engines run choked at take-off and climb power, so the pressure term is producing real thrust rather than being a mathematical curiosity.

Worked Example — What is pressure thrust worth?

A nozzle of exit area $A_e = 0.8$ m² is choked, with an exit static pressure of 130 kPa against a sea level ambient of 101.3 kPa:

$A_e(P_e - P_0) = 0.8 \times (130{,}000 - 101{,}300) = 0.8 \times 28{,}700 = 22{,}960$ N, approximately **23 kN**.

Added to a momentum thrust of, say, 80 kN, that is more than a fifth of the total. Ignoring it would make the engine look substantially weaker than it is. It also explains why the exhaust nozzle exit area is a controlled dimension: distortion, an unapproved repair, a badly seated exhaust cone or a burned-back nozzle lip all change A_e, and changing A_e changes both the pressure thrust and the whole running line of the engine behind it.

Check the Units: A quick unit check catches most arithmetic errors in this equation. Mass flow is kg/s and velocity is m/s, so $\dot{m}\,\Delta v$ has units of kg m/s², which is the definition of the newton. Area is m² and pressure is N/m², so $A_e(P_e - P_0)$ is also newtons. Both terms are forces, which is why they may be added directly. If a calculation produces anything other than newtons, a quantity has been entered in the wrong units — most often a mass flow given in kg per minute, or a pressure given in bar or psi rather than pascals.

Air, the Gas Laws and Bernoulli's Theorem

Newton's laws explain what thrust *is*. The gas laws explain what the engine has to do to the air in order to produce it. Every pressure rise across a compressor stage, every temperature reading on a flight-deck gauge, and every performance shortfall on a hot day at a high-altitude airfield is a consequence of the behaviour set out below. It is also the material that turns a memorised order of the gas path into an explanation of why it has to be in that order.

Air as the Working Fluid

Dry air is approximately 78% nitrogen and 21% oxygen by volume, with argon, carbon dioxide and traces of other gases making up the remainder, plus a variable quantity of water vapour. Only the oxygen takes part in combustion — about a fifth of the air by volume — and even that is only partly consumed, because far more air is drawn in than is needed to burn the fuel. The surplus oxygen and all of the nitrogen are not passengers. They are the mass that carries the heat away from the flame, the mass that is expanded through the turbine, and the mass that is finally accelerated to make thrust. That is why the correct answer to "what is the working fluid of a gas turbine?" is **air**, and why kerosene, which is merely the energy source, is not.

Over the range of temperatures and pressures a turbine engine works in, air behaves closely enough to an **ideal gas** for the ideal-gas relationships to be used with confidence. Two of its properties appear in almost every calculation and are worth fixing in mind: the specific gas constant for air, $R = 287$ J per kilogram per kelvin, and the ratio of specific heats, γ, which is about **1.4** for cold air and falls to roughly **1.33** for hot combustion gas. That change is not a detail. It is one of the reasons the compressor and the turbine cannot be treated as mirror images of one another, and it is why a turbine handling hot gas extracts a different amount of work per unit of pressure ratio than a compressor of the same pressure ratio absorbs.

Mass, Weight and Why the Words Get Mixed Up

Mass is the quantity of matter in a body, measured in kilograms, and it does not change with location. **Weight** is the force that gravity exerts on that mass, $W = mg$, measured in newtons, and it does change with location because g does. A 10 kg component has a mass of 10 kg on the ground, in the air and in orbit; its weight is about 98 N on the ground and effectively nothing in orbit. In Newton's second law and in every gas law it is *mass* that appears, never weight.

Engineering practice nonetheless uses the two words loosely, and the usage is so entrenched that examinations reproduce it — density, for instance, is very commonly defined as "weight per unit volume" when what is strictly meant is mass per unit volume. The two definitions differ only by the constant g, so at any one location they are proportional and no practical error results. Recognise the loose usage rather than fighting it: if a question offers "weight divided by volume" as the definition of density, that is the answer it wants, and the alternatives on offer will be specific volume or something that is not a density at all.

Density, Pressure and Temperature

Density is mass per unit volume — weight divided by volume — written $\rho = m/V$, with the kilogram per cubic metre as its SI unit. Its inverse, volume divided by mass, is *specific volume*; offering volume over mass as the definition of density is a standard distractor and offering pressure over volume is not a density relationship at all. At International Standard Atmosphere sea level conditions, 15°C and 1013.25 hPa, the density of air is **1.225 kg per cubic metre**.

Density is the atmospheric property that matters most to an engine, because a compressor is essentially a volumetric machine: at a given rotational speed it swallows a roughly fixed *volume* of air per second, so the *mass* it passes is proportional to the density of what it is swallowing. Since thrust is proportional to mass flow, anything that reduces density reduces thrust — and both altitude and ambient temperature reduce it, for reasons that come out of a single relationship:

Ideal Gas Equation: $PV = mRT$ or, written in terms of density, $P = \rho RT$
Where P = absolute pressure in pascals, V = volume in cubic metres, m = mass in kilograms, R = 287 J per kilogram per kelvin for air, T = **absolute** temperature in kelvin, and ρ = density in kilograms per cubic metre.

Worked Example — Checking standard sea level density
Standard sea level pressure is 101,325 Pa and standard temperature is 15°C, which is $15 + 273.15 = 288.15$ K. Rearranging $P = \rho RT$ gives $\rho = P/(RT)$:
$\rho = \dfrac{101{,}325}{287 \times 288.15} = \dfrac{101{,}325}{82{,}699} = 1.225$ kg per cubic metre.
The published standard value falls straight out of the equation, which is a useful check that the units were handled correctly. Note that the temperature had to be converted to kelvin first. Using 15 in place of 288.15 would have produced a density roughly twenty times too high — an error large enough to be obvious, but the same mistake made with a 300°C compressor temperature produces an error small enough to look plausible. **Every gas law calculation uses absolute temperature, without exception.**

Two absolute temperature scales are in use and both appear in engine documentation. The **kelvin** is the SI scale and is obtained by adding 273.15 to a Celsius temperature. The **degree Rankine** is its imperial counterpart, obtained by adding 459.67 to a Fahrenheit temperature, and it still turns up in North American engine manuals and performance charts. Both start at absolute zero, which is what makes them valid in the gas laws; Celsius and Fahrenheit do not, and substituting either into a gas law gives a wrong answer.

The Standard Atmosphere and Why Altitude Costs Thrust

Because engine performance depends on the state of the air presented to it, that state has to be defined by agreement before any two engines can be compared. The **International Standard Atmosphere** (ISA) does

that. At sea level it specifies 15°C, 1013.25 hPa and a density of 1.225 kg per cubic metre. Temperature then falls at a lapse rate of about 1.98°C per 1,000 ft, or 6.5°C per kilometre, up to the tropopause at 11,000 m — approximately 36,000 ft — above which the standard temperature stays constant at about –56.5°C. Pressure falls continuously throughout, and much faster than temperature does.

⚛ Worked Example — Air density at cruise altitude

At the standard tropopause, 11,000 m, the pressure is 22,632 Pa and the temperature is 216.65 K. Using $\rho = P/(RT)$:

$\rho = \dfrac{22{,}632}{287 \times 216.65} = \dfrac{22{,}632}{62{,}179} = 0.364$ kg per cubic metre.

That is only $0.364 \div 1.225 = 30\%$ of the sea level value. An engine turning at the same speed swallows about the same *volume* of air per second, so it passes only about **30% of the mass flow** it would pass at sea level. That is the dominant term in the loss of thrust with altitude, and no amount of engine design can recover it.

A competing effect works the other way and is worth naming. The air up there is also very cold, and cold air takes less work to compress for a given pressure ratio, so the cycle itself becomes *more* efficient with altitude. The net result is the familiar one: thrust falls a long way, but the fuel used per newton of thrust improves. The loss of thrust wins on the raw numbers; the gain in efficiency is why aircraft climb to cruise in the first place.

The same reasoning explains the effect of a hot day, but with a caution about how much of the effect the gas laws alone account for. At sea level on an ISA+30 day the temperature is 45°C, or 318.15 K, so $\rho = 101{,}325/(287 \times 318.15) = 1.110$ kg per cubic metre. That is about **9% below** the standard value, and by itself it would cost roughly 9% of the mass flow and therefore of the thrust. The thrust actually lost on such a day is considerably greater than that, because two further effects pile on top of the density loss: hotter air requires more work to compress for the same pressure ratio, and the raised compressor delivery temperature brings the engine up against its turbine temperature limit sooner. Density is the first of three contributors, not the whole story, and the full accounting belongs with engine performance.

Boyle's Law, Charles's Law and the General Gas Law

The ideal gas equation is the general case. The three named laws are simply what it reduces to when one of the three properties is held constant, and each of them describes something real that happens somewhere in the engine.

Law	Held constant	Relationship	Where it is seen in the engine
Boyle's Law	Temperature	$P_1V_1 = P_2V_2$ — pressure and volume are inversely proportional	Charging and discharging a fire-extinguisher bottle or an oxygen cylinder once it has settled back to ambient temperature

Law	Held constant	Relationship	Where it is seen in the engine
Charles's Law	Pressure	$\frac{V_1}{T_1} = \frac{V_2}{T_2}$ — volume is directly proportional to absolute temperature	The combustor, where heat is added at essentially constant pressure and the gas expands in volume instead of rising in pressure
Pressure Law (Gay-Lussac)	Volume	$\frac{P_1}{T_1} = \frac{P_2}{T_2}$ — pressure is directly proportional to absolute temperature	A sealed bottle or accumulator warming through the day; also the constant-volume combustion of a piston engine, which is what the gas turbine deliberately avoids
General (combined) gas law	Nothing	$\frac{P_1 V_1}{T_1} = \frac{P_2 V_2}{T_2}$	Any real process in the gas path, where all three properties change together

Worked Example — Charles's Law

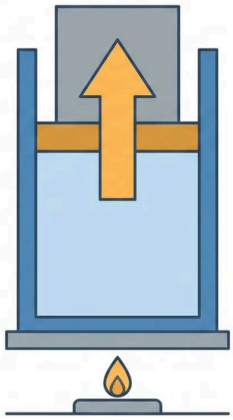
A mass of air occupies 546 cubic feet at 273 K. What volume does it occupy at 274 K, the pressure being held constant?

Charles's law gives $\frac{V_1}{T_1} = \frac{V_2}{T_2}$, so $V_2 = V_1 \times \frac{T_2}{T_1} = 546 \times \frac{274}{273}$.

Because $546 \div 273 = 2$ exactly, every single kelvin of temperature change is worth exactly 2 cubic feet, so $V_2 = 548$ cubic feet — **2 cubic feet greater**. The underlying rule is that a gas at constant pressure changes its volume by $1/273$ of its volume at 273 K for each kelvin of temperature change. Note what has *not* changed: the mass is identical, and only the volume it occupies has altered, so the density has fallen. Nothing was gained or lost by weight.

Specific Heat and the Heat Added in the Combustor

One more property is needed before the cycle can be worked in numbers: the **specific heat capacity**, which is the quantity of heat required to raise one kilogram of a substance by one kelvin. A gas has two of them, because it can be heated either while it is free to expand or while it is held at fixed volume. Heating at **constant pressure** requires more heat, because some of the energy goes into the work of pushing the surroundings back as the gas expands, and its value is written c_p. Heating at **constant volume** requires less, because no expansion work is done, and its value is written c_v. The difference between them is exactly the gas constant, $c_p - c_v = R$, and their ratio is the γ already met above, $\gamma = c_p / c_v$. For air c_p is about 1.005 kJ per kilogram per kelvin, rising to roughly 1.15 for hot combustion gas.



Heated at constant pressure, the gas expands and pushes its piston back, so part of the heat goes into work.

Because a gas turbine adds its heat at constant pressure, it is c_p that governs the combustor, and the heat needed follows directly:

Heat Added at Constant Pressure: $Q = \dot{m}\, c_p\, \Delta T$
Where Q is the heat input rate in kW, $\dot{m}$ the mass flow in kg/s, c_p the specific heat at constant pressure in kJ per kilogram per kelvin, and ΔT the temperature rise in kelvin.

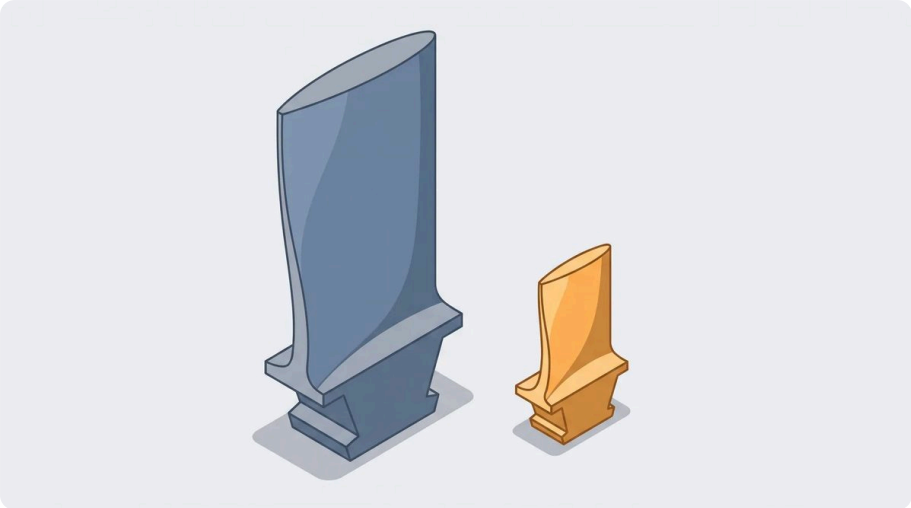
Worked Example — How much heat does the combustor add?
Take a core flow of 100 kg/s entering the combustor at 900 K and leaving at 1,700 K, with $c_p = 1.15$ kJ per kilogram per kelvin for the hot gas:
$Q = 100 \times 1.15 \times (1{,}700 - 900) = 100 \times 1.15 \times 800 = 92{,}000$ kW, that is **92 MW**.
At 43 MJ per kilogram, the fuel flow needed to supply that is $92{,}000 \div 43{,}000 = 2.14$ kg/s. A real engine burns rather more than this, for two reasons that are worth naming: combustion is not perfectly complete, and a significant share of the compressor delivery air is bled away for turbine cooling and sealing and so never passes through the flame at all. The calculation nonetheless gets the order of magnitude right in one line, and it shows clearly why turbine entry temperature and fuel flow are so tightly coupled — asking for 100 K more turbine entry temperature at this mass flow costs about $100 \times 1.15 \times 100 = 11.5$ MW of extra heat, and therefore about a quarter of a kilogram per second of extra fuel.

Continuity — Why the Gas Path Changes Shape

One further relationship governs the physical shape of the engine, and it follows directly from the fact that mass cannot be created or destroyed. In a steady flow, the same mass must pass every cross-section of the duct in every second:

Continuity Equation: $\dot{m} = \rho A V$

Where $\dot{m}$ = mass flow in kg/s, ρ = density in kilograms per cubic metre, A = cross-sectional area in square metres, and V = axial velocity in m/s.



A front-stage compressor blade beside a last-stage blade: the denser air at the rear needs far less area.

Worked Example — Why the compressor annulus tapers

Take a compressor face of annulus area 2.0 m² with air arriving at standard sea level density and 150 m/s axial velocity. The mass flow is:

$\dot{m} = 1.225 \times 2.0 \times 150 = 367.5$ kg/s

Now follow that same air to compressor delivery on a 40:1 engine, where the pressure is $40 \times 101{,}325 = 4{,}053{,}000$ Pa and the temperature is about 901 K. Its density there is $\rho = P/(RT) = 4{,}053{,}000/(287 \times 901) = 15.7$ kg per cubic metre — roughly **thirteen times** the density at the face.

Because $\dot{m}$ and the axial velocity are both held roughly constant along the compressor by design, the area must shrink by about the same factor of thirteen. That is exactly what a compressor annulus does: it tapers steeply from front to rear, the blades getting shorter stage by stage, and the last stages of a large engine carry blades only a few centimetres long. The taper is not styling; it is the continuity equation made out of metal.

Adiabatic, Isentropic and Isothermal Processes

Three words describe the way a gas is squeezed or allowed to expand, and the examination tests all three against one another.

- **Adiabatic** — no heat crosses the boundary of the gas. None is added to it from outside and none is taken away.

- **Isentropic** — adiabatic *and* reversible, that is, adiabatic with no losses. This is the idealised process used in cycle calculations. A real compressor is close to adiabatic but never truly isentropic, because friction, turbulence, tip leakage and shock losses all add entropy.
- **Isothermal** — temperature held constant. For a gas being compressed this can only be achieved by continuously removing the heat of compression as it is generated, which requires a cooling arrangement that an aero-engine compressor does not have and could not carry.

Compression in a gas turbine is treated as **adiabatic**, and the justification is simply speed. The air is through the compressor in a few hundredths of a second, far too quickly for any meaningful quantity of heat to be conducted away through the casings, so essentially all the work done on the air stays in the air. That work raises the internal energy of the gas, and raised internal energy shows itself as raised temperature. **Compressor delivery temperature therefore rises because work has been done on the air** — not because heat has travelled forward from the combustor through the casings, which is negligible by comparison, and not because the air rubs against the casing wall, which contributes almost nothing.

✈ **"Heat is gained" and "no heat is gained" can both be correct:** Two statements about compression look as though they contradict each other, and both are true because they are describing different things. *Compressing a volume of air gains heat* is a statement about the heat **content** of the gas: its temperature and internal energy go up. *Adiabatic compression involves no loss or gain of heat* is a statement about heat **transfer across the boundary**: nothing has been added from outside. The temperature rise comes entirely from the mechanical work put in by the blades, which is precisely why the process can be adiabatic and get very hot at the same time. When a question uses one of these phrasings, read which of the two senses it is asking about.

Isentropic Temperature Rise: $\frac{T_2}{T_1} = \left(\frac{P_2}{P_1}\right)^{\frac{\gamma-1}{\gamma}}$

Temperatures are absolute, and for air with $\gamma = 1.4$ the index $(\gamma - 1)/\gamma$ works out at 0.286.

⚛ Worked Example — How hot is compressor delivery air?

A modern engine has an overall pressure ratio of **40:1** and takes in air at standard sea level temperature, 288 K. For an ideal, isentropic compression:

$T_2 = 288 \times 40^{0.286} = 288 \times 2.87 = 827$ K, which is **554°C**.

A real compressor is not isentropic. Taking an isentropic efficiency of about 88%, typical of a good multi-stage axial compressor, the actual temperature rise is the ideal rise divided by that efficiency: $(827 - 288)/0.88 = 613$ K, giving an actual delivery temperature of $288 + 613 = 901$ K, or approximately **628°C**. The real machine ends up *hotter* than the ideal one for the same pressure ratio, because the energy that went into the losses had nowhere to go except into the air.

Two consequences follow immediately. First, no fuel has yet been burned and the air is already glowing-hot by everyday standards, which is why the rear compressor stages cannot be made of titanium — its useful limit in this application is around 450 to 500°C — and are made of nickel-based alloys or steels instead. Second, air bled from that stage for cabin conditioning has to be cooled before it can be used, and air bled to cool turbine blades is only "cooling" air in the relative sense that the gas it is protecting them from is hotter still.

Bernoulli's Theorem and the Shape of a Duct

Bernoulli's theorem states that **the total energy of a flow remains constant** along a streamline. It does not say that static and dynamic pressure are equal, and it does not say that one of them is always larger than the other. It says that their *sum* is fixed, so that whatever one gains the other must give up. Three quantities are involved:

- **Static pressure** — the pressure the gas exerts on the walls of the duct, and the pressure that an observer drifting along with the flow would measure. This is the pressure a compressor exists to raise.
- **Dynamic pressure** — the pressure equivalent of the flow's kinetic energy, $\frac{1}{2}\rho v^2$ at low speeds. It exists only because the gas is moving, and it disappears the instant the gas is brought to rest.
- **Total pressure**, also called stagnation or pitot pressure — the sum of the two. It is what a probe facing directly into the flow measures, because the flow is brought to rest at the probe mouth and all of the dynamic pressure reappears as static.

Bernoulli, incompressible form: $P_{static} + \frac{1}{2}\rho v^2 = P_{total} = \text{constant}$

The practical consequence is the behaviour of ducts. In **subsonic** flow — which is where nearly every duct in a gas turbine operates — a **convergent** duct, whose cross-sectional area reduces, accelerates the gas and its static pressure *falls*. A **divergent** duct, whose area increases, decelerates the gas and its static pressure *rises*. A divergent duct used deliberately to slow a flow down and recover static pressure from it is called a **diffuser**. Total pressure, meanwhile, never rises in a plain duct: real ducts always surrender a little of it to friction and turbulence, and the quality of a duct is measured by how little total pressure it loses.

Where in the engine	Duct shape	What it does to subsonic flow	What goes wrong if the shape is disturbed
Subsonic intake duct	**Divergent**	Slows the air, raises static pressure, delivers it evenly to the fan	Dents, lip erosion or a displaced acoustic panel distort the flow and can provoke fan flutter or surge
Compressor stator passages	**Divergent**	Slows the gas the rotor has just accelerated, converting its kinetic energy into static pressure	Blade erosion or fouling opens the passage and loses pressure rise, reducing surge margin
Pre-combustor diffuser	**Divergent**	Slows the delivery air enough for a flame to be stabilised in it	Too high a velocity at the burner blows the flame downstream and can cause flame-out
Turbine nozzle guide vanes	**Convergent**	Accelerates the gas and turns it onto the turbine blades at the correct angle	Bowing, cracking or burning changes the effective throat area and shifts the whole engine's running line
Propelling nozzle	**Convergent**	Converts the remaining pressure energy into jet velocity	A changed exit area alters both jet velocity and pressure thrust, and mis-schedules the engine

⚠ **The rule reverses above Mach 1:** Everything above assumes subsonic flow. Once the flow passes the local speed of sound, any further increase in velocity makes the density fall faster than the velocity rises, so the product of the two falls and the duct area must *increase* to pass the same mass. The relationship therefore inverts: a *convergent* passage now decelerates supersonic flow, and a *divergent* passage accelerates it. That is why a nozzle required to produce a supersonic exit velocity must be convergent-divergent, with the flow reaching exactly Mach 1 at the throat and continuing to accelerate through the divergent portion beyond it. Applying the subsonic rule to a supersonic flow — or the supersonic rule to a subsonic one — gives exactly the wrong answer.

Bernoulli in the simple form above assumes constant density, which is a fair approximation at low speed and a poor one inside a compressor or a jet pipe, where density changes a great deal. The principle survives the correction; what changes is the bookkeeping, because the energy account then has to include the *internal* (thermal) energy of the gas alongside its pressure and velocity. The general statement that survives is the **steady flow energy equation**: in a duct with no work input and no heat transfer, the total energy of the gas — thermal plus kinetic — is unchanged from one end to the other. Slowing the gas down therefore raises its temperature as well as its pressure. This is why the air arriving at the compressor face of an aircraft in cruise is already measurably warmer than the outside air before a single blade has touched it, and it is the origin of the ram temperature rise that makes an intake total temperature probe read higher than the outside air temperature.

The Speed of Sound and Why It Matters Inside the Engine

Because the subsonic and supersonic rules are opposites, knowing where the local speed of sound lies is not optional. The speed of sound in a gas is $a = \sqrt{\gamma RT}$ — it depends on the **absolute temperature and nothing else** for a given gas. It does not depend on pressure, and it does not depend on density independently of temperature. At standard sea level conditions, 288 K, that gives $\sqrt{1.4 \times 287 \times 288} = 340$ m/s. At the compressor delivery conditions calculated above, 901 K, it gives $\sqrt{1.4 \times 287 \times 901} = 602$ m/s.

That near-doubling has a direct engineering consequence. The rear stages of a compressor run in air where the speed of sound is very much higher than it is at the front, so their blades can be driven at high speeds without the flow over them going supersonic and shocking. It is also why the front fan of a large engine, working in cold air where the speed of sound is at its lowest, is the stage that runs closest to its aerodynamic limit — and why fan tip speed, not core speed, is usually what caps the rotational speed of the low pressure spool. **Mach number**, the ratio of local flow speed to the local speed of sound, is therefore the parameter that actually matters in a duct, and the same physical velocity can be comfortably subsonic in one part of the engine and supersonic in another.

Worked Example — The ram temperature rise at cruise

The steady flow energy equation says that bringing a moving gas to rest converts its kinetic energy into internal energy, so a flow that is decelerated gets hotter. For a gas being brought to rest from Mach number M, the total temperature relates to the static temperature as $T_{total} = T_{static}\left(1 + \frac{\gamma - 1}{2} M^2\right)$, which for air with $\gamma = 1.4$ is $T_{total} = T_{static}(1 + 0.2M^2)$.

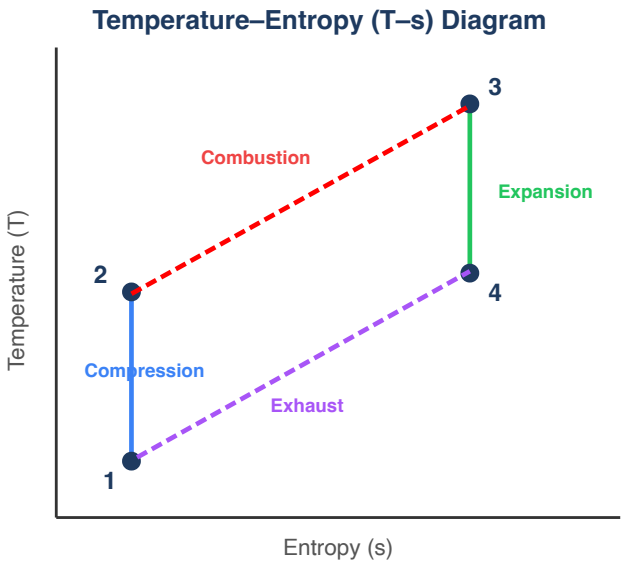
At the standard tropopause the static air temperature is 216.65 K. An aircraft cruising there at Mach 0.85 sees:

$T_{total} = 216.65 \times (1 + 0.2 \times 0.85^2) = 216.65 \times (1 + 0.1445) = 216.65 \times 1.1445 = 248$ K

That is a rise of about **31 K**, from roughly –56°C to roughly –25°C, achieved purely by slowing the air down and before any compressor blade has touched it. Two consequences matter. First, the intake has done part of the compressor's work for nothing, which is the ram effect that makes high-speed flight kinder to a jet engine than the momentum drag term alone would suggest. Second, an intake total temperature probe placed in that decelerated flow reads the higher figure, not the outside air temperature — which is exactly what the engine control system needs, because the engine responds to the air it is actually being given.

The Brayton Cycle (Gas Turbine Thermodynamic Cycle)

Every gas turbine engine operates on the **Brayton cycle** (also called the Joule cycle). It is an *open* thermodynamic cycle with continuous flow, unlike the Otto cycle (petrol engines) or Diesel cycle which are intermittent. The four stages of the Brayton cycle correspond directly to the four main sections of a gas turbine engine:



Brayton Cycle → Engine Sections

1→2 Compression
Compressor section raises pressure & temperature

2→3 Combustion
Fuel burned at constant pressure; huge temp rise

3→4 Expansion
Turbine extracts energy to drive compressor (+ nozzle)

4→1 Exhaust
Gas exits to atmosphere

What the Horizontal Axis Means

The diagram is plotted against **entropy**, and it is worth knowing what that quantity is rather than treating it as a label on an axis. Entropy is a measure of how much of a system's energy is no longer available to be turned into useful work. Add heat to a gas and its entropy rises; the amount it rises by depends on the temperature at which the heat was added, because $\Delta S = Q/T$ for a reversible process — the same quantity of heat added at a high temperature raises entropy less than if it were added at a low one. That is the formal reason why a cycle that adds its heat at the highest practical temperature and rejects it at the lowest practical temperature is the most efficient one, and it is why turbine entry temperature matters as much as pressure ratio.

Two consequences make the diagram readable. First, on temperature-entropy axes the **area beneath a process line is the heat transferred** during that process, so the area under the 2 to 3 leg is the heat put in by the fuel, the area under the 4 to 1 leg is the heat thrown away in the exhaust, and the *enclosed* area of the cycle is the net work obtained. A cycle drawn taller or wider encloses more area and does more work. Second, entropy is created — never destroyed — by every real irreversibility: friction, turbulence, mixing streams at different temperatures, and unrestrained expansion. That is what the second law of thermodynamics says, and its practical meaning inside an engine is blunt: **every unit of entropy generated is work that has been permanently lost** and cannot be recovered downstream by any clever piece of design. Cleaning up a loss is always cheaper than trying to recover the energy it destroyed.

The Four Processes in Detail

The numbered points on the diagram are *cycle states*, not physical positions in the engine, and each numbered leg between them is a distinct thermodynamic process with a name of its own:

- **1 to 2 — adiabatic, ideally isentropic, compression.** The compressor does work on the air. Pressure rises, temperature rises, volume falls, and in the ideal case entropy does not change, which is why the leg is drawn as a vertical line on temperature-entropy axes. No heat is added from outside; the entire temperature rise is the work put in.
- **2 to 3 — constant-pressure heat addition.** Fuel is burned continuously in the combustor. Temperature climbs steeply and the gas expands in volume, but the pressure is held essentially constant because the gas is free to flow onwards rather than being sealed in. Entropy rises, because heat has been added, which is why this leg slopes to the right.
- **3 to 4 — adiabatic, ideally isentropic, expansion.** The hot, high-pressure gas expands through the turbine and gives up energy as shaft work. Pressure and temperature fall; entropy ideally does not change, so this leg is drawn vertical as well.
- **4 to 1 — constant-pressure heat rejection.** The exhaust gas leaves the engine and surrenders its remaining heat to the atmosphere at ambient pressure, closing the cycle so that it can begin again with fresh air at state 1.

One feature of the diagram is worth reading carefully, because it carries a real engineering consequence: the height of point 4 against point 2. The heat carried away along the 4 to 1 leg, everything between point 4 and point 1, is the single largest loss in the cycle, and it is the reason a gas turbine's thermal efficiency is nowhere near 100% however well its components are made. Whether any of it can be reclaimed depends on which of points 4 and 2 is the higher, and that in turn depends on the pressure ratio. The diagram here draws point 4 *above* point 2, which is the case at the modest pressure ratios of the early jet engines and of most stationary plant: the exhaust leaves hotter than compressor delivery, so a heat exchanger can return part of its heat to the air on its way to the combustor. Raise the pressure ratio at a given turbine entry temperature and point 2 climbs while point 4 falls. For the 1,700 K turbine entry temperature used earlier in this note the two cross at a pressure ratio of about 22:1, and at the 40:1 of a modern engine the ideal figures are 827 K at compressor delivery against 592 K at the fully expanded exhaust, so there is nothing left for a heat exchanger to move. That is why recovery by heat exchanger belongs to low pressure ratio stationary plant. A steam bottoming cycle carries no such restriction, which is how a large modern industrial machine still takes useful work from its exhaust; an aero engine can carry neither installation, so it recovers only what the exhaust can still contribute as jet thrust.

The Same Cycle on Pressure-Volume Axes

The same four processes can be drawn on pressure against volume instead of temperature against entropy, and the two views answer different questions. On pressure-volume axes the compression leg 1 to 2 climbs steeply to the left as pressure rises and volume falls. Heat addition, 2 to 3, runs *horizontally* to the right — that horizontal line is what "constant pressure" looks like, and it is the single most recognisable feature of the diagram, because on the Otto cycle of a piston engine the equivalent line is *vertical*: constant volume, with pressure shooting up. Expansion, 3 to 4, falls to the right as the gas gives up pressure and grows in volume, and heat rejection, 4 to 1, returns horizontally at ambient pressure.

On these axes the enclosed area is again the net work of the cycle, but now it is the work done by the gas *on the machine* rather than the heat balance. Between them the two diagrams make the same point from two directions: the temperature-entropy plot shows where the heat goes, and the pressure-volume plot shows where the work is done. If only one of them is remembered, the useful one for identifying a cycle from a picture is pressure against volume, because the orientation of the combustion line — horizontal for Brayton, vertical for Otto — settles the identification immediately.

Open, Continuous and Constant-Pressure — Three Different Claims

Three descriptions of the Brayton cycle appear repeatedly in examination questions, and each is answering a different question about it. Confusing them is easy, because all three are true at once.

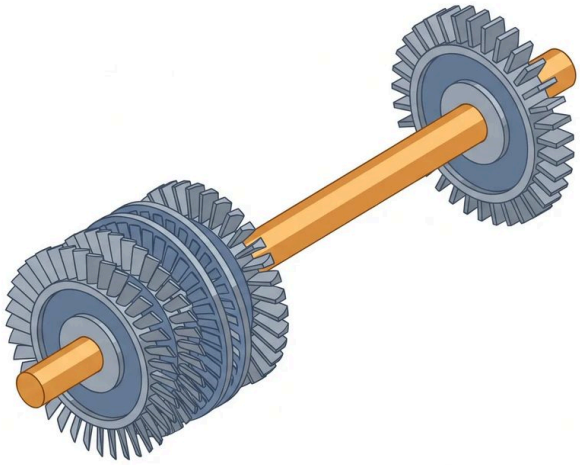
- **Open** describes the working fluid. It is not recirculated: fresh air is taken from the atmosphere at one end and exhaust gas is discharged to the atmosphere at the other, and the atmosphere itself provides the heat-rejection leg from 4 back to 1. A *closed* cycle would need a heat exchanger to cool the same fluid back down and return it to the compressor inlet, which is how a closed-cycle industrial or nuclear gas turbine works and is not how any aero engine works.
- **Continuous** describes the flow. Nothing is trapped, valved or timed. There is not a single valve anywhere in a gas turbine's main gas path, which is one of the reasons it has so few moving parts compared with a piston engine of the same power.
- **Constant pressure** describes the heat addition, and it is the property that gives the cycle its other name, the **constant pressure cycle**. It is the single feature that most sharply distinguishes it from the piston engine's **constant volume** Otto cycle, in which the charge is sealed inside a cylinder by closed valves while it burns, so that pressure rockets up while volume barely changes at all.

Describing the heat addition as taking place at constant pressure carries a qualification that should not be glossed over. A practical combustor loses a small fraction of the compressor delivery pressure to friction, to the pressure drop across the swirlers and dilution holes, and to the momentum change involved in heating the gas. The pressure at the combustor exit is therefore slightly *lower* than at its inlet, typically by a few per cent, not identical to it. The cycle is still called constant-pressure because that small drop is negligible beside the forty-fold rise the compressor has just produced — and because the alternative label, constant-volume, would be spectacularly wrong.

Where the Turbine's Work Actually Goes

A figure worth pausing on is how much of the turbine's output never reaches the aircraft at all. The turbine's first duty is to drive the compressor, and compressing air is expensive. On a simple turbojet the turbine is sized to extract little more than the compressor and accessories need, and on a modern high pressure ratio core the power the compressor demands is itself of the order of two thirds of the rate at which the fuel releases heat — a figure derived from first principles in the worked example below. That power circulates

rather than disappears, because the turbine takes it out of the gas and the compressor puts it straight back into the same gas stream, so it is a measure of how hard the turbine has to work and not of how much of the fuel's energy has been used up. A further small slice is taken through the accessory gearbox for the fuel pump, the oil pumps, the electrical generators and the hydraulic pumps. Only what survives all of that is available as useful output.



Compressor and turbine share one shaft: the turbine's first duty is to drive the compressor.

Worked Example — One kilogram of air right through the cycle

Take a gas generator passing 100 kg/s, drawing air at standard sea level temperature, running an overall pressure ratio of 40:1 and a turbine entry temperature of 1,700 K. Use $c_p = 1.005$ for the cold air in the compressor and $c_p = 1.15$ for the hot gas in the turbine, because the working fluid is genuinely different on the two sides of the flame.

- **1 to 2, compression.** The delivery temperature was calculated earlier as about 901 K, a rise of 613 K. The work the compressor absorbs is $\dot{m}\, c_p\, \Delta T = 100 \times 1.005 \times 613 = 61{,}600$ kW, that is **61.6 MW**.
- **2 to 3, heat addition.** Raising the gas from 901 K to 1,700 K takes $100 \times 1.15 \times 799 = 91{,}900$ kW, that is **91.9 MW** of heat from the fuel.
- **Expansion, the first part of 3 to 4.** The turbine has to deliver the compressor's 61.6 MW. The temperature drop that costs is $\Delta T = 61{,}600 \div (100 \times 1.15) = 536$ K, so the gas leaves this turbine at about $1{,}700 - 536 = 1{,}164$ K, roughly 891°C.

The striking number is the ratio: $61.6 \div 91.9 = 0.67$. **The compressor's power demand is two thirds of the rate at which the fuel is releasing heat**, and the turbine has to produce every watt of it before it can produce anything else. That is the two-thirds figure quoted above, arrived at independently from the gas properties rather than taken on trust.

It does not follow that two thirds of the fuel's energy has been used up. The 61.6 MW leaves the gas at the turbine and is put straight back into the same gas at the compressor, on the same shaft and in the same flow, so none of it has left the cycle; with the turbine sized to deliver exactly what the compressor absorbs, the gas arriving at the gas generator turbine exit still carries essentially the whole of the heat the fuel added to it. What the ratio measures is the size of the circulating load the turbine has to carry before it can produce anything at all, not a share of the fuel's energy that has been spent.

Two cautions about reading the 1,164 K. It is the temperature leaving the turbine that drives *this* compressor, so it is a gas generator exit temperature, not the temperature at the tailpipe of a complete engine. On a turbofan or a turboshaft, further turbine stages downstream extract the fan or shaft power and drop the gas a great deal further before it reaches the exhaust. And the figure ignores turbine cooling air and mechanical losses, both of which a real engine has to pay for. The arithmetic is right; what it describes is the core of the machine, not the whole of it.

What "useful output" means is the single choice that separates the four engine configurations described later in this note, and it is worth stating as a rule. If the turbine is sized to extract *only* what the compressor and the accessories need, then the surplus energy stays in the gas as pressure and heat, and the propelling nozzle converts it into a fast jet — that is a turbojet. If extra turbine stages are added so that more work is taken out than the compressor needs, the surplus appears at a shaft, where it can drive a fan, a propeller or a rotor, and correspondingly less energy is left in the exhaust to make a jet. There is no third possibility, and no way to have both: **every joule the turbine takes out of the gas is a joule the nozzle cannot use, and every joule left in the gas is a joule the shaft does not get.**

✈ **Aviation Context:** The Brayton cycle is **continuous** — unlike a piston engine's intermittent cycle. All four processes happen simultaneously in different parts of the engine at all times. This is why gas turbines produce smooth, vibration-free power and can handle very high power-to-weight ratios. A modern turbofan may process over **500 kg of air per second**.

Pressure Ratio and Thermal Efficiency

The thermal efficiency of an ideal Brayton cycle turns out to depend on one thing only — the pressure ratio the compressor achieves:

Ideal Brayton Thermal Efficiency: $\eta_{th} = 1 - \frac{1}{r_p^{\frac{\gamma-1}{\gamma}}}$

Where r_p is the overall pressure ratio and the index $(\gamma - 1)/\gamma$ is 0.286 for air.

🔥 **Worked Example — What a higher pressure ratio buys**

An early turbojet with an overall pressure ratio of **10:1**: $10^{0.286} = 1.93$, so $\eta_{th} = 1 - 1/1.93 = 1 - 0.518 = 0.482$, about **48%**.

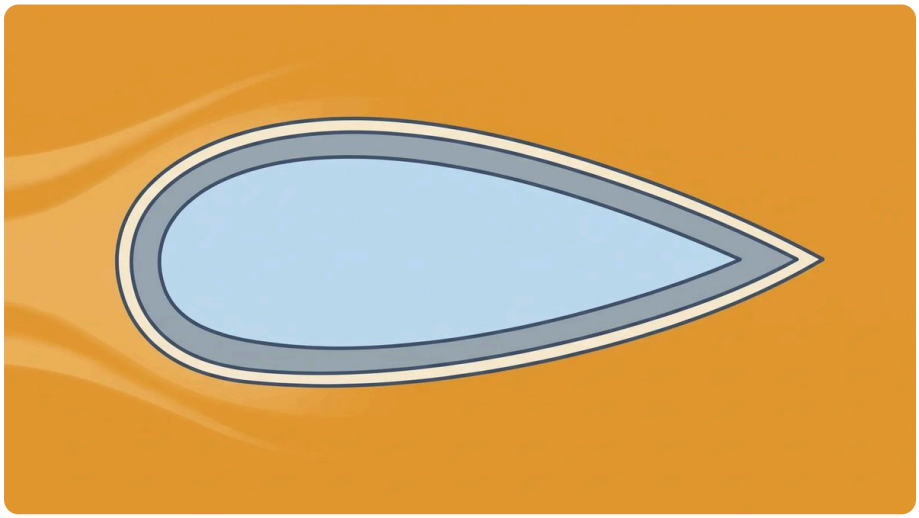
A modern engine at **40:1**: $40^{0.286} = 2.87$, so $\eta_{th} = 1 - 1/2.87 = 1 - 0.348 = 0.652$, about **65%**.

Quadrupling the pressure ratio has lifted the ideal thermal efficiency by roughly seventeen percentage points. These are ceilings and not achievements — real component losses, the air bled from the compressor for cooling and the pressure drop through the combustor all pull the delivered figure well below them. But the *direction* of the relationship is what matters, and it is the entire reason overall pressure ratios have climbed decade after decade.

Pressure ratio cannot simply be raised without limit, and the reason is a competing effect that governs the whole of modern engine design. Raising the pressure ratio raises the compressor delivery temperature along with it; the worked example earlier in this note put a 40:1 machine at around 900 K at delivery. That matters twice over. First, compressor delivery air is the coolant for the turbine, so as it becomes hotter it becomes less effective as a coolant, and a greater quantity of it has to be bled off to do the same job — air that is then unavailable to make thrust. Second, the temperature window available for heat addition, which is the gap between the compressor delivery temperature and the highest temperature the turbine can survive, is squeezed from below. Higher pressure ratio improves the cycle; higher compressor delivery temperature penalises the components. The pressure ratio actually selected is the point at which those two lines cross for the materials and cooling technology of the day, which is why it has crept upwards only as turbine materials and blade cooling have improved.

The ceiling at the top of the cycle is the **turbine entry temperature** — the gas temperature at state 3, where the flow first meets the turbine. On modern civil engines it works across a band of roughly 1,200–1,600°C, sitting near the lower end in the cruise and rising to the upper end at take-off power, and the most highly rated

designs exceed 1,600°C at take-off. The top of that band is above the melting point of the alloys the turbine is made from, which is only possible because the blades are internally cooled and film-cooled with compressor air and protected by ceramic thermal barrier coatings. Turbine entry temperature is therefore not a thermodynamic limit at all. It is a materials and cooling limit, and it is the number that has risen furthest as engines have improved.



Internal cooling air and a ceramic coating let a turbine blade sit in gas hotter than its own melting point.

The Absolute Ceiling: the Carnot Limit

Before leaving efficiency, it is worth knowing what the ultimate ceiling is, because it explains why engineers push turbine entry temperature so hard. No heat engine of any design can convert more of its heat input into work than:

Carnot Efficiency: $\eta_{carnot} = 1 - \dfrac{T_{cold}}{T_{hot}}$

Where both temperatures are absolute, T_{hot} is the temperature at which heat is taken in, and T_{cold} the temperature at which it is rejected.

Take a turbine entry temperature of 1,700 K and an ambient of 288 K: $\eta_{carnot} = 1 - 288/1{,}700 = 0.83$, so **83%**. No engine working between those two temperatures can beat that, whatever its pressure ratio and however perfect its components. The ideal Brayton cycle at 40:1 reached about 65% of the heat input, and a real engine achieves appreciably less again. Each step down from 83% is paid for by a specific, identifiable cause: the Brayton cycle does not add all its heat at the peak temperature the way the ideal Carnot cycle does, and then component losses, cooling air and pressure drops take their own share.

The useful lesson is in the shape of the formula. T_{cold} is fixed by the atmosphere and cannot be changed. The only variable available is T_{hot}, which is turbine entry temperature — so every kelvin gained there raises the

ceiling for everything downstream. That is why turbine cooling technology, single-crystal blade casting and thermal barrier coatings have received so much development effort: they do not improve the engine directly, they raise the temperature at which the whole cycle is permitted to operate.

Why Small Efficiency Changes Have Large Effects

A gas turbine has an unusual and somewhat uncomfortable property, and it follows directly from how much of the energy the compressor consumes: wherever the turbine is sized to produce surplus shaft work, that surplus is the small *difference* between two large numbers. The turbine produces a great deal of work and the compressor consumes most of it, so anything that changes either one is magnified in what is left over. Engineers call the fraction of turbine work consumed by the compressor the **back work ratio**, and for a gas turbine it is very high — a steam plant, by comparison, spends only a per cent or two of its turbine output on the feed pump, because pumping a liquid is vastly cheaper than compressing a gas.

Worked Example — The magnifying effect of a small loss

Suppose a shaft-power engine's turbine produces 100 units of work while the compressor absorbs 65 of them. The net output available at the shaft is $100 - 65 = 35$ units.

Now let the turbine deteriorate so that it produces 5% less, 95 units, with the compressor unchanged. The net output becomes $95 - 65 = 30$ units — a fall of $5/35 = 14.3\%$.

A 5% change in one component has produced nearly a 15% change in the engine's useful output. This is why component efficiencies are protected so jealously, why compressor washing is worth doing on a schedule rather than when performance has visibly collapsed, and why turbine tip clearance control is worth the complexity of an active system. It is also why a gas turbine deteriorates in a way the crew notices only late: the engine compensates by burning more fuel and running hotter long before the thrust actually becomes unavailable.

Efficiencies, and What Each One Measures

Several different efficiencies are used about a gas turbine and they are not interchangeable. Each answers a different question, and quoting the wrong one is a common source of confusion.

Efficiency	The question it answers	Typical value
Isentropic (component)	How close is this compressor or turbine to the ideal, loss-free process for the same pressure ratio?	High, and the last few points are what manufacturers compete over
Combustion	What proportion of the fuel's chemical energy is actually released as heat in the combustor?	Very high at cruise and take-off; it falls away at idle, which is where unburned hydrocarbons and carbon monoxide appear
Mechanical	How much shaft work survives the bearings, seals, gearboxes and accessory drives?	High, and the losses appear as heat in the oil

Efficiency	The question it answers	Typical value
Thermal (cycle)	How much of the fuel's heat becomes kinetic energy in the gas stream?	Governed mainly by pressure ratio and turbine entry temperature
Propulsive	How much of that kinetic energy becomes useful work on the aircraft?	Governed by the ratio of jet velocity to flight speed
Overall	How much of the fuel's energy ends up propelling the aircraft? It is the thermal and propulsive figures multiplied together	The number that actually shows up in the fuel figures

Following the Fuel Energy Through the Engine

Putting the pieces together, it is possible to say roughly what becomes of the 150.5 MW of chemical energy released by the fuel flow calculated earlier in this note. The exact split depends on the engine, the power setting and the flight condition, so the ordering below is the one that holds at the take-off condition that fuel flow belongs to, and the proportions should be read as indicative only.

- **The largest single share leaves as hot exhaust gas.** This is the 4 to 1 leg of the cycle: heat rejected to the atmosphere because the cycle has no way to use gas that has already expanded to ambient pressure. It is the loss the Carnot limit is really describing.
- **The next largest share is kinetic energy left in the jet** that never acts on the aircraft, which is exactly what propulsive efficiency measures and exactly what a high bypass ratio is designed to reduce.
- **Useful propulsive work** — thrust multiplied by flight speed — is what remains after those two, and it is a minority of the input in every flight condition.
- **Smaller identifiable losses** make up the rest: heat carried away in the oil from bearings and gearboxes, power taken by the accessory drives for the aircraft's electrical and hydraulic systems, air bled from the compressor for cabin conditioning and anti-icing, pressure losses in the intake, ducts and combustor, and acoustic energy radiated as noise.

The middle two items are the pair that moves. They sit in that order at take-off, where the aircraft is barely moving and propulsive efficiency is therefore near zero, so almost all of the kinetic energy put into the jet is wasted. In the cruise they change places: the propulsive efficiency worked example earlier in this note puts Engine B at 83% at 250 m/s, which is five parts of useful propulsive work for every one part left behind in the jet. The first item holds its position at every condition, and useful propulsive work stays a minority of the fuel energy at all of them.

Two things follow from that ordering. The first is that anything the airframe takes from the engine — bleed air, generator load, hydraulic demand — is paid for in fuel, which is why bleed and electrical loads are managed rather than ignored on a modern aircraft. The second is that the two largest losses are both thermodynamic rather than mechanical, which is why engine development has been dominated for sixty years by pressure ratio, turbine entry temperature and bypass ratio, and not by better bearings.

The Real Cycle Against the Ideal One

The diagram above shows the ideal cycle, in which the compression and expansion legs are exactly vertical. A real cycle drawn on the same axes leans to the right on both of them, because every irreversibility — blade friction, tip leakage, shock losses, turbulence, the mixing of cooling air into the mainstream — adds entropy. Once that leaning is seen, the consequences are entirely predictable:

- A real compressor delivering a given pressure ratio arrives at a **higher** exit temperature than the ideal one, so it absorbs *more* work than the ideal calculation predicts.
- A real turbine expanding through a given pressure ratio also arrives at a **higher** exit temperature than the ideal one, so it delivers *less* work than the ideal calculation predicts.
- Both errors act in the same unhelpful direction, so the real cycle's net work output and thermal efficiency are always below the ideal figures, never above.

That asymmetry — losses cost you work at the compressor and then cost you work again at the turbine — is why component efficiencies of a percentage point or two are pursued so hard by manufacturers, and it is why an engine whose compressor is fouled with dust and insect debris, or whose turbine blade profiles have eroded, shows up as raised fuel burn and raised exhaust gas temperature long before it shows up as any loss of thrust the crew would notice.

✈ **Why a turbine needs far fewer stages than a compressor:** A turbine expands gas; a compressor squeezes it. Expansion works *with* the natural pressure gradient, since the gas is travelling from high pressure towards low pressure, which is the direction it wants to go anyway. Compression works *against* it, pushing gas towards higher pressure through what aerodynamicists call an adverse pressure gradient. Flow in an adverse gradient tends to separate from the blade surfaces, so a compressor stage can only be given a modest pressure rise before it stalls. A single turbine stage can comfortably drive a whole group of compressor stages, which is why a high pressure spool carrying nine or ten compressor stages may need only one or two turbine stages to drive them.

Cycle States Against Engine Station Numbers

One source of confusion is worth clearing up before it takes root: the numbers 1 to 4 used for the cycle states above are not the same numbers used to label physical positions along the engine. Engine documentation uses a **station numbering** convention in which station 0 is the undisturbed free stream ahead of the aircraft, station 1 is the intake entry, station 2 is the fan or compressor face, station 3 is compressor delivery at the combustor inlet, station 4 is turbine entry at the combustor outlet, station 5 is turbine exit, and higher numbers run aft through the jet pipe to the nozzle. Manufacturers differ in the sub-numbers they insert between these — an intermediate compressor exit might be station 2.5 on one engine and something else on another — so read the numbering from the manual for the engine in front of you rather than assuming it.

The practical value of the convention is that it makes flight-deck and maintenance parameters unambiguous. A temperature labelled T3 is compressor delivery temperature and nothing else; a pressure labelled P3 is compressor delivery pressure measured at that one plane. When a troubleshooting chart asks for a P3 reading it is asking for the pressure at a specific station, and taking it from a convenient nearby tapping instead will produce a plausible number and a wrong diagnosis.

The last three rows of the efficiency table earlier in this section explain something counter-intuitive about the pure turbojet. Its cycle is perfectly respectable; what lets it down at subsonic speed is the second multiplier, its propulsive efficiency. An engine can be excellent at one and poor at the other, and when the high bypass turbofan displaced the turbojet on transport aircraft it did so without improving the cycle at all — it improved only the way the cycle's output was delivered to the aircraft.

Force, Work, Power and Energy

Quantity	Definition	Formula	SI Unit
Force	A push or pull that changes a body's state of motion	$F = ma$	Newton (N)
Work	Force applied over a distance	$W = Fd$	Joule (J)
Power	Rate of doing work	$P = W/t$	Watt (W)
Energy	Capacity to do work (kinetic + potential)	$E_k = \frac{1}{2}mv^2$	Joule (J)
Velocity	Rate of change of displacement (speed + direction)	$v = d/t$	m/s
Acceleration	Rate of change of velocity	$a = \Delta v/t$	m/s²

The table gives the definitions. The *distinctions* between them are what the examination actually tests.

Force Is Not Work, and Work Is Not Power

A force that moves nothing does no work. Push against a parked aircraft with all your strength and, if it does not move, the work done on it is exactly zero however tired you become. Work appears only when the force acts *through a distance*, which is why its formula is $W = Fd$ and why its unit, the joule, is identical to the unit of energy — work is simply energy transferred by a force. Power then adds time to the account: it is the *rate* at which work is done, measured in watts, where one watt is one joule per second. The three quantities cannot be interchanged, and their units are a reliable check on any answer: **newtons for force, joules for work and energy, watts for power**. The horsepower is still met alongside the watt in engine work, and one horsepower is 745.7 watts.

⚠ **Thrust and Power Are Different Quantities:** A turbojet or turbofan is rated in units of **force** — newtons, decanewtons or pounds of thrust. A turboprop or turboshaft is rated in units of **power** — kilowatts or shaft horsepower. The two cannot be compared directly, because power is thrust multiplied by speed, $P = F \times V$. An engine held stationary on a test bed produces its full rated thrust and delivers *zero* propulsive power, because the point at which the force is applied is not moving. The same engine producing 60 kN in cruise at 250 m/s is delivering $60{,}000 \times 250 = 15$ MW of propulsive power. Quoting a turbofan in horsepower, or a turboshaft in pounds of thrust, is a category error rather than a rounding problem.

Shaft Power, Torque and Rotational Speed

For any rotating shaft — a turboprop output shaft, a helicopter rotor drive, an accessory drive — power is delivered as the product of torque and rotational speed. The same power can therefore be delivered as a large torque at low speed or a small torque at high speed, and that equivalence is the entire purpose of a reduction gearbox: it converts a turbine's high-speed, low-torque output into the low-speed, high-torque input that a propeller or rotor needs, at very nearly constant power, losing only what friction takes. It is also why **torque** is the parameter used to set power on a shaft-power engine. With the propeller governed at a constant speed, torque is directly proportional to power, so a torque gauge *is* a power gauge — which is not true of a thrust-producing engine, where no single mechanical parameter stands in for power in the same way.

Velocity, Speed and Acceleration

Speed is a scalar: a magnitude with no direction attached. **Velocity** is a vector: a magnitude *and* a direction. The distinction is not pedantry inside an engine, because a change of direction with no change of speed is still a change of velocity and therefore still demands a force. Gas turned through a large angle by a row of turbine nozzle guide vanes has been accelerated even if it leaves at exactly the speed it entered, and the force that turned it is felt by the vanes as a load their attachments and locating features have to carry. This is one reason nozzle guide vanes are among the most heavily loaded static components in the engine despite doing no shaft work at all.

Acceleration is the rate of change of velocity with respect to time, measured in m/s². It is neither a force nor an average speed — a force is what *produces* an acceleration, through Newton's second law, and the two must never be swapped in an answer. The accelerations inside a propelling nozzle are extraordinary by everyday standards, and it is worth putting a number on one.

Worked Example — Acceleration in a propelling nozzle

Gas enters a nozzle at 150 m/s and leaves at 600 m/s. Take the accelerating length as 0.9 m for the purposes of the calculation. Using $v^2 = u^2 + 2as$, rearranged to $a = (v^2 - u^2)/2s$:

$$a = \frac{600^2 - 150^2}{2 \times 0.9} = \frac{360{,}000 - 22{,}500}{1.8} = \frac{337{,}500}{1.8} = 187{,}500 \text{ m/s}^2$$

That is roughly **19,000 times the acceleration due to gravity**. The force producing it comes from the pressure difference across the nozzle, and the reaction to that force is part of the thrust. It also explains why anything shed upstream — a fragment of blade, a piece of combustor liner, a tool left in the gas path — leaves the engine at a speed that makes the exhaust hazard area a serious matter rather than a formality.

One more quantity completes the set. **Momentum** is mass multiplied by velocity, measured in kilogram metres per second, and it is what Newton's second law is really about: force is the rate at which momentum changes. For a steady stream of gas, the momentum crossing any plane each second is $\dot{m}v$, so the thrust equation given earlier is nothing more than the momentum leaving the engine minus the momentum entering it, plus the pressure term. Reading it that way makes it considerably harder to misremember, and it makes clear why raising the mass flow and raising the velocity change are equally valid routes to the same thrust.

Units, Symbols and the Conversions Worth Knowing

Engine work is carried out in a mixture of SI and imperial units, because manufacturers, operators and regulators in different parts of the world settled on different conventions and the documentation has never been fully harmonised. Being fluent in both is a practical requirement rather than an academic nicety: a thrust setting chart, a torque limit and a pressure gauge on the same aircraft may each use a different family of units.

Quantity	Symbol	SI unit	Also met, and the conversion
Force / thrust	F	newton (N)	pound-force; 1 lbf = 4.448 N. Also the decanewton (daN), 1 daN = 10 N
Work / energy	W, E	joule (J)	British thermal unit; 1 BTU is about 1,055 J
Power	P	watt (W)	horsepower; 1 hp = 745.7 W, so 1 kW is about 1.341 hp
Pressure	P	pascal (Pa), one newton per square metre	bar (100,000 Pa), hectopascal (100 Pa), psi (about 6,895 Pa); standard sea level is 1013.25 hPa or 14.7 psi
Temperature	T	kelvin (K)	degrees Celsius (add 273.15 for kelvin); degrees Fahrenheit and degrees Rankine (add 459.67 for Rankine)
Mass flow	$\dot{m}$	kilogram per second (kg/s)	pounds per hour is common for fuel flow; watch for the change of both mass unit and time unit at once
Torque	T or Q	newton metre (N m)	pound-feet and pound-inches; also expressed as a percentage of a reference torque on many turboprops

⚠ **The dot and the symbol clash:** Two notation traps recur throughout engine work. The dot over a symbol, as in $\dot{m}$, means "per unit time" — m is a mass in kilograms while $\dot{m}$ is a mass *flow* in kilograms per second, and the two are not interchangeable in any equation. Separately, the letter T is used for both temperature and torque and the letter P for both pressure and power, so which is meant has to be read from the context. A calculation that produces a nonsensical magnitude has very often swapped one of these pairs.

Engine Architecture: Gas Path, Spools and Modules

The cycle described above has to be built out of hardware, and the way that hardware is arranged is as much a part of the fundamentals as the thermodynamics. Two engines can run near-identical cycles and still be very different machines to operate and maintain, depending on how many shafts they have, where the drives are taken from, and whether the major assemblies can be separated from one another without dismantling everything else.

The Gas Path and the Order of the Sections

Air passes through a basic turbojet in one fixed order, front to rear: **intake, compressor, combustion chamber, turbine, exhaust nozzle**. That order is not a design preference; it is forced by the cycle. The compressor must come before the combustor, because there is no point burning fuel in air that has not been compressed — pressure ratio is what makes the cycle efficient at all. The turbine must come after the combustor, because it is hot gas that carries the energy to drive it. The nozzle must come last, because its job is to convert whatever pressure energy remains into velocity. Any answer that offers a different sequence — the turbine before the combustor, the nozzle near the front, a recirculating loop inside the core — is describing a machine that cannot run.

Of those five, three are described as the **main sections** of any gas turbine: the **compressor**, the **combustion** section and the **turbine**. Items such as the diffuser, the stator vanes, the nozzle guide vanes and the exhaust cone are components *within* those sections rather than sections in their own right, and offering one of them as a main section alongside the other three is a standard distractor. The intake and the exhaust are usually treated as part of the installation rather than as core sections, which is why they appear in the gas-path sequence but not in the list of three.

The Cold Section and the Hot Section

The engine is also divided into a **cold section** and a **hot section**, and the dividing line is the point at which fuel is burned. Everything upstream of combustion handles air at compressor temperatures; everything downstream handles gas at combustion temperatures.

Section	What it contains	Typical materials	Characteristic deterioration
Cold section	Intake, fan, all compressor stages, the pre-combustor diffuser, the accessory drives and gearbox	Aluminium alloys and composites at the front, titanium in the fan and forward compressor, steels and nickel alloys in the hot rear stages	Foreign object damage, erosion, corrosion, fouling by dust and insect debris, blade tip rub
Hot section	Combustor, all turbine stages, exhaust	Nickel- and cobalt-based superalloys, single-crystal castings, ceramic thermal barrier coatings	Creep, thermal fatigue cracking, oxidation, sulphidation, coating loss, burning and distortion

The pre-combustor **diffuser** is the item that sits closest to that dividing line. It sits immediately in front of the combustor and its job is to slow the compressor delivery air before it reaches the flame, but it is upstream of ignition and therefore belongs to the *cold* section. A statement placing the diffuser in the hot section is wrong, and a statement placing the turbine in the cold section is wrong for the same reason in the opposite direction. The test is always the same: has this component seen burning fuel?

The division is far from academic. It drives the choice of materials, as the table shows. It drives inspection intervals and inspection methods, because the failure mechanisms are entirely different — a cold section problem is usually mechanical damage you can see, while a hot section problem is usually time-at-temperature damage you have to look for. And it drives the shape of a shop visit, because hot section refurbishment is a far heavier and more expensive task than cold section work and is very often the event that defines when an engine comes off wing.

The Gas Generator

The combination of compressor, combustor and the turbine stage or stages that drive that compressor is called the **gas generator**, or simply the **core**. Its job is exactly what the name says: to produce a supply of hot, high-pressure gas. On its own a gas generator is not a propulsion device at all — it is an energy source, and what is attached behind it decides what kind of engine it becomes. A propelling nozzle makes it a turbojet. A fan, plus the extra turbine stages needed to drive that fan, makes it a turbofan. A power turbine driving a shaft makes it a turboprop or a turboshaft. Recognising the identical core inside all four configurations is the quickest route to understanding why they behave so differently while sharing so much hardware, and it is why manufacturers routinely develop one core into several very different products.

What a Gas Generator Needs Before It Is an Engine

A compressor, a combustor and a turbine will not run on their own, and the supporting systems are not accessories in the dismissive sense — several of them are what makes the machine survivable. Five are always present in some form, whatever the engine type.

- A **fuel system** to deliver fuel at the pressure the burners need and, far more importantly, to *schedule* how much is delivered. The fuel control is what prevents the engine over-temperaturing or surging while it accelerates, so it is a protective device as much as a supply system.
- A **lubrication system** to feed and scavenge the bearing chambers and gearboxes. Its second job matters as much as its first: the oil is a coolant, carrying heat away from bearings running in chambers surrounded by very hot structure.
- An **internal air system**, taking air from chosen compressor stages to cool turbine discs and blades, to pressurise the bearing chamber seals, and to balance the axial thrust loads on the rotors. This air never makes thrust and every kilogram of it is a cost, which is why it is metered as carefully as fuel is.
- A **starting and ignition system**, because a gas turbine cannot start itself: something external has to turn the high pressure spool up to a self-sustaining speed while igniters light the fuel.
- An **indication system**, because almost nothing inside a running engine can be seen. Speeds, temperatures, pressures, fuel flow, oil condition and vibration are the entire evidence base on which the engine is operated and maintained.

Each of these is a subject in its own right and is developed later in this module. The point to carry forward from the fundamentals is that they are all consequences of the cycle rather than additions to it: the internal air system exists because turbine entry temperature exceeds what metal can survive, the oil system exists because the shafts must be supported inside a hot machine, and the fuel schedule exists because the compressor has a surge line that must not be crossed.

Spools: One Shaft, Two or Three

A **spool** is a compressor, the turbine that drives it, and the shaft joining them, all rotating together as a single assembly at a single speed. The number of spools an engine has is one of its defining architectural features, and the spool arrangement determines which turbine drives which compressor.

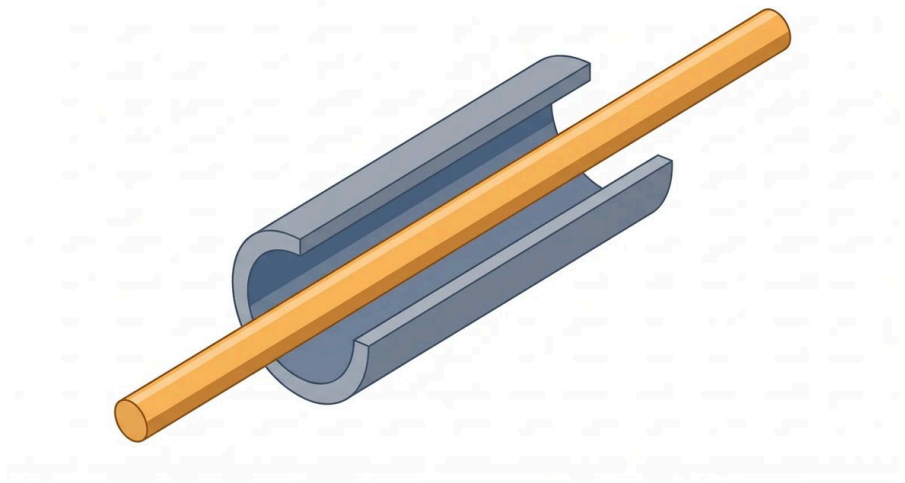
The rule governing which turbine drives which compressor is simple and never varies: **the turbine immediately behind the combustor drives the highest-pressure compressor**, and each successive turbine further downstream drives the next compressor further forward. The reason is energy. Gas leaving the combustor is at its hottest and highest pressure, so the first turbine stage has the greatest energy available to it — and the high pressure compressor is precisely the one that needs the most work per kilogram of air. Working outwards from the flame in both directions keeps the biggest job matched to the biggest energy supply.

Arrangement	Which turbine drives what	Speed designations	Where it is found
Single spool	One shaft carries every compressor stage and every turbine stage; all rotate together at one speed	N1 only	Early turbojets, small auxiliary power units, some small turboshafts

Arrangement	Which turbine drives what	Speed designations	Where it is found
Twin spool	The first turbine drives the HP compressor; the later turbine drives the fan and any booster stages	LP = **N1**, HP = **N2**	The great majority of civil turbofans
Triple spool	The first turbine drives the HP compressor; the next drives the IP compressor; the last drives the fan	LP = **N1**, IP = **N2**, HP = **N3**	Rolls-Royce large civil turbofans

Read the speed designations carefully. **N1 always means the low pressure spool**, which on a front-fan engine is the fan. On a two-spool engine the high pressure spool is N2 — but on a three-spool engine N2 is the *intermediate* spool and the high pressure spool is **N3**. Reading N2 as "the high pressure spool" without first establishing how many spools the engine has is a real trap on the flight line, because the same label on two different engine types refers to different hardware.

A direct consequence of the definition of a spool is that everything mounted on the same shaft turns at the same speed. On a conventional dual-spool front-fan engine — one with no fan drive gearbox — the fan, the low pressure compressor and the low pressure turbine form one rotating assembly, so the **fan turns at exactly the same speed as the low pressure turbine that drives it**. It does *not* turn at the same speed as the high pressure compressor, which is on a physically separate shaft. The two shafts are concentric, the LP shaft running through the bore of the HP shaft, and they are entirely free to run at different speeds and even, on some designs, in opposite directions. Where a fan drive gearbox is fitted, as it is on a geared turbofan, the fan turns at roughly a third of the low pressure turbine's speed while still being driven by it; that arrangement is described later in this note.



The low pressure shaft runs through the bore of the high pressure shaft, free to turn at its own speed.

The reason for splitting the machine at all is that different parts of it want to run at different speeds. A large fan is limited by the tip speed of its blades: once the tips approach and pass the local speed of sound, efficiency

falls away and noise climbs steeply, so a big fan has to turn slowly. A high pressure compressor is small in diameter and needs high blade speed to do its work, so it has to turn fast. Tying them to one shaft forces a compromise that suits neither. Splitting them lets each run near its own optimum, and it brings a second benefit that is at least as valuable: at part speed the spools automatically settle at different speed ratios from those they hold at full power, which widens the range of conditions over which the compressor stays clear of stall and surge. High overall pressure ratios became practical only once multi-spool architecture was available to make them stable.

Each spool has to be supported on its own bearings, and a concentric arrangement means some of those bearings sit between two rotating shafts rather than between a shaft and static structure. These **inter-shaft bearings** see the *difference* between the two shaft speeds rather than either speed by itself, and both their races are moving, which makes their oil supply and their vibration signature more complicated than a conventional bearing's. It is one of the practical costs of a multi-spool engine and one of the reasons a vibration survey has to be interpreted against the engine's actual architecture.

What Happens if a Shaft Shears

Because a spool is a matched pair, breaking the shaft that joins them produces a violently asymmetric result. Consider the low pressure shaft of a turbofan shearing in flight. The **LP turbine** instantly loses its load — the fan it was driving — while keeping its full supply of hot gas, so the net accelerating torque on it becomes enormous and it accelerates with nothing to restrain it. This is **turbine runaway**, and it can reach burst speed within a couple of seconds. The **fan**, meanwhile, has lost its driving torque while still being loaded by the airflow through it, so it slows down. The two halves of one spool therefore move in opposite directions: turbine overspeed at the back, compressor underspeed at the front.

A turbine disc that bursts releases fragments with enough energy to leave the engine and the nacelle entirely, which makes this one of the small number of genuinely uncontained failure modes, so engines are protected against it. Electronic protection watches the rate of change of shaft speed, or watches for a mismatch between speeds that ought to be locked together, and shuts off the fuel. Mechanical protection is also used: some designs allow the unloaded turbine assembly to move rearwards under gas load so that its blades foul static structure and are braked before burst speed can be reached. The maintenance consequence is straightforward — an unexplained speed mismatch between the two ends of a single spool, or an unexplained vibration change accompanying it, must be treated as a possible shaft problem and never written off as an indication fault until that has been eliminated.

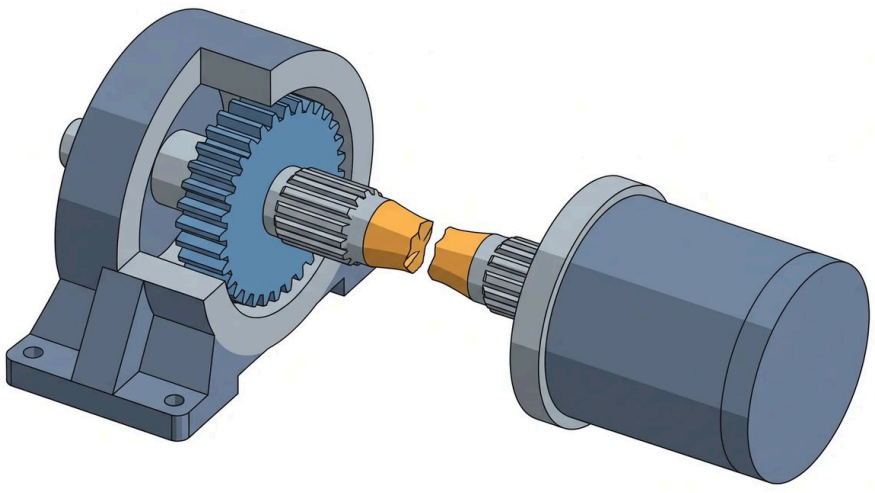
The Accessory Gearbox and Its Drive

The engine has to drive its own services and the aircraft's — fuel pumps, oil pressure and scavenge pumps, hydraulic pumps, electrical generators, and usually a dedicated alternator supplying the engine control system. All of them hang on an **accessory gearbox**, and the power for it is taken from the **high pressure**

spool through a bevel gear and a **radial drive shaft**, often called a tower shaft, running outwards through a strut in the engine casing.

The drive comes from the HP spool for two reasons and both matter. First, the HP spool has the narrowest operating speed range of any spool in the engine — it never runs slowly the way a fan can — so accessories driven from it see a far more consistent input speed than they would from the fan. A generator or a fuel pump driven off the fan would face a range of input speeds it could not usefully accommodate. Second, the HP spool is the one the starter has to turn to get the engine going, so routing the starter drive and the accessory drives through the same gearbox is the tidiest possible arrangement. On a high bypass engine that gearbox is mounted on the **high pressure compressor casing**, where the radial drive can reach it directly; on some large engines the drive is carried onwards by a transfer gearbox to a main gearbox slung lower on the core for accessibility, but the power still originates at the HP spool in every case.

Every accessory drive on that gearbox is a potential path along which damage could travel back into the engine, and the standard protection is a **waisted drive shaft**: a shaft machined with a deliberately reduced-diameter section over part of its length. If the driven unit seizes or is grossly overloaded, the shaft twists off at that waist instead of transmitting the load into the gearbox and the engine behind it. It is a mechanical fuse, and like an electrical fuse its whole purpose is to fail first, cheaply and predictably, so that something expensive and hard to reach does not fail instead. A waisted shaft is not there to save weight and not there to assist dynamic balance, and a sheared one is evidence about the accessory it was driving, not about the shaft itself — fitting a new shaft without establishing why the old one sheared simply arms the same failure again.



A waisted drive shaft: its narrowed section twists off if the accessory seizes, so the gearbox is spared.

Modular Construction

A **modular** engine is built as a set of self-contained assemblies — typically a fan module, a booster or intermediate module, a core module containing the high pressure compressor, combustor and HP turbine, a low pressure turbine module and an accessory drive module. The defining property is that any one of them

can be removed and replaced *without disturbing the rest of the engine*. The older, non-modular alternative requires the engine to be progressively stripped from one end in order to reach a component in the middle.

The advantage is directly financial. If one module has reached its life limit or has been damaged, only that module goes to the workshop; the rest of the engine remains serviceable, and a replacement module drawn from stores can be fitted and the engine returned to service in a fraction of the time and cost of a full strip and overhaul. It also allows modules of different ages to be combined in one engine, so that remaining life is used up rather than discarded along with the module that happened to fail first.

The price is bookkeeping and balance. Because modules move between engines, the maintenance records have to follow the *module* and not merely the engine serial number: hours and cycles are tracked per module, and life-limited parts inside a module are tracked individually within it. Rotating assemblies also have to be balanced as a module and, on many engines, checked or trim balanced after installation, because a module built correctly to its own drawing tolerances is not automatically in balance in combination with the modules on either side of it. One thing modular construction is *not* is a statement about how the engine is physically put together on a stand: building an engine vertically is a workshop convenience adopted by some manufacturers, and it has nothing to do with whether the engine is modular.

Balance, Critical Speeds and Why Vibration Is Watched

A machine whose only moving parts rotate ought to run perfectly smoothly, and the fact that it does not is worth understanding, because vibration is one of the few windows into an engine that stays open while it is running. Any rotor carries some residual out-of-balance: manufacturing tolerances, blade-to-blade weight variation, and later erosion or damage all leave the centre of mass fractionally off the axis of rotation. That offset generates a rotating force proportional to the offset mass, to its radius, and to the *square* of the rotational speed. The square is the important part — the same tiny imbalance that is undetectable at idle is significant at take-off speed.

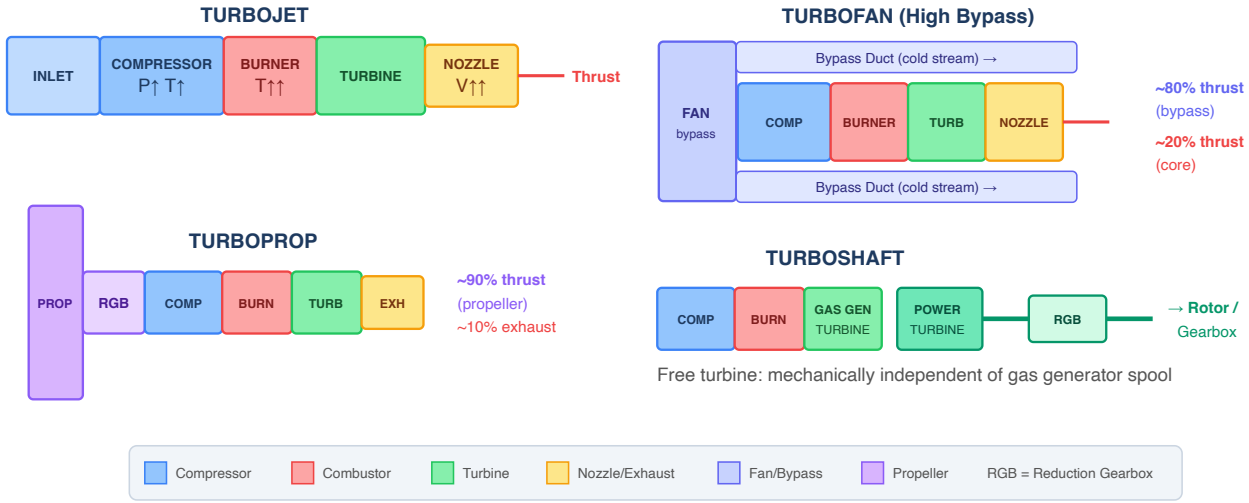
Rotors also have natural bending frequencies, and the speed at which a rotor's rotational frequency coincides with one of them is called a **critical speed**. At that speed the out-of-balance force is applied in step with the rotor's own tendency to flex, so the deflection builds instead of being resisted. Engines are designed so that no critical speed falls inside the normal continuous operating range, but a rotor may have to pass through one during acceleration and deceleration, and this is one of the reasons certain speed bands are marked for continuous operation to be avoided and why an engine is not left dwelling at an arbitrary speed during a ground run.

The practical consequences reach the flight line in three ways. Bearings are supported in **squeeze-film** oil damping arrangements on many engines, so that a thin film of oil between the bearing housing and the structure absorbs vibration energy rather than passing it straight into the casings. Fan and compressor blades are moved, matched or replaced in balanced sets, and a **trim balance** is carried out by adding small weights at a defined position after a blade change. And a recorded rise in the vibration level at a particular rotor's own

frequency identifies which spool is responsible, because each spool turns at its own speed and therefore signs its own contribution to the signal.

Gas Turbine Engine Types

All gas turbine engines share the same core — compressor, combustor, turbine — but differ in how they convert the energy extracted by the turbine into useful output. The four main types encountered in aviation are:



It helps to see those four as points on a single continuum rather than as four unrelated machines. The variable that orders them is how much air is worked on for every kilogram that passes through the core. A pure turbojet works on nothing but its core flow — an effective bypass ratio of zero. A low bypass military turbofan works on perhaps half as much again as its core swallows. A civil high bypass turbofan works on several times more. A turboprop, whose propeller disc is many times the diameter of its intake, works on an air mass equivalent to a bypass ratio of many tens to one. Move along that line and the propulsive device gets larger while the velocity it adds gets smaller, and the flight speed at which the arrangement stops being competitive comes down with it. Nothing else about the four types is as fundamental as that single trade, and the sections below are best read as four different resolutions of it.

Turbojet

The simplest gas turbine configuration. **All** of the air entering the intake passes through the core (compressor → combustor → turbine). The turbine extracts only enough energy to drive the compressor; all remaining energy is converted to a high-velocity exhaust jet that produces thrust. Turbojets are efficient at very high speeds (Mach 2+) but are **noisy and fuel-hungry** at subsonic speeds. They are rarely used on modern civil aircraft but are still found on some military aircraft and older designs.

"All of the air passes through the core" is the defining statement, and everything else about the turbojet follows from it. Because the core flow is small — tens of kilograms per second rather than the many hundreds a large turbofan passes — the only way to obtain useful thrust is to give that small flow a very large velocity increase. Jet exit velocities of the order of 600 m/s and above are normal, and at subsonic flight speeds that leaves an enormous velocity difference between the jet and the air it is being thrown into.

The mechanism behind the fuel consumption is the one set out in the worked example earlier in this note: the energy poured into the jet rises with the *square* of the velocity change while the thrust obtained rises only in proportion to it, so a fast jet is an expensive way to buy thrust at subsonic speed. The mechanism behind the noise is separate and just as unforgiving. Jet mixing noise — the roar of a high-velocity jet shearing into still air — rises with a very high power of the jet velocity, classically the eighth power, so a modest reduction in jet velocity produces a very large reduction in noise. Between them, those two relationships drove the turbojet off civil aircraft: noise legislation and fuel price were attacking the same design feature from two directions.

The picture reverses at high speed. As the aircraft's own velocity climbs, a slow bypass jet runs out of margin over the flight speed and can no longer add momentum usefully, while a turbojet's high jet velocity still leaves plenty of room to work in. Ram compression in the intake also becomes increasingly valuable, doing part of the compressor's job for nothing: the temperature rise it produces grows with the square of the flight Mach number, and the pressure ratio it produces grows faster still. Above roughly Mach 2 the turbojet, or a very low bypass turbofan, is the appropriate architecture. Above about Mach 3 the compressor and turbine become an encumbrance rather than a help, because ram compression alone is sufficient, and that is the point at which the ramjet takes over. Military installations also frequently add reheat downstream of the turbine for short periods of greatly increased thrust.

Maintenance considerations follow the architecture directly. A turbojet's entire mass flow passes through the hot section, so the hot section is worked hard for every newton produced and hot section life tends to dominate the maintenance programme. There is no cool bypass stream washing the outside of the core, so nacelle and jet pipe skin temperatures are higher, fire zone ventilation and fire protection are correspondingly more demanding, and the exhaust hazard area behind a running turbojet is both longer and hotter than behind a high bypass engine of the same thrust.

Turbofan

The dominant engine type in modern commercial aviation. A large **fan** at the front of the engine is driven by a turbine (usually the low-pressure turbine via a shaft running through the centre of the engine). The fan accelerates a large mass of air, most of which **bypasses** the core through an annular duct. The **bypass ratio (BPR)** is the ratio of bypass air mass flow to core air mass flow:

Bypass Ratio: $BPR = \frac{\dot{m}_{bypass}}{\dot{m}_{core}}$

Modern high-bypass turbofans (e.g., CFM LEAP, Rolls-Royce Trent XWB, GEnx) have BPRs of **9:1 to 12:1**, meaning 9–12 kg of air pass through the bypass for every 1 kg through the core. Approximately **75–85%** of total thrust comes from the bypass (cold) stream. High BPR engines are quieter and more fuel-efficient than turbojets because they produce thrust by moving a large mass of air at moderate velocity rather than a small mass at high velocity.

A fan is a propeller that has been placed inside a duct and given many more, much shorter, much stiffer blades. The duct is what makes it a different machine. It allows the blade tips to run at transonic speeds without the losses an unducted propeller would suffer there, it controls the flow entering and leaving the blades so that the working conditions along the span can be designed rather than accepted, and it lets the bypass stream be exhausted through a nozzle of chosen area. None of those is available to an open propeller, and together they are why a fan can be used at the high subsonic cruise speeds where a propeller cannot.

Worked Example — Where a turbofan's thrust actually comes from

A turbofan on a static take-off run passes a total of **600 kg/s** of air at a bypass ratio of **5:1**. The total splits as $600 \div (5 + 1) = 100$ kg/s through the core and $600 - 100 = 500$ kg/s through the bypass duct. Take the cold stream as leaving at 330 m/s and the hot stream at 480 m/s, with the aircraft stationary so that $V_0 = 0$:

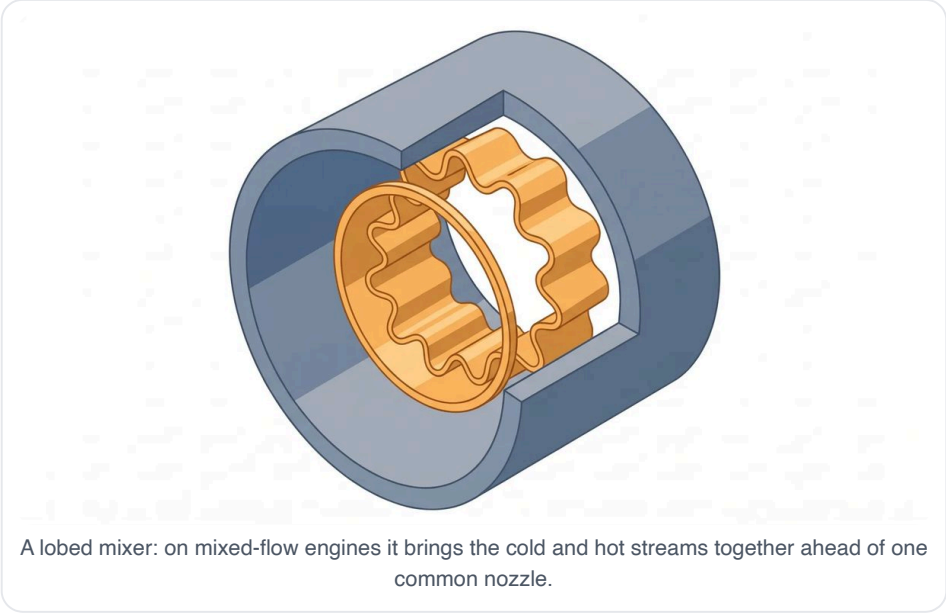
- Bypass (cold) stream thrust: $500 \times 330 = 165{,}000$ N
- Core (hot) stream thrust: $100 \times 480 = 48{,}000$ N
- Total: $213{,}000$ N, that is about **213 kN**

The bypass stream is producing $165{,}000 \div 213{,}000 = 77\%$ of the thrust from air that never sees a flame. Notice that it is doing so at a substantially *lower* velocity than the hot stream — the propulsive efficiency argument made concrete in one calculation. Repeat the arithmetic at a higher bypass ratio and the cold stream's share climbs further still, which is why a modern high bypass engine sits above the 77% obtained here. The core has become, in effect, a gas-fired engine whose real job is to drive the fan.

Turbofans are classified by bypass ratio, and the classification is useful because it predicts almost everything else about the engine. **Low bypass** engines, with ratios below about 2:1, are found on military aircraft and on older civil types; their jet velocity is high, they are slim, and they tolerate supersonic flight. **Medium bypass** covers roughly 2:1 to 4:1. **High bypass** begins around 5:1 and reaches the figures quoted above for current civil engines, with geared designs going higher again. As the ratio rises, the fan diameter grows, the jet velocity falls, fuel consumption and noise both improve, and the engine becomes progressively less suited to high subsonic and supersonic flight because its slow bypass jet has less and less margin over the aircraft's own speed.

The two streams may be exhausted separately or mixed. In a **separate-flow** or unmixed installation, typical of large high bypass engines, the cold stream leaves through its own annular nozzle around the core while the hot stream leaves through the core nozzle further aft. In a **mixed-flow** installation, usual on lower bypass engines and on many business-jet engines, the two streams are brought together in a lobed mixer ahead of a single common nozzle. Mixing recovers a small amount of thrust and reduces jet noise by evening out the

velocity profile leaving the engine, but it costs length, weight and duct pressure loss, which is why it is not universal.



A lobed mixer: on mixed-flow engines it brings the cold and hot streams together ahead of one common nozzle.

Driving the fan is a demanding job in its own right. The fan absorbs a large amount of power at a low rotational speed, while a turbine is happiest producing power at high speed, so the low pressure turbine has to be given a large annulus and several stages to extract that power at a speed the fan can accept. On a large conventional turbofan the low pressure turbine may have five or more stages and account for a substantial share of the engine's weight. The alternative is to insert a reduction gearbox between the low pressure turbine and the fan, so that each can run at its own best speed and the turbine can be made smaller, faster and lighter with fewer stages; that arrangement is dealt with among the alternative turbine constructions elsewhere in these notes.

Two further points are worth carrying to the flight line. First, the bypass duct air does considerably more than make thrust: it cools the outside of the core, it ventilates the fire zones, and on most large engines it is the working stream for the thrust reverser. An obstruction or a damaged duct panel is therefore a multi-system problem, not a simple thrust loss. Second, the fan is the largest rotating assembly on the aircraft and the first thing any ingested object meets. Fan blade condition, blend repair limits, blade root lubrication and fan trim balance after a blade change or a blade set rotation are routine and consequential tasks, and an out-of-limits vibration indication traced to the fan is far more common than one traced to the core.

One historical variant is worth knowing because it clarifies what the definition of a turbofan actually requires. Early bypass engines were sometimes built as **aft-fan** designs, in which the fan was placed behind the turbine rather than in front of the compressor and was driven not by a shaft but by an extra row of turbine blades formed on the outer ends of the fan blades themselves, with the hot gas driving the inner portion and the fan pumping cold air through the outer portion. The arrangement avoided a long shaft and kept the fan out of the way of the core, but it put the fan blade roots in the hot gas stream, which limited both the temperature and the life that could be achieved. Front-fan designs displaced it entirely. The point that survives is the definition: what

makes an engine a turbofan is the presence of a bypass flow driven by turbine power, not the position of the fan.

Turboprop

Uses the gas turbine core to drive a **propeller** through a **reduction gearbox** (RGB). The turbine extracts almost all available energy from the gas flow, leaving very little for exhaust thrust (typically only ~10%). The propeller generates ~90% of the total thrust by accelerating a very large mass of air at low velocity. Turboprops are very efficient at speeds below **400 knots** and at lower altitudes, making them ideal for regional airliners (e.g., ATR 72, Bombardier Dash 8) and utility aircraft.

The reason so little energy is left over for the exhaust is a deliberate design choice, not an accident of layout. Extra turbine stages are fitted behind the ones that drive the compressor, sized specifically to extract shaft power, and the more work the turbine takes out the less pressure and heat remain in the gas to make a jet. The residual jet thrust is nonetheless real, which means a turboprop's output cannot be described honestly by shaft power alone. The industry answer is **equivalent shaft horsepower** (ESHP), which adds the residual jet thrust — converted into an equivalent power at the aircraft's speed — to the shaft horsepower actually delivered to the propeller. Quoting only the shaft power understates the engine; quoting the engine as thrust alone is meaningless, because thrust and power are different quantities and cannot be exchanged without knowing the speed.

The reduction gearbox is not an optional refinement. A gas turbine's power turbine runs in the tens of thousands of revolutions per minute, while a propeller is limited to roughly one to two thousand by the speed of its blade tips. The tip sees the vector sum of its rotational speed and the aircraft's forward speed, so a propeller turning too fast has supersonic tips, loses efficiency sharply and becomes extremely noisy. The gearbox reconciles the two by trading speed for torque at very nearly constant power. That also makes it a heavily loaded, closely monitored assembly with its own oil supply, its own chip detectors and its own inspection schedule — on a turboprop, gearbox condition is as much a part of powerplant maintenance as the engine itself, and a metal indication from the gearbox is as serious as one from an engine bearing chamber.

Two shaft arrangements exist and it is worth knowing which one you are looking at. In the **fixed-shaft**, or direct-drive, arrangement drawn in the schematic above, the propeller is geared to the same shaft that carries the compressor and turbine, so propeller speed, compressor speed and turbine speed are locked together in a fixed ratio. In the **free-turbine** arrangement, a separate power turbine sitting downstream of the gas generator turbine drives the propeller through the gearbox and turns independently of it. The free turbine allows the gas generator to be started and run without turning the propeller, which makes starting easier and allows the propeller to be held stationary on the ground; the fixed shaft is mechanically simpler and gives a more direct relationship between power lever movement and engine response. Both arrangements are in widespread service, and neither is a modern replacement for the other.

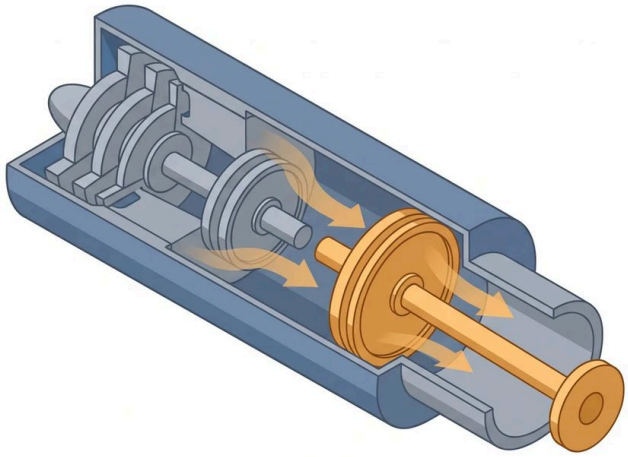
The propeller is also what sets the aircraft's practical speed limit, and the mechanism is the same tip-speed problem seen from the other end. Propeller efficiency holds up well while the blade tips remain comfortably subsonic, but the tip Mach number combines rotational and forward speed, so as the aircraft flies faster the tips approach the speed of sound no matter what the propeller itself is doing. Past that point efficiency collapses and noise climbs steeply. This is why the turboprop's advantage is confined to a lower speed band — and why, inside that band, it is genuinely and substantially more economical than a turbofan rather than merely cheaper to buy.

On the flight line the practical consequence is that a turboprop is set up and monitored differently from a turbofan. Power is set on **torque** rather than on a fan speed or an engine pressure ratio, for the reason given above: with the propeller governing at a constant speed, the torque gauge is reading power directly. The propeller and the engine are controlled as a single system, which means a fault that presents as an engine problem may in fact lie in the propeller control and the reverse is equally true, so troubleshooting has to treat the two together from the outset rather than clearing one before starting on the other.

Turboshaft

Functionally similar to a turboprop but the output shaft drives a **rotor system** (helicopter) or other machinery (ship, generator, pump) rather than a propeller. In a turboshaft, the engine is designed to produce **shaft horsepower** (SHP) with virtually zero residual jet thrust. Most turboshaft engines use a **free turbine** (also called a power turbine) — a separate turbine stage that is mechanically independent of the gas generator spool. This allows the power turbine to rotate at a different speed from the gas generator, which is essential for helicopter rotor speed control.

The mechanical independence of the free turbine is worth spelling out, because it is what makes a turboshaft practical for a helicopter at all. There is no shaft connecting the power turbine to the gas generator; the two are coupled only by the gas flowing from one into the other. That means the gas generator can be accelerated and decelerated to follow the power demand while the power turbine, and therefore the rotor, is held at constant speed by its governor — exactly what a helicopter needs, since rotor speed must stay within a narrow band whatever the collective is doing. It means the engine can be started, run and brought up to idle with the rotor stationary and the rotor brake applied. And in an autorotation the rotor goes on turning while the gas generator idles, but not because it is driving the power turbine round with it. The rotor drive carries a freewheel unit, a one-way clutch between the engine output shaft and the main gearbox, and the moment the rotor tries to turn faster than the engine is driving it that unit releases and disconnects the rotor from the engine altogether. Rotor speed then rises clear of power turbine speed on the gauges — the classic needle split — which is the opposite of the two being carried round together. It is that freewheel unit rather than the free turbine that makes autorotation possible: the same unit is fitted whatever the engine, which is why piston-engined and fixed-shaft turbine helicopters autorotate too. What the free turbine contributes is a light power turbine left carrying nothing once the drive has released, and a gas generator that can go on running at idle without the rotor having to turn it.



In a free turbine engine the power turbine has its own shaft, coupled to the gas generator only by the gas.

Because the two shafts are not mechanically connected, each needs its own speed indication, and the naming convention here differs from the turbofan convention. On a two-spool turbofan, N1 is the low pressure spool and N2 the high pressure spool. On many turboshaft and turboprop engines the *gas generator* spool is designated N1 or Ng, and the free power turbine is N2, Nf or Np. The labels are the same but the hardware they point at is not, so establish which engine architecture a manual or a question is describing before interpreting any speed designation on it.

"Virtually zero residual jet thrust" is a design requirement rather than a passive observation. On a helicopter, residual thrust from the exhaust would be a force acting in whatever direction the exhaust happened to point, which would have to be trimmed out and which would change with every power setting. Turboshaft exhausts are therefore arranged to minimise it, and where they are angled upwards or to the side the reason is clearance, infra-red signature or the avoidance of hot gas reingestion rather than any attempt to make thrust. The turbine is sized to leave as little energy in the gas as it usefully can, which is the exact opposite of the turbojet's sizing philosophy despite both engines using the same kind of core.

The same architecture is used well beyond rotorcraft. An auxiliary power unit is a small turboshaft whose load is an electrical generator and a bleed-air supply rather than a rotor. Marine propulsion, pipeline pumping and emergency generating sets use larger versions of the same idea. In every case the engine is rated in power rather than thrust, its health is assessed by a power assurance check that compares achieved torque and gas generator speed against the ambient conditions of the day rather than by any thrust measurement, and its operating limits are expressed as torque, gas generator speed and turbine temperature.

Example — Engine Types in Service:

- **Turbojet:** General Electric J79 (F-4 Phantom), Rolls-Royce Olympus 593 (Concorde)
- **Turbofan:** CFM International LEAP-1A (A320neo), GE90-115B (Boeing 777)
- **Turboprop:** Pratt & Whitney Canada PW127 (ATR 72), GE H80 (L-410)
- **Turboshaft:** Safran Arriel 2E (Airbus H145), GE T700 (Black Hawk)

Choosing Between Them

Comparing the four is easiest if the comparison avoids the properties already described and concentrates instead on how each one is *rated*, *geared* and *maintained*, because those are the differences an engineer meets first.

Type	Propulsive device	Turbine sized to extract	Output rated in	Reduction gearing	Dominant maintenance driver
Turbojet	The propelling nozzle alone	Only what the compressor and accessories need	Force (N, daN or lbf)	None	Hot section life, since the whole flow is hot flow
Turbofan	Ducted fan, plus the core jet	Compressor, accessories, and the fan as well	Force (N, daN or lbf)	None on a conventional design; a fan drive gearbox on a geared one	Fan and low pressure system condition alongside the hot section
Turboprop	Unducted propeller, plus a small residual jet	Compressor, accessories, and almost all remaining energy	Power (kW, SHP or ESHP)	Always, within the powerplant	Propeller and reduction gearbox as well as the engine
Turboshaft	An output shaft driving an external load	Compressor, accessories, and essentially everything left	Power (kW or SHP)	Always, within the engine, the airframe transmission, or both	Power assurance results, free turbine and transmission condition

The choice between them in service is not a question of one being better than another. It is a question of matching the propulsive device to the mission, and the drivers are usually practical rather than thermodynamic: the length and surface of the runways to be used, the altitude and temperature of the airfields, the sector length, the local noise rules, the acquisition and maintenance cost the operator can carry, and the performance required after an engine failure. A turboprop's low-speed thrust and short field performance can matter far more to a regional operator than the cruise speed it gives up, while a long-haul operator will pay for cruise efficiency and accept a longer take-off run.

One consequence of the common core is worth knowing, because it explains a great deal about parts commonality and about type training. Manufacturers routinely develop a single gas generator into several

products: the same core, in different states of tune, may appear in a regional turbofan, an industrial power generation set and a marine engine. Modules, accessories, control philosophies and even maintenance practices can then carry across between apparently unrelated applications, which is why an engineer who understands one core has a substantial head start on any engine built around it.

⚠ **Exam Note:** Do not confuse a turbofan with a turboprop. A turbofan's fan is *ducted* (enclosed in a nacelle) and operates at transonic tip speeds; a turboprop's propeller is *unducted* and operates at subsonic tip speeds. The reduction gearbox ratio for a turboprop is much higher (typically 15:1 to 20:1) than for a geared turbofan (~3:1) because propellers turn much slower than fans.

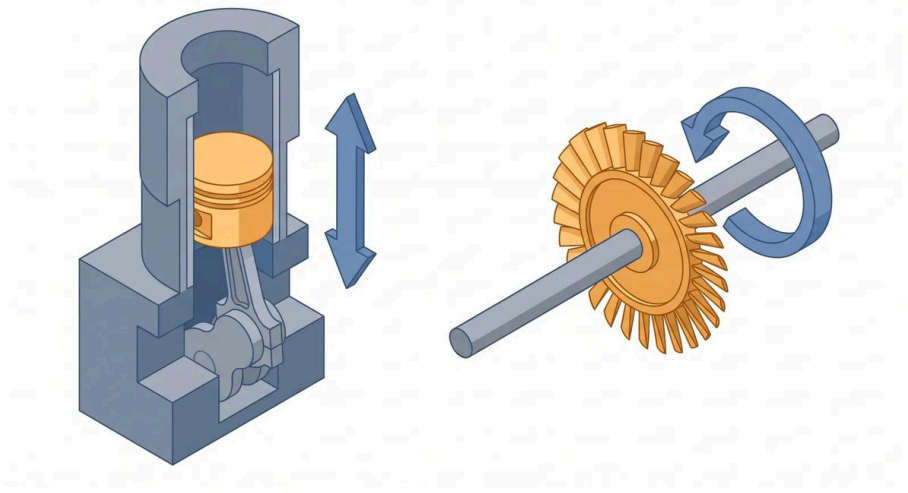
The Gas Turbine Compared with the Piston Engine

The gas turbine did not replace the piston engine on large aircraft because of the cycle alone. It replaced it because of a set of mechanical and operational consequences that follow from the cycle, and setting the two side by side is the quickest way to see which properties are fundamental and which are incidental.

Characteristic	Piston engine	Gas turbine
Motion of the working parts	Pistons reverse direction twice per revolution; a crankshaft converts that to rotation	Rotation only, continuously and in one direction
Combustion	Intermittent, one burn per cylinder every two revolutions on a four-stroke	Continuous, in a flame that is never extinguished in normal running
Valves in the gas path	Inlet and exhaust valves, cams, springs and their drives	None at all
Vibration	Inherent, from reciprocating inertia and from separate firing impulses	Low, arising only from residual rotor imbalance and aerodynamic excitation
Power to weight	Modest	High, and it improves as the engine gets larger
Frontal area per unit of power	Large	Small
Fuel	Aviation gasoline, tightly specified for octane rating and detonation resistance	Kerosene, less volatile, safer to handle and tolerant of a wider specification
Principal limiting parameter	Cylinder head temperature and the margin against detonation	Turbine gas temperature and rotor speed
Response to a power lever movement	Rapid	Lagged, because rotor inertia must be overcome and surge margin preserved
Efficiency at low power	Falls modestly below its best	Falls sharply, because pressure ratio and turbine temperature both drop away

Characteristic	Piston engine	Gas turbine
Usual maintenance philosophy	Fixed overhaul life	On condition, driven by trend monitoring and modular replacement

The most-quoted advantage is smoothness, and the reason for it is purely mechanical: **a gas turbine has no reciprocating parts**. In a piston engine every piston is brought to a complete standstill and reversed twice per revolution, and the inertia forces required to do that are large, cyclic and impossible to balance out entirely; the firing impulses then arrive at the crankshaft as a series of separate blows. A gas turbine's rotating assemblies simply turn, at a steady speed, with a steady flow of gas through them. What vibration remains comes from residual imbalance in the rotors and from aerodynamic excitation, not from the working cycle itself. That is the root of the very low healthy vibration baseline described earlier, and it is what makes a turbine's vibration signature such a sensitive instrument.



A piston stops and reverses twice per revolution; a turbine rotor simply turns, always in one direction.

It is worth being explicit about what the smoothness does *not* come from, because the alternatives are offered as examination answers. It does not come from better lubrication: a turbine's oil system is excellent, but a well-lubricated piston engine still shakes. And it does not come from running at a lower temperature: a gas turbine's hot section runs very much hotter than any piston engine's cylinder head. The absence of reciprocating masses is the principal reason.

Several operational differences follow from the same root. A gas turbine has no mixture control, because the fuel is metered into a continuously burning flame rather than into a measured charge of air, and the combustor runs with a very large excess of air at all times. It has no ignition system running continuously either: the igniters fire only during starting and when continuous ignition is deliberately selected for protection against flame-out in heavy rain, turbulence or icing conditions. And it responds to a power lever movement with a lag rather than immediately, because the fuel control has to accelerate the rotors without pushing the compressor into surge — which is why turbine-powered aircraft are flown with the spool-up time anticipated rather than discovered.

Starting shows the same divergence. A piston engine needs only to be turned fast enough for the first charge to fire, after which it is self-sustaining almost immediately. A gas turbine has to be driven up to a speed at which the compressor is delivering enough air for combustion to be stable and at which the turbine can produce more work than the compressor absorbs; below that speed the engine cannot sustain itself, so the starter must keep assisting well after the fuel has lit and the flame has established. That is why a turbine start is a timed, monitored sequence with a defined light-up point and a defined starter cut-out speed, rather than a single event. It is also why the failure modes at start have their own names and their own distinct causes, and why the fundamental parameters watched during a start are gas generator speed and turbine temperature — the two quantities that say whether the compressor is delivering and whether the combustor is behaving.

The gas turbine's weaknesses are real and they are the reverse side of the same characteristics. It is thermodynamically poor at low power settings, because both the pressure ratio and the turbine entry temperature fall away as the engine is throttled back, and the ideal-cycle relationship set out earlier shows immediately what that costs. It is expensive to manufacture, because its hot section parts are made from costly alloys by costly processes. It is intolerant of the sort of neglect a piston engine will survive, since a small quantity of the wrong debris in the gas path can end an engine's life in seconds. And its efficiency advantage depends on flying fast and high, which is why the piston engine remains entirely sensible in a light aircraft cruising at 120 knots and entirely hopeless in an airliner.

One habit of mind carries forward from all of this into the rest of the module, and it is the practical reason the physics in this note matters. A gas turbine is normally assessed without being opened: its condition is *inferred* from a handful of parameters read at the flight deck and on a test run. Because the cycle behaves in predictable ways, those inferences are sound. If a compressor becomes fouled or a turbine's tip clearances open up, its component efficiency falls; the cycle must then burn more fuel to reach the turbine work the engine is being asked for, so for the same thrust the exhaust gas temperature **rises** and the margin to the temperature limit shrinks. Nobody has to see the blades to know that. Every diagnosis of that kind is an application of the four processes, the gas laws and Newton's laws set out above, which is why they are worth understanding rather than merely memorising.

Module 16 — Piston Engine

16.1 — Fundamentals

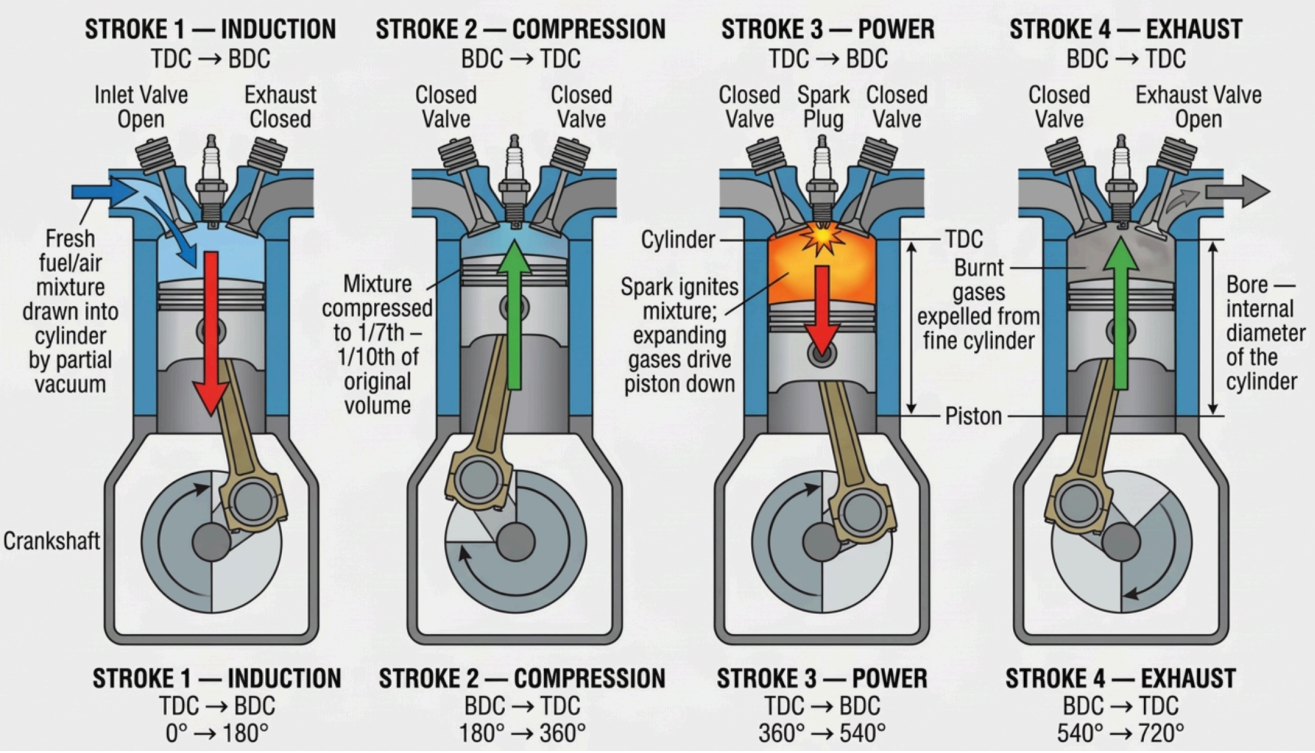
Introduction

The **piston engine** (reciprocating engine) converts the chemical energy in aviation fuel into mechanical energy through a repeated cycle of combustion events inside cylinders. Despite being one of the oldest forms of aircraft powerplant, piston engines remain the dominant choice for light aircraft, trainers, and many general aviation types. Every Part 66 engineer working on these aircraft must have a thorough understanding of piston engine theory.

Operating Principles

The Four-Stroke (Otto) Cycle

The vast majority of aircraft piston engines use the **four-stroke Otto cycle**. Each cylinder completes four distinct strokes (two complete crankshaft revolutions = 720°) to produce one power event:



Memory aid — "Suck, Squeeze, Bang, Blow": This classic mnemonic summarises the four strokes: **Induction** (suck mixture in), **Compression** (squeeze it), **Power** (bang — ignition and expansion), **Exhaust** (blow gases out).

Stroke-by-Stroke Detail

Stroke	Piston	Inlet Valve	Exhaust Valve	Crankshaft	Action
1 — Induction	TDC → BDC	Open	Closed	0° → 180°	Fresh fuel/air mixture drawn into cylinder by partial vacuum
2 — Compression	BDC → TDC	Closed	Closed	180° → 360°	Mixture compressed to 1/7th – 1/10th of original volume
3 — Power	TDC → BDC	Closed	Closed	360° → 540°	Spark ignites mixture; expanding gases drive piston down
4 — Exhaust	BDC → TDC	Closed	Open	540° → 720°	Burnt gases expelled from cylinder

Key terms:

- **TDC** — Top Dead Centre: piston at highest point in cylinder
- **BDC** — Bottom Dead Centre: piston at lowest point in cylinder
- **Stroke** — distance between TDC and BDC
- **Bore** — internal diameter of the cylinder

The Otto Cycle as a Thermodynamic Cycle

The four strokes describe what the hardware does. The **Otto cycle** is the thermodynamic idealisation of what the gas inside the cylinder does, and it is the form the examination tests. The ideal cycle has four processes, of which only two are strokes in the mechanical sense:

- **1 → 2 Isentropic compression.** The trapped charge is compressed with no heat entering or leaving it. Pressure and temperature both rise; the temperature rise comes entirely from the work done on the gas, not from any heat source.
- **2 → 3 Heat addition at constant volume.** The charge burns while the piston is assumed to be stationary at TDC, so the volume does not change while the heat is released. Pressure and temperature jump almost vertically on a pressure–volume diagram.
- **3 → 4 Isentropic expansion.** The high-pressure gas drives the piston down, doing work on it, again with no heat crossing the boundary.
- **4 → 1 Heat rejection at constant volume.** In the idealisation the residual heat is simply dumped at BDC. In the real engine this is the blowdown through the opening exhaust valve followed by the exhaust stroke.

The single defining feature, and the one the examiner is looking for, is that **the Otto cycle adds its heat at constant volume**. It is often described in one phrase as a **constant-volume combustion cycle**. Everything else about it — spark ignition, a pre-mixed charge, a modest compression ratio — follows from wanting the whole charge to burn in the very short window while the piston is near the top of its travel.

Air-standard Otto efficiency:

$$\eta_{ideal} = 1 - \frac{1}{r^{\gamma - 1}}$$

Where r is the compression ratio and γ is the ratio of specific heats of the working fluid, taken as 1.4 for air.

Note what is **not** in this expression: fuel, mixture strength, engine speed, cylinder count. In the ideal cycle the thermal efficiency of a spark-ignition engine depends on the compression ratio alone.

Worked example — the ideal ceiling at three compression ratios:

At $r = 7$:

$$\eta = 1 - 7^{-0.4} = 1 - 0.459 = 0.541 = 54.1\%$$

At $r = 8.5$:

$$\eta = 1 - 8.5^{-0.4} = 1 - 0.425 = 0.575 = 57.5\%$$

At $r = 10$:

$$\eta = 1 - 10^{-0.4} = 1 - 0.398 = 0.602 = 60.2\%$$

Raising the compression ratio from 7:1 to 10:1 lifts the ideal ceiling by about six percentage points. That is the whole reason designers push compression ratios upward, and the whole reason they are stopped by detonation before they get very far.

A real spark-ignition aero engine delivers around half of that ideal figure. The shortfall is not one large loss but a stack of unavoidable ones: heat conducted into the cylinder head and barrel while the gas is hottest; combustion that takes tens of crankshaft degrees rather than happening instantly at TDC; the exhaust valve opening before BDC so that some expansion work is thrown away to reduce the pumping effort of the exhaust stroke; the work of drawing the charge past the throttle butterfly and pushing the burnt gas out again; chemical dissociation at peak temperature; incomplete combustion; and, at high power, deliberate mixture enrichment used purely to cool the charge and the exhaust valve. On top of all of that, the practical figure quoted for an engine is normally referred to shaft power, so the friction of the engine itself is inside the gap as well.

Real Valve and Ignition Timing

The stroke table above places every event exactly at TDC or BDC. No engine is built that way. Gas has mass and therefore inertia, valves take time to lift and seat, and the charge takes a measurable time to burn, so every event is moved away from the dead centre a textbook cycle would put it at.

- **Inlet valve opens before TDC.** The valve is off its seat and flowing by the time the piston starts down, so the cylinder is not trying to draw through a barely-cracked valve during the fastest part of the induction stroke.
- **Inlet valve closes after BDC.** The incoming charge is moving at speed; its momentum keeps it flowing into the cylinder for a while after the piston has stopped descending and started back up. Holding the valve open captures that extra charge. This is also why the point of inlet closure, not BDC, marks the true start of compression.

- **Exhaust valve opens before BDC.** Some expansion work is deliberately sacrificed so that most of the cylinder pressure has blown down before the piston starts up. The pumping work saved on the exhaust stroke is worth more than the expansion work given away.
- **Exhaust valve closes after TDC.** The outgoing gas column has momentum too, and continues to scavenge the cylinder briefly after the piston has reached the top.

Because the inlet opens before TDC and the exhaust closes after it, there is a period around TDC at the end of the exhaust stroke when **both valves are open together**. This is **valve overlap**, and it is deliberate. The exhaust gas leaving at speed lowers the pressure at the top of the cylinder and helps pull fresh charge in; the fresh charge sweeping past the exhaust valve carries heat away from the hottest component in the combustion chamber. At the speeds an engine spends most of its life at, overlap is worth having.

The cost appears at the other end of the speed range. At idle the gas velocities are low, so instead of scavenging, overlap allows exhaust gas to be drawn back into the cylinder and inlet tract. The charge is diluted with inert burnt gas, combustion becomes less reliable, and the engine idles roughly and needs a richer mixture to fire at all. That is why a large-overlap engine is lumpy at idle and cleaner once the throttle is opened.

Ignition timing is displaced for the same reason. Burning a cylinder full of mixture is not instantaneous; the flame front takes an appreciable slice of crankshaft rotation to cross the chamber. To have the pressure peak arrive where it can do useful work — a little after TDC, of the order of ten to fifteen degrees, when the crank has enough leverage on the connecting rod — the spark must occur while the piston is still rising. On a fixed-timing magneto installation the spark is typically set somewhere in the region of twenty to twenty-five degrees before TDC, with the exact figure fixed by the engine type certificate data sheet and the engine maintenance manual, never by a rule of thumb.

Getting the timing wrong, in both directions:

- **Too far advanced** — combustion is largely complete while the piston is still coming up, so the peak pressure is higher and arrives earlier, opposing the piston. Cylinder head temperatures climb, exhaust gas temperature tends to fall because more of the heat is released into the cylinder rather than into the exhaust, and the end-gas is squeezed harder for longer, which is exactly the condition that provokes detonation.
- **Too far retarded** — the charge is still burning well down the expansion stroke. Peak pressure is lower and comes too late to do full work, so power falls, and the gas is still hot when the exhaust valve opens, so exhaust gas temperature rises and the exhaust valve runs hotter.

The two faults therefore move cylinder head temperature and exhaust gas temperature in opposite senses, which is one of the more useful diagnostic pairs available from the cockpit instruments.

Why the Power Delivery Is Not Smooth

A four-stroke cylinder produces one power stroke for every two crankshaft revolutions, so at 2,400 rpm each cylinder fires 1,200 times a minute. The torque delivered to the crankshaft is not steady: it is a series of impulses separated by three strokes during which that cylinder is absorbing work rather than producing it. Only two things smooth this out — the number of cylinders, which is why the power impulses are spread evenly

around the two revolutions, and the flywheel effect of the rotating masses, which on an aircraft engine is largely provided by the propeller itself.

This is worth holding on to, because it explains why the propeller on a direct-drive light aircraft engine is not simply a load on the output shaft. It is also the engine's flywheel: it is what carries the crankshaft through the three strokes in which any given cylinder is absorbing work rather than producing it, and it is a large part of why a piston engine will not idle at all with nothing substantial attached to its output shaft.

Charge Motion, Flame Travel and Dual Ignition

Nothing in the four processes of the ideal cycle says anything about how the charge is arranged inside the cylinder, yet that arrangement decides whether the engine burns cleanly or detonates. Two things matter: how the charge is moving when the spark occurs, and how far the flame then has to travel.

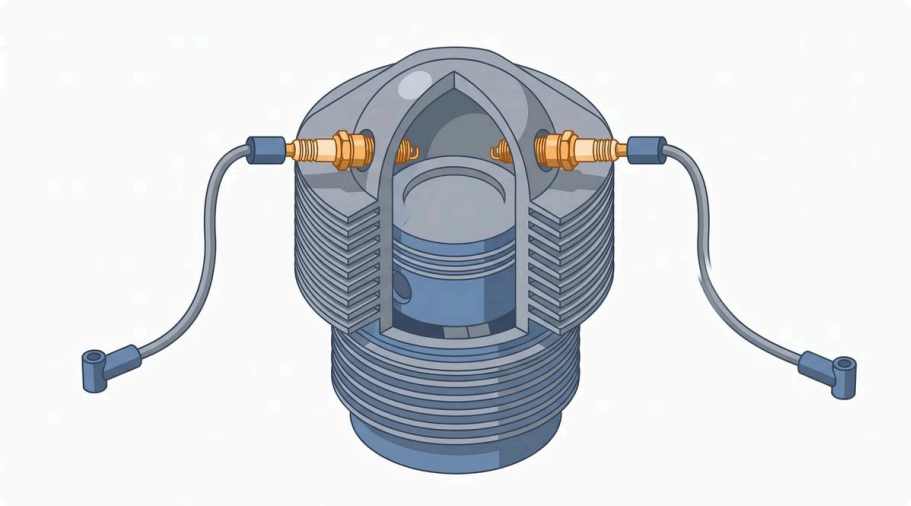
The charge does not enter the cylinder quietly. It arrives through the inlet valve at high velocity and is deliberately given organised motion by the shape of the port and the chamber:

- **Swirl** — rotation of the charge about the cylinder axis, generated by the direction the inlet port aims the incoming gas.
- **Tumble** — rotation about an axis across the cylinder, which breaks down into fine-scale turbulence as the piston rises.
- **Squish** — the charge squeezed out of the narrow gap between the flat part of the piston crown and the corresponding flat area of the head as the piston approaches TDC, producing a jet of gas across the chamber at exactly the moment ignition occurs.

All three exist for one reason: **a still charge burns far too slowly**. Turbulence wrinkles and stretches the flame front so that it consumes the mixture in a fraction of the time a smooth flame would take, which is what allows a cylinder full of mixture to be burnt within the few milliseconds available at 2,400 rpm. It also has a direct bearing on knock resistance, because the faster the flame reaches the last of the mixture, the less time that last portion spends being heated and compressed while waiting to be burnt.

That last portion has a name worth knowing: the **end gas**. It is the part of the charge furthest from the point of ignition, and it is compressed and heated both by the piston and by the expanding burnt gas behind the advancing flame. If it reaches its auto-ignition temperature before the flame arrives, it ignites spontaneously and all at once instead of being burnt progressively. That is detonation, and it is the phenomenon that sets the practical ceiling on the compression ratio of a spark-ignition engine.

Shortening the distance the flame must travel is therefore a design objective, and it is one of the two reasons every certified spark-ignition aircraft piston engine has **two spark plugs in every cylinder**:



Each cylinder carries two spark plugs on opposite sides of the chamber, each fed by its own ignition lead.

- **Redundancy.** Two complete and independent ignition systems, each with its own source of ignition energy, harness and set of plugs, mean that the failure of one leaves the engine running. A magneto is also self-generating, so the ignition system does not depend on the aircraft electrical supply once the engine is turning.
- **Combustion performance.** Two flame fronts starting from opposite sides of the chamber meet in the middle, so each has to travel only about half as far as a single flame would. The charge is burnt in less time, the end gas has less opportunity to auto-ignite, and combustion is more complete.

The second point is what makes the pre-flight magneto check meaningful. Selecting one magneto leaves the cylinder firing on a single plug, so the flame path is longer and the burn slower; combustion is finished later in the expansion stroke, less work is extracted, and the engine loses a little power. On a fixed-pitch installation that shows as a small drop in rpm, and the maintenance manual states both the maximum permitted drop and the maximum permitted difference between the two magnetos. **A drop that is too small, or none at all, is as much a defect as a drop that is too large** — it usually means the magneto being switched off is not actually being earthed, so the engine remains live even with the switch off. That is a serious hazard to anyone near the propeller, and it is caught by exactly this check.

Two plugs are not a spare pair. Both plugs in a cylinder fire on every cycle in normal operation; neither is a standby. That is why a single failed plug or lead produces a rough-running engine and a measurable power loss rather than no symptom at all, and why the magneto check compares the engine's behaviour on each system separately rather than simply confirming that it still runs.

The Two-Stroke Cycle

In a **two-stroke engine**, the cycle completes in just two strokes (one crankshaft revolution = 360°). The piston itself acts as a valve, covering and uncovering ports in the cylinder wall:

- **Stroke 1 (upward):** Piston compresses mixture above while drawing fresh charge into the crankcase below

- **Stroke 2 (downward):** Combustion drives piston down; near BDC, exhaust port uncovered, then transfer port admits fresh mixture from crankcase

Two-stroke engines have a **higher power-to-weight ratio** (one power stroke per revolution vs. one per two revolutions), but they are less fuel-efficient and produce more emissions. They are rarely used in certified aircraft but appear in some ultralight and microlight types.

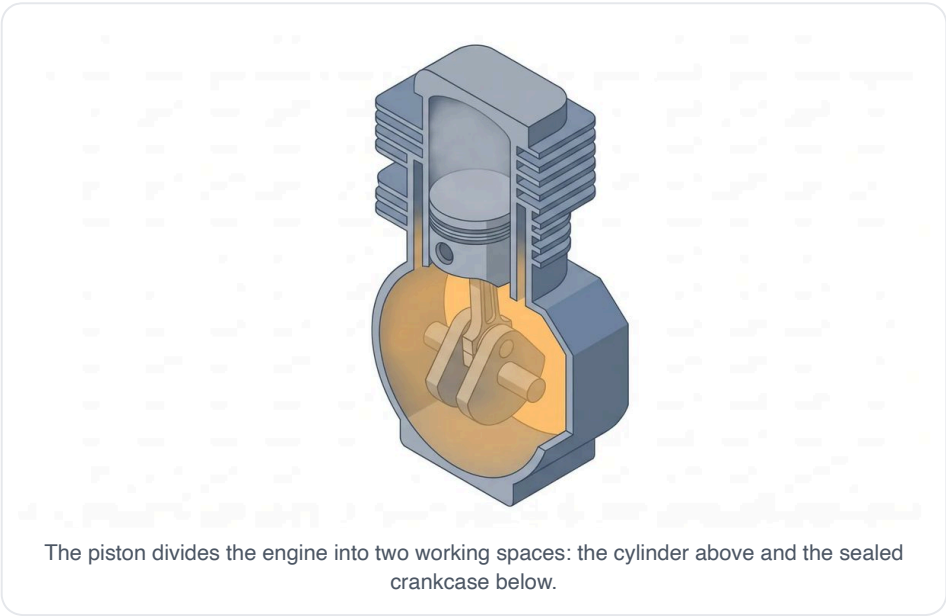
Inside the Two-Stroke: Ports, Transfer and Scavenging

The two-stroke engine gets rid of the camshaft, the pushrods, the rockers and the poppet valves, and replaces them with holes in the cylinder wall that the piston itself covers and uncovers. Everything else about the engine follows from that one decision, including all of its weaknesses.

Three ports do the work of the valve train:

- **Inlet port** — admits fresh mixture into the **crankcase**, not into the cylinder. It may be uncovered by the piston skirt, or controlled by a reed valve or a rotary disc valve driven off the crankshaft.
- **Transfer port** — a passage cast into the cylinder that carries the mixture from the crankcase up into the cylinder above the piston.
- **Exhaust port** — uncovered by the piston crown near the bottom of its travel, before the transfer port opens.

The crankcase is therefore not an oil reservoir but a pump. As the piston rises it creates a depression in the sealed crankcase below it, which draws the fresh charge in through the inlet port; as the piston descends after combustion, it pressurises that same crankcase and pushes the charge up the transfer passage into the cylinder. Every two-stroke engine of this type has two working spaces separated by the piston, and both matter.



The piston divides the engine into two working spaces: the cylinder above and the sealed crankcase below.

That has an immediate and far-reaching consequence for lubrication: **a crankcase being used as a pump cannot also hold a sump of oil**. The lubricant must therefore travel with the charge, either pre-mixed with the fuel at a ratio commonly around 50 parts fuel to 1 part oil, or metered into the induction by a small injection pump. It is burnt with the charge and lost out of the exhaust, which is why this arrangement is called **total-loss**

lubrication. The engineer sees the consequences directly: oil consumption is by design rather than by defect, the exhaust smokes, spark plugs and exhaust systems accumulate deposits, and there is no oil pressure system of the kind a four-stroke engine is monitored by.

Why Two-Stroke Scavenging Is Difficult

In a four-stroke engine the exhaust stroke pushes the burnt gas out and the induction stroke pulls the fresh charge in, as two separate events. A two-stroke has to do both at once, in the short window near BDC when the ports are open, using nothing but the pressure difference between the crankcase and the cylinder. Getting the burnt gas out without simply blowing the fresh charge out with it is the central problem of the design, and it is never solved completely.

- **Deflector piston (cross) scavenging** — an early solution in which the piston crown carries a raised deflector that turns the incoming charge upward, away from the exhaust port. Simple, but the deflector adds mass and creates a hot spot on the crown.
- **Loop (Schnuerle) scavenging** — the transfer ports are angled so that the incoming charge sweeps up the far wall, across the top of the cylinder and back down towards the exhaust port, pushing the burnt gas ahead of it. It allows a flat piston crown and is the normal arrangement on modern small two-strokes.
- **Uniflow scavenging** — the charge enters through ports at the bottom and leaves through a valve in the head, so the gas travels one way only. It scavenges best and is used in large two-stroke compression-ignition engines, at the cost of reintroducing a valve and its drive.

Because the piston opens and closes every port, the port timing is symmetrical about BDC: a port opens the same number of degrees before BDC as it closes after it. The exhaust port opens first and therefore also closes last, which means **the exhaust port is still open when the transfer port has already shut**. Some fresh charge inevitably escapes straight out of the exhaust. This short-circuiting is why a two-stroke burns more fuel per horsepower and emits more unburnt hydrocarbon than an equivalent four-stroke, and why it is the four-stroke that has the advantage in fuel efficiency and emissions.

Two further effects follow from the ports being in the cylinder wall. The charge is never fully compressed over the whole geometric stroke, because compression cannot begin until the exhaust port is covered on the way up; the ratio actually achieved on the trapped charge is therefore lower than the geometric compression ratio the dimensions suggest. And a proportion of burnt gas always remains behind to dilute the next charge, so combustion is less clean and less repeatable than in a well-scavenged four-stroke.

The tuned exhaust: a two-stroke expansion chamber is not a silencer with a fancy shape. Its diverging cone sends a negative pressure wave back to the exhaust port to help pull the burnt gas out, and its converging cone returns a positive wave timed to push escaping fresh charge back into the cylinder just before the port closes. It works only over a narrow band of engine speeds, which is one reason a two-stroke has a much narrower usable power band than a four-stroke, and why the exhaust system on such an engine is a tuned component that must not be substituted or modified.

Why a Two-Stroke Does Not Give Twice the Power

The arithmetic looks compelling: twice as many power strokes per revolution ought to mean twice the power from the same capacity. In practice a well-developed two-stroke produces appreciably more power than an equal-capacity four-stroke, but well short of double, for four separate reasons that all pull the same way:

- Part of the fresh charge is lost out of the exhaust port before it closes, so less of it is trapped and burnt.
- The charge that is trapped is diluted with residual exhaust gas, which lowers the peak pressure it can produce.
- The effective compression and expansion strokes are shortened by the height of the ports, so less work is extracted per event.
- The crankcase pumping is itself a parasitic load on the engine.

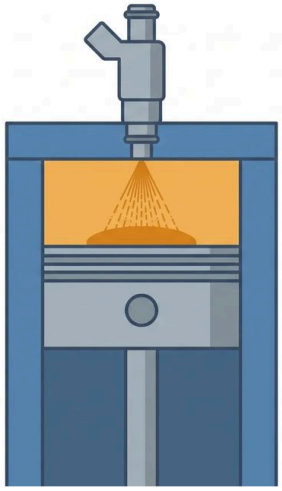
Set against that, the design removes the entire valve train — camshaft, tappets, pushrods, rocker gear, valve springs and the accessory drive to turn them. It is that removal, more than the extra firing events, which gives the two-stroke its reputation for a high power-to-weight ratio. It also removes the parts that most commonly need scheduled attention, which is why two-stroke engines are attractive to the ultralight and light-sport market where simplicity and installed weight matter more than fuel burn or overhaul life.

Feature	Four-stroke	Two-stroke
Crankshaft revolutions per power stroke	2	1
Charge admission	Poppet valves driven by a camshaft	Ports uncovered by the piston, sometimes with a reed or rotary valve
Lubrication	Pressure fed from a sump or dry-sump tank, oil recirculated and filtered	Total loss where the crankcase is the pump — oil pre-mixed with the fuel or injected into the induction
Specific fuel consumption	Lower	Higher, because of charge short-circuiting
Exhaust emissions	Lower	Higher, including unburnt fuel and oil
Installed weight for a given power	Higher	Lower
Usable speed range	Wide	Narrow, and dependent on a tuned exhaust
Typical aircraft application	Certified light aircraft of every class	Ultralight, microlight and light sport types

Manufacturers serving the light end of the market often build both. Rotax, for example, is a single manufacturer whose range has included both two-stroke and four-stroke aircraft engines, so the name alone tells an engineer nothing about which cycle is in front of him — the model designation and the engine data plate do.

The Diesel (Compression Ignition) Cycle

In a **diesel cycle** engine, there is no spark plug. Air alone is compressed to a very high ratio (typically 15:1 – 22:1), raising its temperature above the ignition point of the fuel. Fuel is then **injected directly** into the cylinder near TDC, where it auto-ignites on contact with the hot compressed air.



Air alone is squeezed until it is hotter than the fuel's ignition point, and the injected fuel lights itself.

Feature	Otto (Spark Ignition)	Diesel (Compression Ignition)
Compression ratio	7:1 – 10:1	15:1 – 22:1
Ignition method	Spark plug	Auto-ignition (compression heat)
Fuel type	AVGAS (100LL)	Jet-A / diesel
Mixture formation	Carburettor or fuel injection (pre-mixed)	Direct injection into cylinder
Thermal efficiency	~25–30%	~35–42%
Weight	Lighter	Heavier (stronger construction needed)
Aircraft examples	Lycoming O-360, Continental IO-550	SMA SR305, Continental CD-300

The Diesel Cycle as a Thermodynamic Cycle

Just as the Otto cycle is the idealisation behind the spark-ignition engine, the **Diesel cycle** is the idealisation behind the compression-ignition engine, and the two differ in exactly one process:

- **1 → 2 Isentropic compression** of **air alone**. No fuel is present, so there is nothing in the cylinder that can ignite itself no matter how hard it is squeezed.
- **2 → 3 Heat addition at constant pressure.** Fuel is injected as the piston begins to move away from TDC. The rate of burning is matched to the rate at which the volume is increasing, so the pressure stays approximately constant while the heat is released.
- **3 → 4 Isentropic expansion** of the products of combustion.
- **4 → 1 Heat rejection at constant volume**, representing blowdown and exhaust.

So the examination distinction is a single sentence: **the Otto cycle adds heat at constant volume; the Diesel cycle adds heat at constant pressure.** Everything the two engines differ about physically follows from that. The Otto engine must have its whole charge present and mixed before ignition, so it can only compress as hard as the fuel will tolerate without igniting itself. The Diesel engine keeps the fuel out until the moment it is wanted, so the compression ratio is limited by mechanical strength rather than by the fuel, and can go far higher.

> **A comparison that is easy to state backwards:** at the **same** compression ratio the Otto cycle is the more efficient of the two, because adding heat at constant volume releases it at the smallest volume and therefore at the highest possible pressure and temperature, whereas the Diesel cycle is still releasing heat while the volume is growing. The compression-ignition engine nevertheless wins in practice, because it is not competing at the same compression ratio — it runs at roughly twice the ratio of a spark-ignition engine, and the efficiency gained from that far outweighs the efficiency lost to constant-pressure heat addition. Both effects are real; the compression ratio is the one that decides the outcome.

The real engine sits between the two idealisations. Injection is not slow enough to hold the pressure exactly constant, so part of the charge burns almost at constant volume and the rest at roughly constant pressure. The cycle that describes this properly is the dual or mixed cycle, and every modern high-speed compression-ignition engine follows it more closely than it follows the textbook Diesel cycle. For the examination, however, the constant-pressure answer is the one that identifies the Diesel cycle.

What a Compression-Ignition Engine Demands of the Installation

An aviation diesel is not a petrol engine with the spark plugs removed. Compressing air to a ratio in the region of 15:1 to 22:1 and then burning fuel in it produces peak cylinder pressures of the order of two to three times those in a spark-ignition engine of comparable size, and almost every distinctive feature of the installation exists to deal with that:

- **Structural weight.** The crankcase, crankshaft, bearings, connecting rods and cylinders must all be built for those pressures. This is the direct reason for the extra weight noted in the comparison table above, and it is why the claim that a diesel is always the lighter installation is false.
- **A reduction gearbox.** A propeller reduction gearbox is common rather than exceptional on this type of engine, and the reason lies with the propeller rather than with the engine: propeller tip speed sets a firm ceiling on the speed at which a propeller may be turned, and where an engine is designed to run above that ceiling its power can reach the shaft only through gearing. Most certificated aviation diesels are in that position; those that are not, such as the SMA SR305 named in the comparison table above, are designed to turn slowly enough to drive the propeller directly. With the gearbox come torsional vibration considerations, gearbox oil, chip detection and its own inspection schedule.
- **Turbocharging.** Compression-ignition aero engines are almost always turbocharged. There is no throttle butterfly restricting the air path, so the engine takes to boosting readily, and boosting is also what restores sea-level power at altitude.
- **Electronic control.** Injection timing and quantity have to be scheduled precisely against speed, load and temperature, so full authority digital engine control with a single power lever is the norm. Mixture and

propeller controls of the traditional kind are absent.

- **High-pressure fuel systems.** Injection pressures are far above anything in a carburetted installation. The safety precautions, the pipe standards and the leak-check procedures are correspondingly different, and an injector or rail must never be treated like a low-pressure fuel line.

Glow plugs do not ignite the charge. Many compression-ignition engines carry glow plugs or an intake air heater, and it is easy to mistake these for an ignition system. They are a **starting aid** only: with a cold engine the compressed air loses too much heat to cold cylinder walls to reach the fuel's auto-ignition temperature, so the heater raises the starting temperature enough to get the first firings. Once running, the engine ignites its fuel purely by the heat of compression, and the heaters switch off. The ignition method of a compression-ignition engine is the heat of compression, at every stage after the start.

Why Jet-A Is the Attraction

The fuel is often the commercial reason for choosing a compression-ignition engine in the first place. Jet-A and Jet A-1 are available at essentially every airfield that handles turbine traffic, whereas leaded aviation gasoline is a shrinking, specialised supply in many parts of the world and faces continuing pressure over its lead content. Kerosene is also far less volatile than gasoline, so it produces much less flammable vapour at ambient temperature — a genuine safety advantage in handling, storage and in a survivable accident.

The trade-offs are equally practical, and an engineer should know them:

- Kerosene begins to form wax crystals at low temperature, which is a real consideration for high-altitude and cold-soak operation; anti-icing additive requirements come from the airframe and engine manuals, not from custom.
- The fuel itself lubricates the injection equipment, so fuel lubricity and cleanliness matter far more than they do to a carburettor. Water and particulate contamination damage injection systems quickly.
- Kerosene is denser than gasoline, so a given tank volume is a greater mass. Loading and centre of gravity calculations must use the correct density for the fuel actually uplifted.
- The low volatility that makes kerosene safe also makes cold starting harder, which is exactly why the starting aids described above exist.

Cycle and Ignition Method Are Independent Choices

One of the most common conceptual errors in this subject is to treat "two-stroke" and "diesel" as if they described the same axis. They do not. The **number of strokes** is a question about how the cylinder is filled and emptied; the **method of ignition** is a question about how the charge is lit. All four combinations exist:

	Spark ignition (Otto)	Compression ignition (Diesel)
Four-stroke	The mainstream general aviation engine	The modern aviation diesel
Two-stroke	Ultralight, microlight and light sport engines	Uniflow-scavenged designs, dominant in large marine engines and used in a small number of aviation designs

Examination distractors exploit this constantly, offering options such as "two-stroke diesel cycle" where the intended answer is simply the four-stroke Otto cycle. Read the two halves of such an option separately: one may be right while the other is wrong, and the option is only correct if both halves are.

Watch for invented terminology. Questions on these two cycles are often built around plausible-sounding phrases that name nothing at all — "constant density combustion" or "variable volume combustion". Only two descriptions belong to the cycles in this section, constant **volume** for Otto and constant **pressure** for Diesel, so an option using any other adjective can be discarded on sight regardless of how reasonable the rest of it sounds.

Piston Displacement and Compression Ratio

Piston Displacement (Swept Volume)

Piston displacement is the total volume swept by all pistons in one complete stroke from TDC to BDC. It is a primary indicator of engine size and power potential.

Single cylinder swept volume:

$$V_{cyl} = \frac{\pi}{4} \times d^2 \times L$$

Total engine displacement:

$$V_{total} = V_{cyl} \times N$$

Where: d = bore diameter, L = stroke length, N = number of cylinders

Example: Lycoming O-360 — 4 cylinders, bore = 5.125 in, stroke = 4.375 in

$$V_{cyl} = \frac{\pi}{4} \times 5.125^2 \times 4.375 = 90.3 \text{ cu in}$$

$$V_{total} = 90.3 \times 4 = 361.2 \text{ cu in} \approx 360 \text{ cu in}$$

This is why it is called the "O-**360**" — the number refers to the displacement in cubic inches.

Bore, Stroke and the Bore/Stroke Ratio

Displacement fixes how much air an engine can swallow, but the same displacement can be built from a large bore with a short stroke or a small bore with a long stroke, and the choice changes the engine's character completely. The comparison is described by the **bore/stroke ratio**, and the terminology is examined directly:

- **Over-square** — the bore is larger than the stroke, so the ratio is greater than 1.
- **Square** — bore and stroke are equal.
- **Under-square** (also called long-stroke) — the stroke is larger than the bore, so the ratio is less than 1.

Two real engines:

- The engine in the worked example above, bore 5.125 in and stroke 4.375 in: ratio = 5.125 ÷ 4.375 = **1.17**, so the engine is over-square.
- A four-cylinder Rotax 912 has a bore of 79.5 mm and a stroke of 61 mm: ratio = 79.5 ÷ 61 = **1.30**, also over-square. Its displacement follows from the same formula as before:

$$V_{cyl} = \frac{\pi}{4} \times 79.5^2 \times 61 = 302{,}800 \text{ mm}^3 = 302.8 \text{ cm}^3$$

$$V_{total} = 302.8 \times 4 = 1{,}211 \text{ cm}^3$$

  which is exactly the 1,211 cc quoted for the engine.

Most modern aircraft piston engines are over-square, for the reasons set out below.

The dominant reason to keep the stroke short is **mean piston speed**. It is a measure of how fast the piston and its rings slide along the cylinder wall, and of how far they travel in every minute of running, so it bears directly on how quickly the bore, the rings and the piston skirt wear. Mean piston speed depends only on stroke and engine speed:

Mean piston speed:

$$v_p = 2 \times L \times N$$

Where L is the stroke and N is the crankshaft speed. The factor of 2 is there because the piston covers the stroke twice in each revolution — once up and once down.

Worked example — what the stroke costs:

An engine with a 4.375 in stroke at 2,700 rpm:

$$v_p = 2 \times \frac{4.375}{12} \times 2700 = 1{,}969 \text{ ft/min}$$

Build the same displacement with a 5.125 in stroke instead, and at the same 2,700 rpm:

$$v_p = 2 \times \frac{5.125}{12} \times 2700 = 2{,}306 \text{ ft/min}$$

That is 17% more rubbing speed at the rings and 17% more piston travel per minute, for exactly the same power output. The inertia loading has not risen with it: the reciprocating inertia force is the piston mass multiplied by the crank throw and by the square of the **crankshaft** speed, the crankshaft speed is unchanged at 2,700 rpm, and the 17% longer throw is offset by the smaller and therefore lighter piston that holding the displacement constant now demands. It is the sliding speed and the distance travelled that have gone up, and that is why an over-square layout is the usual choice where an engine must run for thousands of hours between overhauls.

A large bore brings a second advantage that connects directly to the efficiencies discussed later in this note: it provides more area in the cylinder head, so larger valves can be fitted, and larger valves let the cylinder breathe better at a given speed. A long-stroke engine, by contrast, gains a longer crank throw — but at equal displacement that buys no extra torque, because the smaller bore takes exactly as much area off the piston crown as the longer throw adds in leverage, and the torque a given mean effective pressure produces is set by

the displacement alone. What the long stroke changes is where the torque peak falls on the speed scale: the smaller bore forces smaller valves, so the cylinder stops filling well at a lower speed, and the best torque is therefore made lower down the range. That is the sense in which a long-stroke layout favours torque at low speed. It is also physically taller, which matters in a cowling.

Aero engines are deliberately slow-turning by road-vehicle standards. Direct-drive engines are commonly limited to somewhere below three thousand rpm, and the binding constraint is usually not the engine at all but the **propeller tip speed**, which must be kept well below the speed of sound to avoid a collapse in propeller efficiency and a large increase in noise. Keeping mean piston speed down is a welcome consequence of the same limit rather than the reason for it.

Reading an Engine Designation

The displacement calculation explains the naming convention used by the major American manufacturers of opposed engines: the number in the designation is the displacement in cubic inches, rounded. The letters in front of it describe the configuration and the systems fitted, and they are worth knowing because they tell an engineer a great deal before the cowling comes off:

- **O** — horizontally opposed
- **I** — fuel injected
- **T** — turbocharged
- **S** — supercharged
- **G** — geared, that is, fitted with a propeller reduction gearbox
- **A** — approved for aerobatic flight, with an oil system that continues to function inverted

The letters are read as a group, so a designation such as TSIO tells you at once that the engine is turbo-supercharged, fuel injected and horizontally opposed. It also explains why an engine family that has always been fuel injected carries the I in every one of its designations: there is no carburetted variant for the bare O prefix to describe. European manufacturers do not necessarily follow the convention at all — Rotax model numbers, for instance, are simply type numbers and carry no displacement information — so the engine data plate and the type certificate data sheet remain the authority.

> **Converting between the two systems:** 1 cubic inch is 16.387 cm³. So a 361 cu in engine is 361 × 16.387 = 5,916 cm³, or about 5.9 litres, and a 1,211 cm³ engine is 1,211 ÷ 16.387 = 73.9 cu in. Displacement figures in maintenance data may appear in either system, and mixing them up by a factor of sixteen is an easy and very visible mistake.

From Displacement to Power: Mean Effective Pressure

Displacement on its own does not produce power. What produces power is pressure acting on the piston crown over a distance, repeated many times a minute, and displacement is only the "distance times area" part of that product. The quantity that supplies the missing pressure term is **mean effective pressure**.

Cylinder pressure varies enormously through the cycle, from below atmospheric during induction to its peak shortly after ignition. Mean effective pressure is the single constant pressure which, if it acted on the piston

throughout the whole power stroke, would do exactly the same work as the real, varying pressure does over the complete cycle. It is a bookkeeping device, but an extremely useful one, because it turns a complicated pressure history into one number that can be multiplied by the geometry.

- **Indicated mean effective pressure (IMEP)** is derived from the pressure actually developed in the cylinder, and gives indicated power.
- **Brake mean effective pressure (BMEP)** is the equivalent figure worked backwards from the power measured at the crankshaft, so it already has the engine's friction and pumping losses taken out of it.

Power from mean effective pressure — the PLANK relationship:

$$HP = \frac{P \times L \times A \times N \times K}{33{,}000}$$

Where:

- P = mean effective pressure, lb per square inch
- L = length of stroke, in feet
- A = area of the piston crown, in square inches
- N = number of power strokes per minute **per cylinder**, which for a four-stroke engine is the crankshaft speed divided by two
- K = number of cylinders

The constant 33,000 is the definition of one horsepower in foot-pounds per minute. Use brake mean effective pressure and the answer is brake horsepower; use indicated mean effective pressure and the answer is indicated horsepower.

Worked example. Take the four-cylinder engine used above — bore 5.125 in, stroke 4.375 in — running at 2,700 rpm with a brake mean effective pressure of 146 psi.

Stroke in feet:

$$L = \frac{4.375}{12} = 0.3646 \text{ ft}$$

Piston area:

$$A = \frac{\pi}{4} \times 5.125^2 = 20.63 \text{ in}^2$$

Power strokes per minute per cylinder, four-stroke:

$$N = \frac{2700}{2} = 1{,}350$$

$$HP = \frac{146 \times 0.3646 \times 20.63 \times 1350 \times 4}{33{,}000} = 179.7 \text{ hp}$$

Which is about 180 horsepower — a wholly typical rating for an engine of this size and speed.

Reading the relationship as a designer does makes it far more useful than reading it as an arithmetic exercise, because it names every route to more power and shows what each one costs:

- **Increase** $L \times A \times K$ — that is, increase the displacement. Reliable, and the reason large-displacement engines exist, but it costs weight and frontal area directly.

- **Increase N** — run faster. Free in terms of hardware, but mean piston speed, friction power and propeller tip speed all rise, and volumetric efficiency eventually falls away, as the sections on efficiencies explain.
- **Increase P** — get more pressure out of each cycle. This is what supercharging and turbocharging do by cramming in a denser charge, and what a higher compression ratio does by extracting more work from the same charge. It is the only route that raises power without raising either the size or the speed of the engine, and it is limited by detonation and by the mechanical strength of the parts.

Note also what the four-stroke division by two costs. Because N is half the crankshaft speed for a four-stroke and equal to it for a two-stroke, the same displacement, the same mean effective pressure and the same rpm would give a two-stroke twice the power — which is exactly the theoretical doubling discussed in the previous section, and exactly the doubling that the two-stroke's poorer trapping and lower achievable mean effective pressure prevent it from realising.

Torque and Power Are Not the Same Quantity

Mean effective pressure produces a turning moment at the crankshaft, and the turning moment produces power only in combination with rotational speed. The two are related by a single expression that is worth being able to use in both directions:

Torque and horsepower:

$$HP = \frac{T \times N}{5{,}252}$$

Where T is torque in pound-feet and N is speed in rpm. The constant is 33,000 divided by 2π, and it exists only because the units are mixed.

Worked example: the 180 hp engine above, at 2,700 rpm:

$$T = \frac{HP \times 5252}{N} = \frac{180 \times 5252}{2700} = 350 \text{ lb-ft}$$

Torque is set almost entirely by mean effective pressure and displacement, so it is largely a measure of how well the engine fills and burns each cylinderful. Power adds the question of how often it does so.

This is why a compression-ignition engine can produce high torque at low speed and still not produce headline horsepower: it makes its living on P, not on N. It is also why the propeller, not the engine, usually sets the maximum speed of a piston installation.

Compression Ratio

The **compression ratio** is the ratio of the total cylinder volume (when piston is at BDC) to the clearance volume (when piston is at TDC):

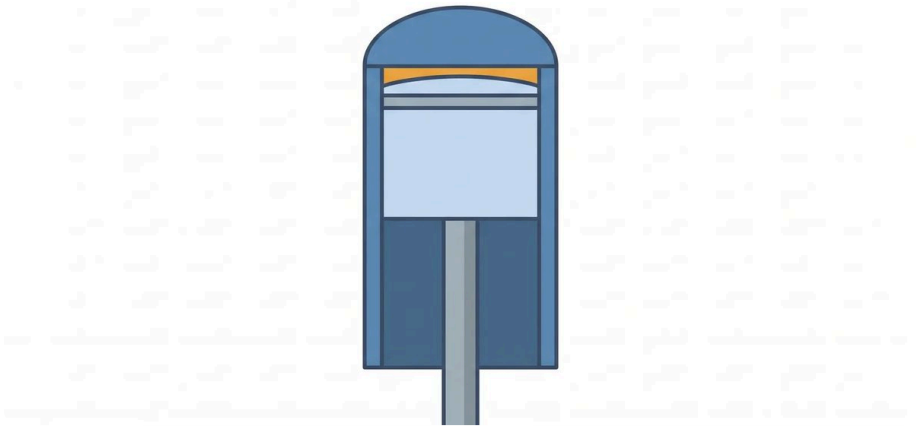
$$CR = \frac{V_{total}}{V_{clearance}} = \frac{V_{swept} + V_{clearance}}{V_{clearance}}$$

Typical values: spark ignition aircraft engines = **7:1 to 10:1**

Higher CR → higher thermal efficiency → more power from same fuel, but increases risk of detonation and requires higher octane fuel.

Clearance Volume and Total Cylinder Volume

Compression ratio is a ratio of two volumes, and the smaller of the two is the one that does the work of defining it. The **clearance volume** is everything that is still enclosed above the piston when the piston is at TDC: the combustion chamber machined into the cylinder head, the space around the valve heads, any dish or valve cut-out in the piston crown, and the small volume between the piston crown and the top compression ring. It is a surprisingly small number, and that is precisely why the compression ratio is so sensitive to it.



With the piston at TDC, only a small enclosed space remains above the crown: the clearance volume.

The relationship between the three volumes is worth stating explicitly, because it is asked in both directions:

Total cylinder volume:

$$V_{total} = V_{swept} + V_{clearance}$$

That is, the volume enclosed with the piston at BDC equals the piston displacement of that cylinder plus the clearance volume. Rearranging the compression ratio expression gives the clearance volume directly from the design figures:

$$V_{clearance} = \frac{V_{swept}}{CR - 1}$$

Worked example — how small the clearance volume really is:

Take the cylinder calculated above, with a swept volume of 90.3 cu in, built to a compression ratio of 8.5:1.

$$V_{clearance} = \frac{90.3}{8.5 - 1} = \frac{90.3}{7.5} = 12.04 \text{ cu in}$$

$$V_{total} = 90.3 + 12.04 = 102.34 \text{ cu in}$$

Checking the result against the definition: 102.34 ÷ 12.04 = 8.50, as required.

So a cylinder that swallows just over 90 cubic inches of mixture squeezes it into a space of about twelve cubic inches — roughly the volume of a small coffee cup. Anything that changes that twelve cubic inches by even a few per cent changes the compression ratio noticeably.

Geometric, Trapped and Effective Compression Ratio

The figure calculated above is the **geometric** compression ratio, derived purely from the dimensions of the parts. Two other ideas share the name, and confusing them is a common source of error.

The **trapped** or dynamic compression ratio recognises that compression does not begin at BDC. As described earlier, the inlet valve stays open well past BDC to capture the momentum of the incoming charge, so the piston has already travelled some way up the bore before the cylinder is sealed. The volume being compressed is therefore less than the full swept volume, and the trapped ratio is lower than the geometric one. Changing the camshaft, or changing valve timing through wear or incorrect assembly, changes the trapped ratio without altering a single dimension of the engine.

The **effective** compression ratio takes account of the pressure the charge starts from. A naturally aspirated engine starts compressing from something below ambient pressure, because the induction system and the throttle both cost pressure. A supercharged or turbocharged engine starts compressing from a manifold pressure well above ambient, so both the pressure and the temperature reached at the end of compression are higher than the geometric ratio alone would suggest. This is the reason — and it runs in the direction that surprises people — that **boosted engines are built with a lower geometric compression ratio than naturally aspirated ones**. The boost has already done part of the compressing, and reducing the geometric ratio is how the designer keeps the end-of-compression temperature below the point at which the fuel will detonate.

The compression-ignition engine escapes this constraint entirely, which is what allows the 15:1 to 22:1 figures in the comparison table earlier in this note. There is no fuel in the cylinder during compression, so there is no end-gas that can auto-ignite prematurely; the limit becomes the mechanical strength of the engine rather than the chemistry of the fuel.

What Changes a Compression Ratio in Service

Compression ratio is a certified design parameter. It appears on the type certificate data sheet, and the fuel grade approved for the engine is chosen to suit it. Several perfectly ordinary maintenance situations nevertheless alter it, and an engineer should be able to see which way each one pushes.

Carbon deposits — clearance volume down, ratio up. Combustion deposits build up on the piston crown and in the combustion chamber, occupying space that used to be clearance volume. Suppose deposits reduce the 12.04 cu in clearance volume of the example above by 5%, to 11.44 cu in:

$$CR = \frac{90.3 + 11.44}{11.44} = \frac{101.74}{11.44} = 8.89$$

A 5% change in the small volume has moved the compression ratio from 8.5:1 to nearly 8.9:1. Worse, the deposits themselves are poor conductors and glow when hot, so the same fault raises the ratio and introduces hot spots at the same time — a combination that leads towards detonation and pre-ignition.

Cylinder base shims — clearance volume up, ratio down. Shims or additional gaskets fitted under a cylinder flange raise the whole barrel relative to the crankshaft, so the piston no longer comes as close to the head at TDC. On a 5.125 in bore the cross-sectional area is

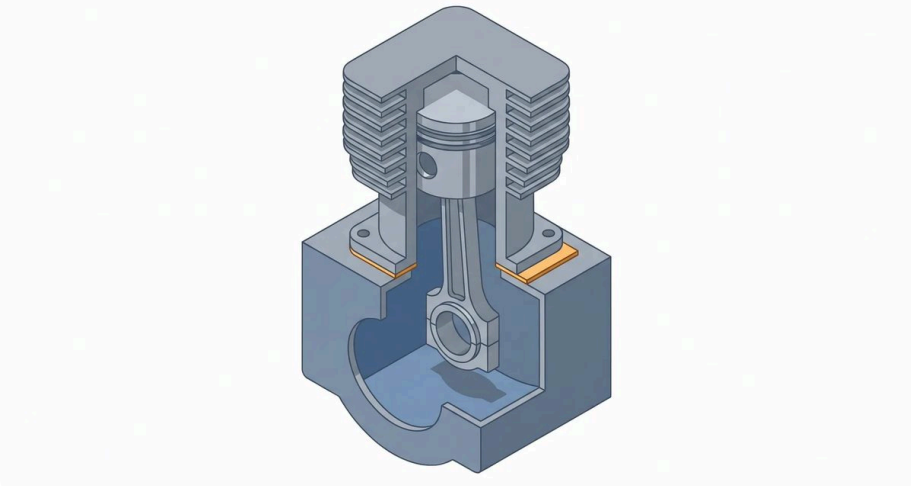
$$A = \frac{\pi}{4} \times 5.125^2 = 20.63 \text{ in}^2$$

so a shim only 0.010 in thick adds 0.010 × 20.63 = 0.21 cu in of clearance volume:

$$CR = \frac{90.3 + 12.25}{12.25} = 8.37$$

Ten thousandths of an inch — about the thickness of three sheets of paper — has taken roughly 0.13 off the compression ratio. It has also changed the piston-to-valve clearance and the deck height, which is why shimming is never a free adjustment and is only done to approved data.

Two other mechanisms are worth recognising. **Valve seat recession**, where the valve gradually sinks deeper into a worn or reworked seat, increases the clearance volume slightly and so lowers the ratio. And fitting a **piston of a different part number** — one with a deeper dish, a dome, or different valve reliefs — changes the clearance volume directly. Pistons are not interchangeable between compression-ratio variants of the same engine even when they look identical and fit perfectly, and a mismatched set can produce an engine that detonates on the fuel it is placarded for.

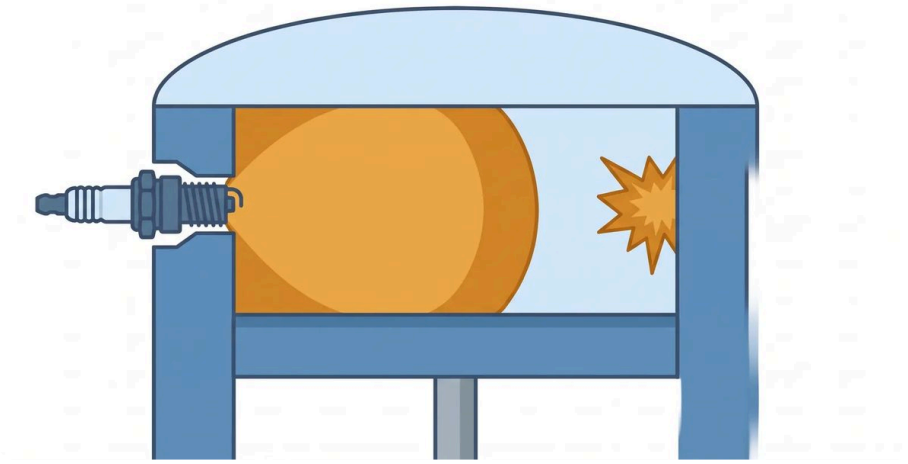


A shim under the cylinder flange lifts the barrel, so the clearance volume grows and the ratio falls.

Compression ratio and fuel grade are a matched pair. The approved fuel grade for an engine is chosen so that the charge does not auto-ignite at the pressure and temperature the compression ratio produces. Raise the ratio, whether by carbon accumulation, by unapproved parts or by removing a shim that was fitted for a reason, and the same fuel that was adequate before is no longer adequate. Fuelling a high-compression engine with a lower grade than it is placarded for has the identical effect from the other direction. Detonation damage — broken ring lands, eroded piston crowns, hammered bearings — can be done in a single high-power take-off.

Detonation and Pre-Ignition: the Two Faults the Compression Ratio Governs

The compression ratio is limited by combustion, not by mechanics, and the two abnormal combustion events that impose that limit are constantly confused with one another. They have different causes, occur at different moments in the cycle, and call for different responses, so the distinction is worth setting out precisely.



In detonation the spark lights the charge normally, but the end gas explodes before the flame front arrives.

	Detonation	Pre-ignition
What happens	The charge is ignited normally by the spark, but the last portion of it — the end gas — reaches its auto-ignition temperature and explodes spontaneously before the advancing flame front gets there	The charge is ignited by a hot surface inside the chamber **before** the spark occurs at all
When in the cycle	After the spark, late in the burn	Before the spark, while the piston is still rising
Typical causes	Fuel grade below that approved for the compression ratio; excessive compression ratio from carbon accumulation or unapproved parts; ignition timing too far advanced; induction air temperature too high; mixture leaned too far at high power; cylinder head temperature already high	A glowing hot spot — a carbon flake, an overheated exhaust valve, a spark plug of the wrong heat range, or a damaged plug electrode

Effect on the engine	Pressure rises almost instantaneously instead of progressively, producing a shock wave that hammers the piston crown, the ring lands, the bearings and the head; cylinder head temperature climbs	Combustion starts far too early, so the rising piston works against a fully developed pressure; power is lost and heating is severe and rapid
Relationship	Prolonged detonation overheats components and can create the hot spot that then causes pre-ignition	Pre-ignition raises pressures and temperatures and can in turn provoke detonation

Each can therefore lead to the other, which is why an engine that has been allowed to detonate under high power can destroy itself remarkably quickly. In an aircraft neither condition is audible from the cockpit over the noise of the propeller and the airframe, so the crew's warning is indirect: rising cylinder head temperature, a loss of power that does not match the power setting, and roughness. The engineer's warning is the physical evidence found afterwards — broken ring lands, an eroded or holed piston crown, hammered bearing shells, and cylinder head damage around the hottest part of the chamber.

Both faults connect back to the compression ratio in the same way. A higher ratio produces a higher pressure and temperature at the end of compression, which brings the end gas nearer to its auto-ignition point before the spark has even occurred, and leaves less margin for anything else that adds heat. Everything that makes detonation more likely — hot induction air, a hot cylinder, advanced timing, a lean mixture at high power, or a low-grade fuel — is simply another contribution on top of what the compression ratio has already used up. This is why the approved fuel grade, the maximum permitted cylinder head temperature, the mixture schedule and the ignition timing are not four independent limits but four aspects of one limit.

Why a rich mixture protects an engine at high power. Fuel that is supplied beyond the amount the available oxygen can burn does not release heat; it absorbs it, evaporating in the cylinder and leaving with the exhaust. Enrichment at take-off power is therefore a deliberate cooling measure for the piston crown and the exhaust valve, and a deliberate detonation margin, paid for in specific fuel consumption. It is also why a partially blocked jet, a mis-set mixture linkage or an induction leak that leans a cylinder is a far more serious defect at high power than at cruise.

Compression Ratio Is Not a Compression Check

The two terms sound alike and are frequently confused, but they measure completely different things.

Compression ratio is geometry. It is fixed when the engine is built, it is stated in the engine's approved data, and no test carried out on a complete engine measures it. The only way to determine it is from the dimensions of the parts.

A **compression check** measures how well the cylinder seals. In the differential form used throughout aviation maintenance, the piston is brought to TDC on its compression stroke with both valves closed, regulated air is applied through the spark plug hole, and the tester compares the pressure held in the cylinder with the pressure supplied to it, conventionally 80 psi. The reading is a leakage measurement, quoted as the cylinder pressure over the supply pressure, and where the escaping air can be heard identifies the leak path: past the

piston rings if it can be heard at the crankcase breather, past the exhaust valve if it can be heard at the tailpipe, past the inlet valve if it can be heard at the induction inlet or carburettor.

A cylinder with a perfectly correct compression ratio can therefore fail a compression check completely, and a worn engine whose rings are leaking still has exactly the compression ratio it was built with. Keep the two ideas apart.

Efficiencies

Mechanical Efficiency

The ratio of **brake horsepower** (BHP — power available at the crankshaft) to **indicated horsepower** (IHP — power developed inside the cylinders):

$$\eta_{mech} = \frac{BHP}{IHP} \times 100\%$$

Typically **85–95%**. The difference is lost to friction in bearings, piston rings, valve gear, and driving accessories.

Where the Mechanical Losses Go

The gap between the power developed on top of the pistons and the power delivered at the propeller flange is real work, and it is worth knowing what consumes it, because almost every item in the list is something maintenance can make better or worse.

- **Rubbing friction.** The piston rings and skirt sliding on the cylinder wall are the single largest contributor, followed by the main and big-end bearings and the camshaft and valve gear. All of it depends on rubbing speed, on the load pressing the surfaces together, and on the oil film separating them.
- **Pumping work.** The engine must drag the charge in past the air filter, the venturi and a partly closed throttle butterfly, and then push the burnt gas out through the exhaust system. At part throttle this is a substantial loss, and it is one an engine running at wide-open throttle and low rpm largely avoids — which is part of why cruise is flown at a high manifold pressure and a low engine speed rather than the other way round.
- **Accessory drive.** Magnetos, alternator, fuel pump, oil pump, vacuum pump, propeller governor and, on some installations, a hydraulic pump are all turned by the engine. None of them appears in the indicated power and all of them come out of the brake power.

Worked example: an engine develops 200 indicated horsepower and loses 20 horsepower to friction, pumping and accessories:

$$BHP = IHP - FHP = 200 - 20 = 180 \text{ hp}$$

$$\eta_{mech} = \frac{180}{200} \times 100\% = 90\%$$

The 20 horsepower that vanished did not disappear; it became heat, and most of it went into the oil. That is why oil temperature is a genuine indicator of how hard an engine is working and of whether something is dragging that should not be.

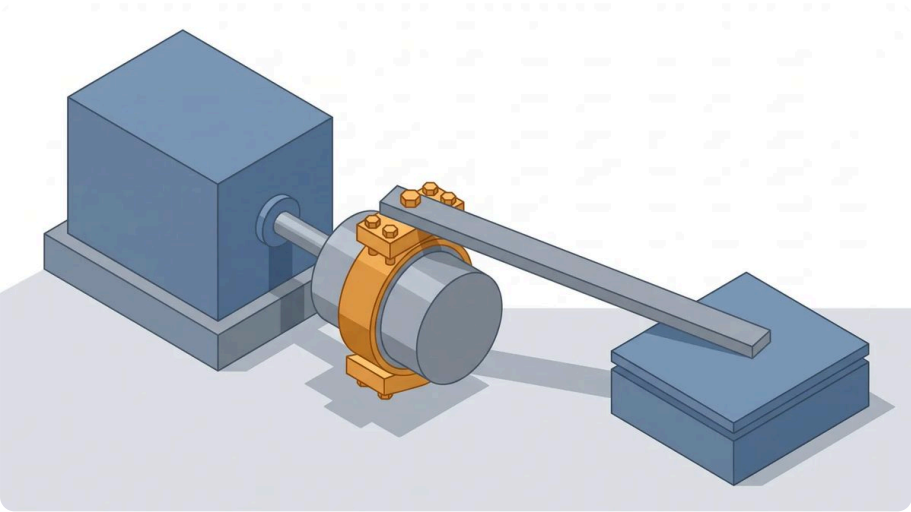
Mechanical efficiency is not a fixed property of an engine. Friction power rises steeply with crankshaft speed — rubbing velocities rise, and the viscous shear in every oil film rises with them — so friction horsepower increases as rpm increases, and it rises faster than in direct proportion. Indicated power, meanwhile, rises with speed only until the engine's breathing starts to fall away. The result is that mechanical efficiency is best somewhere in the middle of the speed range and falls off at both ends: at idle the friction is small in absolute terms but enormous as a proportion of the tiny indicated power being produced, and near the maximum permitted speed friction has grown to consume a serious fraction of a large indicated power.

That rising friction power is only half of what limits an engine to a maximum continuous speed. The other half is what happens to its breathing at the same speeds, which the volumetric efficiency material further down this section deals with.

Oil grade is a mechanical-efficiency decision. An oil that is too heavy for the ambient temperature has a thicker film and shears harder, so friction power goes up and brake power comes down; on a cold morning the effect is large enough to be felt in the acceleration of the engine and seen in the oil pressure. An oil that is too thin at operating temperature fails to keep the surfaces apart and the loss becomes wear rather than drag. Grade selection is stated in the engine manual against ambient temperature ranges for exactly this reason, and it is a maintenance input to an efficiency figure.

Where the Word "Brake" Comes From, and How Power Is Measured

Indicated power can be calculated from a pressure record taken inside the cylinder. Brake power cannot be calculated at all; it has to be measured, by applying a known resisting load to the output shaft and seeing how much the engine can overcome. The oldest device for doing so was the Prony brake, a friction band clamped around the shaft whose restraining arm rested on a weighing scale — and it is from that brake that **brake** horsepower takes its name. Modern test beds use hydraulic or electrical dynamometers instead, but the definition is unchanged: brake power is the power that survives to the output shaft and is available to do external work.



A Prony brake: a friction band clamped round the output shaft, its arm resting on a weighing scale.

Because power depends on the state of the air the engine is breathing, a measured figure means nothing without the conditions it was measured at. A power rating is therefore always quoted **at a stated crankshaft speed and referred to standard atmospheric conditions**, and an observed figure from a test run in real ambient conditions has to be corrected before it can be compared with the rating. The correction is essentially a density correction, and it runs in the direction the induction argument predicts: a hot day or a high field elevation means less dense air, less mass of charge, less fuel that can usefully be burnt, and less power — the engine is not faulty, it is being asked to breathe thinner air.

The maintenance organisation almost never has a dynamometer, so the aircraft's own propeller is used as the load instead. A **fixed-pitch propeller** is a well-behaved absorber: for a given blade setting it absorbs a torque that rises steeply with rotational speed, so there is exactly one speed at which the propeller's demand matches what the engine can supply at full throttle. That is why the static rpm reached at full throttle on the ground is a genuine measurement rather than a formality:

- **Static rpm below the range in the maintenance data** suggests the engine is not making its power — low compression, ignition faults, incorrect timing, a restricted induction or an over-rich or over-lean mixture — or that the propeller is not the one the figures were established for.
- **Static rpm above the range** is equally significant and is often overlooked. It usually means the propeller is not absorbing the power it should, and it raises the possibility of the engine being overspeeded in flight.
- The figures are only comparable if the ambient conditions are accounted for, which is why the maintenance data states them for standard conditions or provides a correction.

On a **constant-speed** installation the picture is different, because the governor holds the rpm at the selected value and varies blade angle to do it. Engine speed then tells you nothing about power, and manifold pressure and fuel flow become the indications that do. This is the practical reason the two kinds of installation are checked differently on the ground, and it follows directly from what a propeller is doing as a load rather than from any difference in the engine.

Power, torque and the instruments that show them. No light aircraft carries an instrument that reads power directly. What the panel offers is a set of proxies: rpm, which fixes the frequency of the power strokes; manifold pressure, which is the best available indication of how much charge each stroke is getting; and fuel flow, which is proportional to the rate at which energy is being supplied. Reading a power setting from a chart is the process of converting those proxies back into the power figure the engine was rated at, which is why the chart is specific to the engine, the propeller and the installation.

Thermal Efficiency

The ratio of useful work output to the total heat energy in the fuel consumed:

$$\eta_{thermal} = \frac{\text{Work output (BHP)}}{\text{Heat energy in fuel consumed}} \times 100\%$$

Typically **25–30%** for spark-ignition aircraft engines. The remaining energy is lost as exhaust heat, cooling, and radiation.

Reading Thermal Efficiency Off the Fuel Flow

Thermal efficiency sounds like a laboratory quantity, but it can be calculated from two numbers that are on the instrument panel of any aeroplane: the power being produced and the fuel being burned to produce it. The bridge between them is **brake specific fuel consumption**, the mass of fuel consumed per unit of brake power per hour.

Thermal efficiency from specific fuel consumption:

$$\eta_{thermal} = \frac{2544}{BSFC \times LHV}$$

Where BSFC is in pounds of fuel per brake horsepower per hour and LHV is the lower heating value of the fuel in BTU per pound. The constant 2,544 is the number of BTU equivalent to one horsepower-hour; the lower heating value of aviation gasoline is about 18,700 BTU per pound.

Worked example: an engine delivering 180 BHP burns 15 US gallons of AVGAS per hour. Aviation gasoline weighs about 6.0 lb per US gallon, so the fuel flow is

$$15 \times 6.0 = 90 \text{ lb/hr}$$

$$BSFC = \frac{90}{180} = 0.50 \text{ lb/hp/hr}$$

$$\eta_{thermal} = \frac{2544}{0.50 \times 18700} = \frac{2544}{9350} = 0.272 = 27.2\%$$

Which lands squarely inside the 25 to 30% band quoted above, and was obtained from nothing more exotic than a fuel flow gauge and a power setting. Cross-checking the same figures the other way round: the fuel carries 90 × 18,700 = 1,683,000 BTU per hour into the engine, and 180 hp × 2,544 = 457,900 BTU per hour comes back out at the crankshaft.

Because efficiency and specific fuel consumption are inversely related, anything that lowers BSFC raises thermal efficiency. Leaning the mixture towards the best-economy setting, which is at peak exhaust gas temperature, gives the lowest specific fuel consumption the engine is capable of; running deliberately rich for cooling at high power gives the highest. That enrichment is a knowing trade of specific fuel consumption for component temperature and detonation margin, and it is measurable: the same engine will show a distinctly worse thermal efficiency during a full-power climb than during a properly leaned cruise, without anything at all being wrong with it.

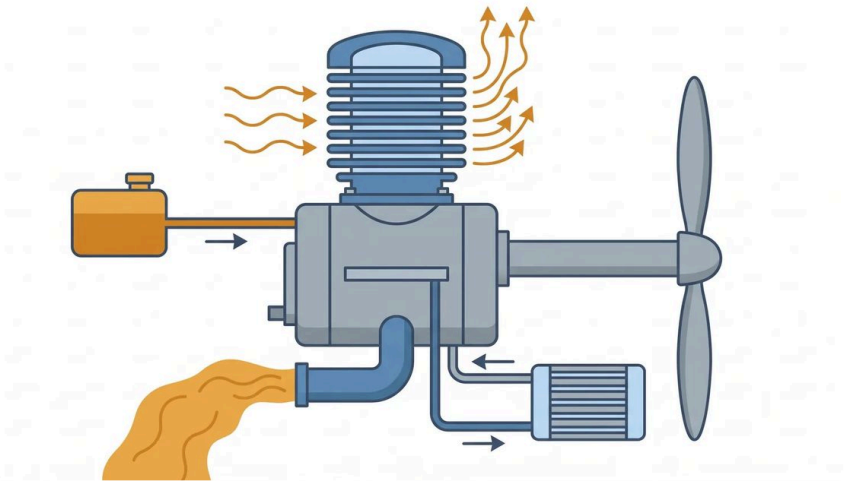
The Air-Standard Limit and the Size of the Gap

The ideal-cycle calculation earlier in this note put the theoretical ceiling for a spark-ignition engine at around 57% for a compression ratio of 8.5:1. The measured figure is a little under half of that. It is worth being able to account for the difference item by item rather than dismissing it as "losses", because each item is a real physical process and several of them are influenced by how the engine is maintained and operated:

- **Heat straight into the metal.** The gas is at its hottest at exactly the moment it is in contact with the smallest, thinnest part of the chamber. A large fraction of the heat released never gets the chance to do work at all.
- **Finite burning time.** The ideal cycle releases all its heat instantaneously at TDC. A real flame front takes tens of crankshaft degrees to cross the chamber, so some of the heat is released when the piston has already moved down and the leverage is gone.
- **Blowdown.** The exhaust valve opens before BDC, so the last part of the expansion is thrown away deliberately.
- **Pumping and friction.** Because the practical figure is referred to brake power, everything counted as a mechanical loss above is also inside this gap.
- **Chemistry.** At peak temperature some combustion products dissociate, absorbing energy that is only partly recovered as the gas cools; and combustion is never quite complete, so some fuel leaves as unburnt hydrocarbon or as carbon monoxide.

The Engine's Heat Balance

Put the same idea in the form an engineer uses on the hangar floor and it becomes a heat balance: of all the energy that arrives in the fuel tank, this is roughly where it goes. The figures below are approximate, vary with power setting and mixture, and are quoted differently by different sources, but the shape of the picture is consistent and examinable.



Fuel energy leaves as crankshaft work, exhaust heat, heat to the cooling air and heat to the oil.

Destination	Approximate share of fuel energy	What it means in practice
Useful work at the crankshaft	25 – 30%	The only part that turns the propeller
Exhaust gases	40 – 45%	The single largest loss; the reason exhaust gas temperature is such a sensitive indicator, and the energy a turbocharger recovers a small part of
Cooling system	20 – 25%	Heat rejected through the cylinder head and barrel fins and carried away by the oil; the job the cowling, baffles and cooling air must actually do
Friction, pumping and radiation	Around 5 – 10%	Appears as heat in the oil and in the engine structure

Two conclusions follow, and both are tested. First, the majority of the wasted energy leaves the engine through the exhaust and the cooling system, not through friction — friction feels significant because it is the part an engineer can touch, but thermodynamically it is the small one. Second, the cooling system is not a minor accessory: it is being asked to move something like a fifth to a quarter of the entire fuel energy of the engine. That share is the bulk of what the engine structure has to reject, and it divides in two. The heat radiated into the cooling airflow from the cylinder head and barrel fins is the larger part, around 15 to 20% of the fuel energy on its own; the rest is heat that has soaked into the pistons, cylinder walls and bearings and leaves the engine through the oil and the oil cooler. A source that quotes only the fin share therefore gives a smaller number for the same engine.

What that means in numbers: for the 180 BHP engine above, burning 90 lb/hr, the fuel supplies 1,683,000 BTU per hour. At 20 to 25%, the cooling system is disposing of roughly 337,000 to 421,000 BTU every hour — a heat load comparable with the crankshaft output itself. That is why a displaced or perished baffle seal, a missing cowl fastener or a blocked cylinder fin is not a cosmetic defect but a direct cause of high cylinder head temperature.

Volumetric Efficiency

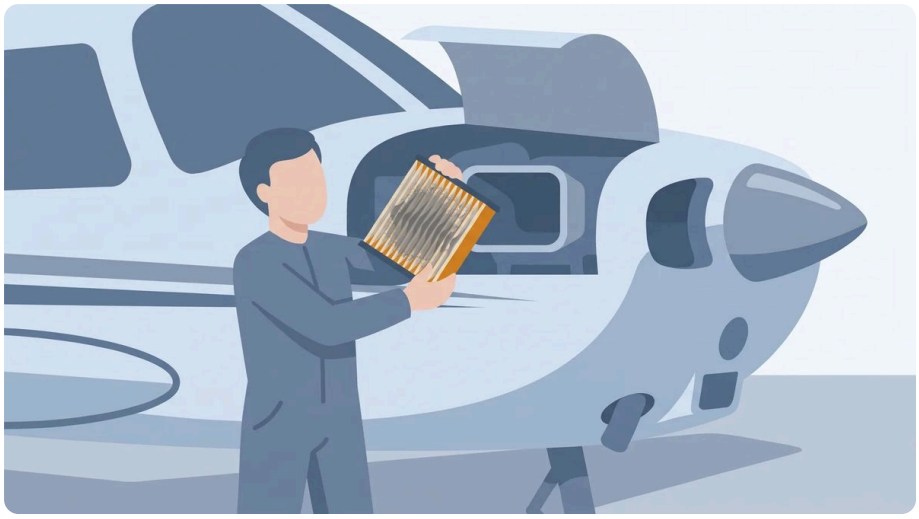
The ratio of the **actual volume of mixture drawn in** to the **theoretical swept volume** at ambient conditions:

$$\eta_{vol} = \frac{V_{actual}}{V_{swept}} \times 100\%$$

Typically **75–85%** for normally aspirated engines. Reduced by intake restrictions, valve timing, heat, and altitude. Can exceed 100% with supercharging.

What Actually Limits Volumetric Efficiency

Volumetric efficiency is a measure of how well the engine breathes, and every restriction between the free airstream and the closed inlet valve subtracts from it. It is useful to walk the air path in order, because that is the order in which an engineer troubleshoots it:



Walking the air path starts at the air filter, the cheapest and most neglected restriction to breathing.

- **Air filter.** A clean filter costs a little pressure; a contaminated one costs a lot. This is the cheapest and most frequently neglected influence on the whole list.
- **Alternate air or carburettor heat path.** When selected, air is drawn from a warmer, usually less direct source, so it arrives both hotter and at a slightly lower pressure.
- **Carburettor venturi.** A venturi works by creating a pressure drop, so the very mechanism that meters the fuel also costs breathing. This is one reason an injected engine breathes marginally better than a carburetted one of the same size.
- **Throttle butterfly.** At part throttle this is by far the largest restriction in the system, and it is deliberate — throttling is how a spark-ignition engine's power is controlled.
- **Induction pipe length, diameter and bends.** Long, narrow or sharply bent runners cost pressure and, on a multi-cylinder engine, distribute the charge unevenly between cylinders.
- **Inlet port, valve size and lift.** The final and often the tightest restriction. It is why a large bore, which allows a large valve, helps breathing.

- **Residual exhaust gas.** Whatever burnt gas is left in the clearance volume at the end of the exhaust stroke occupies space that fresh charge cannot. Anything that raises exhaust back pressure — a collapsed muffler baffle, an obstructed tailpipe, a distorted exhaust gasket — leaves more of it behind.

Charge temperature deserves separate treatment because it is the effect that is easiest to demonstrate. A given volume of hot air contains less mass than the same volume of cold air. Anything that heats the charge on its way in — hot induction pipes, a hot cylinder head and inlet port, or the deliberate selection of carburettor heat — reduces the mass of oxygen that ends up in the cylinder, and with it the amount of fuel that can be burned and the power produced. Selecting carburettor heat therefore always costs power, and the loss is visible in the cockpit as a drop in rpm on a fixed-pitch installation.

Altitude works on volumetric efficiency in the same direction, though the larger part of the power loss with height is simply the lower density of the air itself rather than a change in the ratio. The ratio does fall as well, if less dramatically: the same heat picked up from hot engine surfaces is delivered into a smaller mass of air at each stroke, so it produces a larger proportional fall in the density of the charge.

Working in the other direction, volumetric efficiency can be pushed **above** 100%. Supercharging and turbocharging do it by force, delivering the charge to the inlet valve at above ambient pressure. A naturally aspirated engine can also exceed 100% over a narrow band of speeds, if the induction and exhaust systems are tuned so that the returning pressure waves happen to arrive at helpful moments — the same effect that a two-stroke expansion chamber exploits. Both are genuine, and both mean that the definition of volumetric efficiency has to state what conditions the theoretical volume is measured at, or the number is meaningless.

Volumetric Efficiency and Engine Speed

Worked example — how much air an engine actually needs. Take a 361 cu in four-stroke engine at 2,400 rpm. Each cylinder fills once every two revolutions, so the whole engine draws its displacement once every two revolutions:

$$\dot{V}_{theoretical} = 361 \times \frac{2400}{2} = 433{,}200 \text{ cu in/min}$$

$$= \frac{433{,}200}{1728} = 251 \text{ cu ft/min}$$

At a volumetric efficiency of 80% the engine actually swallows

$$251 \times 0.80 = 201 \text{ cu ft/min}$$

The missing 50 cubic feet a minute is the direct measure of everything the induction system is costing.

Engine speed enters this twice, and in opposite senses. Raising the speed increases the number of induction strokes per minute, which is what raises the airflow and therefore the power. But raising it also gives the cylinder less time to fill on each stroke, while the pressure drop needed to move gas through a restriction rises roughly with the square of the flow velocity. Beyond the speed at which the induction system is tuned, the second effect wins and volumetric efficiency falls away.

So running an engine at excessively high rpm is doubly unrewarding: the friction power is climbing steeply at the same time as the breathing is deteriorating. This is why the power curve of a piston engine flattens and

then falls at the top of the speed range rather than continuing to rise, and why maximum power occurs at a speed the engine can actually fill its cylinders at.

What Does Not Affect Volumetric Efficiency

It is as important to know what is outside a definition as what is inside it, and volumetric efficiency attracts more confusion than the other two.

Oil viscosity does not affect volumetric efficiency. At a given engine speed the mass of charge a cylinder swallows depends on the pressure, temperature and restriction of the air path, none of which the oil is part of. Heavy oil increases the work absorbed by the moving parts and therefore reduces mechanical efficiency, but it does not make the cylinder breathe any less well. Sorting a stated influence into the correct one of the three efficiencies is a standard examination task, and oil viscosity is the classic case of it — it belongs to mechanical efficiency, not to volumetric.

Exhaust noise is a similar trap. A loud exhaust is a **symptom** — usually of a leak, a failed internal baffle or a missing component — and the underlying defect may well be raising back pressure and hurting breathing. The sound level itself changes nothing.

Which Efficiency Does the Engineer Actually Control?

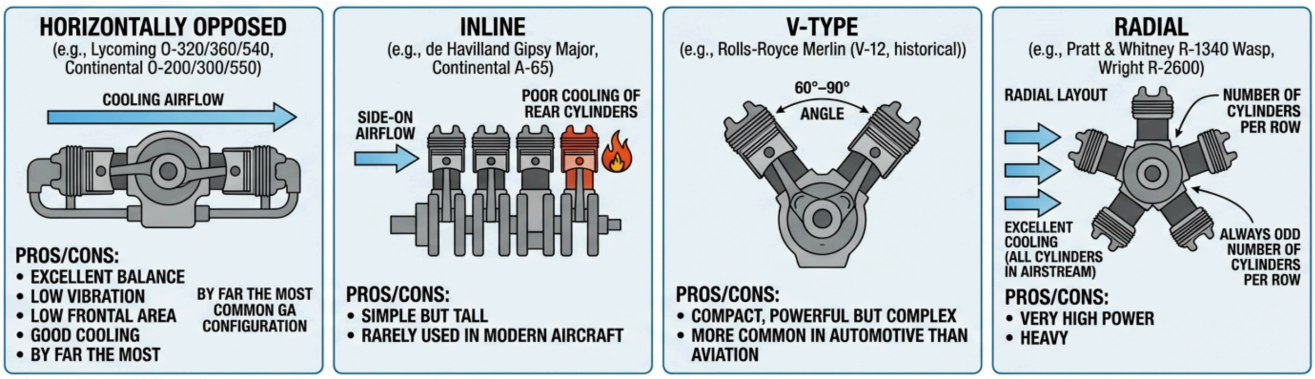
The three efficiencies answer three different questions about the same engine, and it is worth being precise about how they relate, because they are not simply multiplied together. **Volumetric efficiency** asks how much charge the cylinder actually captures, and therefore sets how much power the engine is able to make at all. **Thermal efficiency** asks what fraction of the energy in the fuel reaches the crankshaft as useful work — and because it is referred to brake power, the friction and pumping losses are already inside that figure. **Mechanical efficiency** then isolates that one loss, separating the work done on top of the pistons from the work that survives to the propeller flange. Beyond the flange the propeller performs a conversion of its own, typically achieving of the order of 80% at its design condition, so the efficiency of the complete chain from fuel energy to thrust power is lower again than any single figure quoted in this section.

What makes the distinction practical rather than academic is that each efficiency is influenced by a different set of maintenance actions:

Efficiency	The maintenance actions that decide how close the engine gets to it
Volumetric	Filter servicing to the stated interval; induction system leak and security checks; valve clearances and valve timing set to the manual; exhaust system inspected for internal collapse and obstruction; carburettor heat used only when it is needed
Thermal	Magneto internal timing and engine timing checked and set; plugs cleaned, gapped, tested and rotated; harness insulation resistance checked; combustion deposits controlled by correct operation; only approved pistons, cylinders and fuel grades used, so that the certified compression ratio is the one the engine actually has
Mechanical	Oil grade chosen for the ambient temperature range; oil and filter changed at the stated interval and the filter examined for debris; accessories checked to turn freely; bearing and ring clearances kept within overhaul limits; nothing added to the accessory drive that the engine was not certificated with

None of these actions changes the definition of an efficiency. What they change is how close the engine in front of you gets to the figure the manufacturer published for it, which is the only version of the number that matters in service.

Engine Configuration and Firing Order



Configuration	Characteristics	Examples
Horizontally Opposed	Cylinders in two banks, opposite each other. Excellent balance, low vibration, low frontal area, good cooling. By far the most common GA configuration.	Lycoming O-320/360/540, Continental O-200, O-300 and IO-550
Inline	All cylinders in a single row. Simple but tall, poor cooling of rear cylinders. Rarely used in modern aircraft.	de Havilland Gipsy Major, Ranger L-440
V-type	Two banks of cylinders at 60°–90° angle. Compact, powerful, but complex. More common in automotive than aviation.	Rolls-Royce Merlin (V-12, historical)
Radial	Cylinders arranged in a star pattern around the crankshaft. Excellent cooling (all cylinders in airstream), very high power, heavy. Always odd number of cylinders per row.	Pratt & Whitney R-1340 Wasp, Wright R-2600

The Horizontally Opposed, or Flat, Engine

The horizontally opposed layout is also called a **flat** engine, and the two names describe exactly the same machine — "flat four" and "horizontally opposed four-cylinder" are interchangeable. The cylinders lie on opposite sides of the crankshaft in the same horizontal plane, so **the angle between the two banks is 180 degrees**. That single geometric fact is where almost all of the layout's advantages come from.

- **Balance.** Cylinders in opposing pairs move outward and inward together, so the reciprocating inertia force of one is cancelled almost exactly by its opposite number. What remains is a small rocking couple, because, with each cylinder on its own crank throw, as these engines have, the two cylinders of a pair cannot occupy the same station along the crankshaft and are therefore slightly offset. This is why a flat engine runs smoothly with relatively little balancing mass on the crankshaft.
- **Frontal area.** The engine is wide but very shallow, so it fits inside a slim, low-drag cowling and leaves the pilot a view over the nose.

- **Crankshaft length.** Every cylinder has its own crank throw, so a six-cylinder flat engine has six of them, the same number a six-cylinder inline needs. The difference is how they are spaced along the shaft: the two throws serving a facing pair lie immediately next to one another and are set 180 degrees apart, so the six are grouped into three closely spaced pairs rather than strung out one behind another down the whole length of the crankshaft. The same six cylinders are therefore carried on a much shorter shaft. A short, stiff crankshaft resists torsional vibration far better than a long one.
- **Cooling.** Every cylinder projects into the airstream on its own, rather than sitting in the wake of the one in front, so with correctly fitted baffles each one can be given its own supply of cooling air.

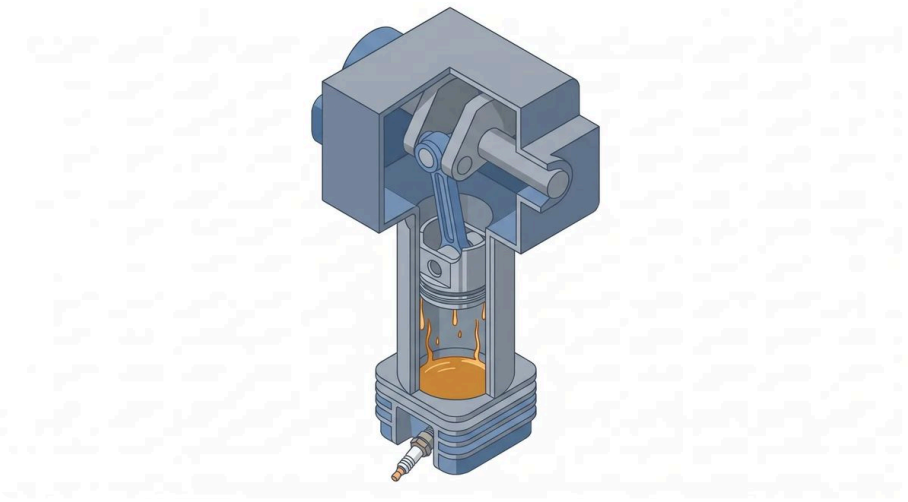
Almost every certified light aircraft engine in current production is a flat four or a flat six; flat eights exist but are far less common. Construction is characteristic: the crankcase is split vertically on the centreline into two halves that clamp the main bearing shells between them, and long through-bolts pass right across the engine, retaining the main bearings and the cylinder base flanges at the same time. This is why cylinder removal on such an engine is not a self-contained job — slackening the wrong fasteners disturbs the main bearing clamp-up, and the torque sequence in the manual exists to protect it.

Inline and Inverted Inline

A single row of cylinders gives the simplest possible crankcase and leaves one whole side of every cylinder accessible, which is why the layout persisted so long in training and touring aircraft. Its weaknesses are the ones the table above names. The engine is tall, and cooling of the rear cylinders is poor because the air reaching them has already passed over — and been heated by — every cylinder ahead of them, so temperatures rise progressively down the row. A long crankshaft, with all its torsional consequences, is a further penalty of putting every cylinder in one line.

Both of the examples given in the table for the inline configuration are **inverted** designs, with the cylinders hanging below the crankshaft. Inverting the engine raises the propeller shaft relative to the bottom of the cowling, which gives better propeller ground clearance and a better forward view, and it puts the accessories and the crankcase where they are easy to reach from above.

Inverted engines and hydraulic locking. When an inverted engine is shut down, gravity works against it: oil that has drained down the cylinder walls, and any fuel that has seeped past a valve, collects in the combustion chambers and the cylinder heads, which are now the lowest point. Liquid does not compress. If enough accumulates and the engine is then turned under starter power, the piston can be stopped dead against it and a connecting rod bent or a cylinder head cracked. The same hazard exists on the lower cylinders of a radial engine and for the same reason. The engine maintenance manual states the procedure — typically involving removal of the lower spark plugs to drain the affected cylinders — and it must be followed before the engine is turned under power after any period of standing.



In an inverted cylinder, oil and fuel drain into the head, the lowest point, where the piston can hit it.

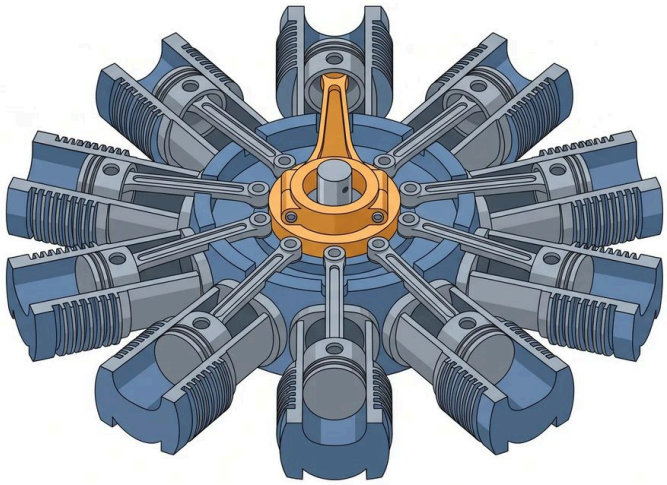
The V-Type

The V layout folds two banks of cylinders onto a single crankshaft. Compared with an inline of the same cylinder count, the engine is roughly half as long, its crankshaft is far shorter and stiffer, and its frontal area is only moderately greater. That combination is what made it the layout of choice for high-power military engines in the piston era, where the cylinder count needed to be high and the crankshaft still had to survive.

The bank angle matters mechanically as well as dimensionally: it determines how the firing impulses of the two banks interleave and how well the primary forces balance. The 60 to 90 degree range noted in the table is not arbitrary. The great V-12 aero engines were also liquid cooled, which is what allowed the banks to be brought so close together — there was no requirement to force cooling air between and around every cylinder. The price was a radiator, its ducting, its drag, its coolant and the vulnerability of the whole system to a single leak.

The Radial

A radial engine arranges its cylinders like the spokes of a wheel around a single crank throw, with all of them equally spaced and all of them directly exposed to the airstream. Its internal arrangement is unlike any other layout: **one master connecting rod** runs from its piston to the crankpin, and every other piston drives an **articulated or link rod** that is pinned to a flange on the big end of that master rod. A consequence worth remembering is that the link-rod pistons do not describe quite the same motion as the master-rod piston, so their effective strokes and compression ratios differ very slightly from cylinder to cylinder.



One master rod runs to the single crankpin; every other piston's link rod pins to its big-end flange.

The requirement for an **odd number of cylinders in each row** follows directly from the four-stroke cycle. For even firing, the engine must fire alternate cylinders around the circle. Starting at cylinder 1 in a nine-cylinder row and taking every second cylinder gives 1, 3, 5, 7, 9, then 2, 4, 6, 8, and only then back to 1 — every cylinder visited exactly once in the 720 degrees of the cycle. Try the same thing with an even number of cylinders and taking every second one returns to the starting cylinder after covering only half of them, leaving the other half unfired. An odd count is therefore not a tradition but an arithmetic necessity.

Where more power is needed than one row can provide, rows are stacked: a twin-row radial of fourteen cylinders is two rows of seven, and one of eighteen cylinders is two rows of nine, with the rear row staggered so that its cylinders sit in the gaps between those of the front row. The two engines named in the table illustrate both arrangements — a nine-cylinder single-row engine and a fourteen-cylinder two-row engine. Multi-row engines cool the rear row far less easily than the front, which is why cooling gills, baffles and cylinder head temperature monitoring became so important on large radials.

The layout's virtues are cooling, a very short and rigid crankshaft, and an excellent power-to-weight ratio for its era. Its vices are the ones that ended its dominance in front-line use: a large frontal area and therefore high drag, considerable weight, and the oil consumption and hydraulic-locking behaviour that come from having cylinders below the crankshaft.

Rotary: Two Different Engines Sharing One Word

"Rotary" is used for two entirely different machines, and the examination will not tell you which one it means.

The **historical rotary** of the First World War — the Gnome and Le Rhone engines and their contemporaries — looks like a radial and is not one. Its crankshaft is bolted to the airframe and does not turn; instead the crankcase, the cylinders and the propeller all rotate around it as a single mass. That gave excellent cooling and a very large flywheel effect, but also enormous gyroscopic forces in a turn, total-loss lubrication with castor oil, and no practical means of throttling. The layout was obsolete by the 1920s.

The **Wankel rotary** is a modern engine with no pistons at all. A roughly triangular rotor with curved faces turns eccentrically inside a housing whose bore is an epitrochoid — a figure-of-eight-like shape rather than a circle. Each of the rotor's three faces forms a chamber with the housing wall, and as the rotor turns, each of those

chambers is carried past the intake port, through compression, past the sparking plugs, through expansion and past the exhaust port. Every face therefore completes a full four-phase cycle in one revolution of the rotor, so the engine produces three power events per rotor revolution. The output shaft is geared to the rotor so that it turns three times for each single revolution of the rotor, which gives one power event per output shaft revolution from each rotor.

Wankel rotary — advantages	Wankel rotary — disadvantages
Very high power-to-weight ratio for its displacement	Higher specific fuel consumption than an equivalent piston engine
Far fewer moving parts — no valves, camshaft, pushrods, connecting rods or reciprocating pistons	Apex seals at the rotor tips run against the housing at high speed and are the principal wear and reliability concern
Exceptionally smooth, because nothing reciprocates and the rotating masses are easily balanced	The long, thin combustion chamber has a poor surface-to-volume ratio, so it loses heat readily and burns incompletely
Compact for the power delivered, with ports instead of a valve train	Higher exhaust emissions, particularly unburnt hydrocarbon, and weaker torque at low speed

Wankel engines have found a place in unmanned aircraft and in a small number of light aircraft installations, where compactness and smoothness are worth more than fuel burn. They are covered further in the study note dealing with alternative engine constructions.

Balance, Vibration and the Number of Cylinders

Configuration and firing order exist in the same section of the syllabus because they solve the same problem from two directions. The configuration decides what forces the moving parts generate and how they can be made to cancel; the firing order decides how the combustion impulses are distributed in time. Vibration is what is left over when neither has been arranged well.

Three distinct sources of vibration act on a piston engine, and they behave differently:

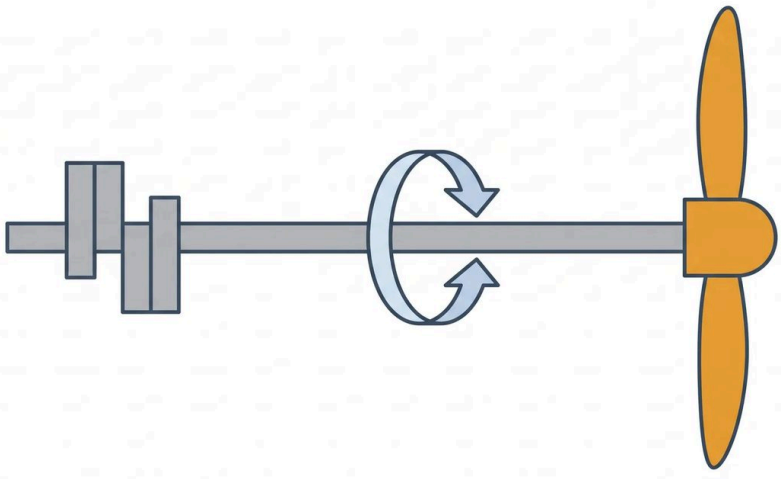
- **Rotating out-of-balance.** The crank throws, the big ends and part of the connecting rod mass all whirl about the crankshaft axis. This is the easiest source to deal with because it can be balanced completely by fitting counterweights of the right mass in the right angular positions.
- **Reciprocating inertia.** The piston, gudgeon pin and the upper part of the connecting rod stop and reverse twice per revolution, and the force required to do that acts along the cylinder axis. It cannot be cancelled by a rotating counterweight alone; it has to be cancelled by another piston moving the opposite way, which is precisely what the opposed configuration provides.
- **Torque reaction and firing impulses.** Each power stroke delivers a sudden twist to the crankshaft, and the crankshaft delivers an equal and opposite reaction to the crankcase and thus to the engine mounts. This is a torsional excitation, and it is why engines are carried on flexible mounts rather than bolted rigidly to the airframe.

The cylinder count acts on the second and third of these at once. It acts on the third because it sets how closely one firing impulse follows the next, and the closer they come the less the crankshaft speed can fall away between them. It acts on the second because more cylinders give the designer more freedom to oppose

one residual couple with another: a flat six can be laid out so that its three opposing pairs cancel each other and is inherently balanced, whereas a flat four is left with a rocking couple it cannot cancel from within itself. That combination of smoother delivery and better balance, rather than extra power alone, is a large part of why the flat six became the standard at the higher end of the light aircraft market.

Torsional vibration deserves separate attention because it is the source that can destroy a crankshaft. The crankshaft, the propeller and everything geared to them form an elastic system with natural frequencies of its own. If the firing impulses arrive at a frequency close to one of those natural frequencies, the twisting oscillation builds up instead of dying away, and the stresses can exceed anything the steady loads would produce. The consequences are visible in the way engines are certificated and operated:

- Some installations carry a **restricted rpm range** that must be transited rather than operated in continuously. It is a placarded limitation and it exists because of a torsional resonance, not because the engine is generally unhappy at that speed.
- Larger and higher-powered engines carry **dynamic counterweights** on the crankshaft — pendulum masses hung on the crank webs, free to swing on oversized pins so that they oscillate out of phase with the firing impulses and absorb the torsional energy. They are tuned to a particular order of vibration, so their pins and bushings are matched components that must not be interchanged.
- **The propeller is part of the system.** Changing to a propeller of different mass or polar moment of inertia changes the natural frequency of the whole assembly, which is why propeller and engine combinations are approved together and why an apparently harmless substitution requires approved data.



Crankshaft and propeller form one elastic system; the propeller's mass helps set its natural frequency.

A rough-running engine is a symptom, not a diagnosis. Uneven running can come from combustion — a dead cylinder, a fouled plug, an induction leak on one cylinder, a valve not sealing — or from the mechanical sources above, such as a failed engine mount, an out-of-balance propeller or a damaged dynamic counterweight bushing. The two families feel different and are found by different means: the first shows up in the cylinder-by-cylinder indications and the magneto check, the second is largely independent of power setting and mixture and often changes character with rpm. Deciding which family the symptom belongs to is the first step, and it is the step most often skipped.

Cylinder Numbering

A firing order is a list of cylinder numbers, so it means nothing without the numbering convention it was written for, and conventions are not universal.

On a **radial** engine the convention is settled and worth memorising: number 1 is the cylinder at the top, and the remaining cylinders are numbered in order around the circle, counted clockwise when the engine is viewed from the rear. On a multi-row radial the front row and the rear row are numbered as part of one continuous sequence rather than separately.

On **horizontally opposed** engines the position of cylinder number 1 is a manufacturer's choice, and the two principal manufacturers of these engines do not start from the same end of the crankcase. An engineer who takes a firing order from one manufacturer's data and applies it to another manufacturer's cylinder numbering will connect the ignition leads to the wrong cylinders while believing the sequence to be correct. Always take the numbering diagram and the firing order from the same document, and that document must be the manual for the engine actually installed.

Firing Order

The **firing order** is the sequence in which cylinders fire. It is designed to:

- **Distribute thermal stress** — avoid adjacent cylinders firing consecutively
- **Minimise vibration** — balance the crankshaft loading
- **Ensure smooth power delivery** — evenly space power strokes

Typical firing orders:

- **4-cylinder opposed:** 1-3-2-4 or 1-4-2-3
- **6-cylinder opposed:** 1-4-5-2-3-6
- **Radial (5-cylinder):** 1-3-5-2-4 (every other cylinder around the circle)

In a 4-stroke engine, power strokes are spaced every **720° ÷ N** of crankshaft rotation (where N = number of cylinders). For a 4-cylinder engine: 720° ÷ 4 = **180°** between power strokes.

How a Firing Order Is Constructed

A firing order is not chosen for elegance. It is the solution to several constraints that have to be reconciled with one another, and understanding them makes the published sequences memorable rather than arbitrary.

- **Even spacing of the power impulses.** The cycle must be divided equally between the cylinders so that the crankshaft receives a push at regular intervals rather than in bursts. This is what makes the torque delivery smooth.
- **Alternation between banks, or around the circle.** Consecutive firings are placed on opposite sides of the engine wherever possible, so that the loads on the crankshaft and crankcase swap from side to side instead of hammering the same region repeatedly.
- **Separation of neighbours in time.** Two cylinders that are physically adjacent should not fire one after the other wherever the crankshaft layout allows it. If they did, one region of the crankcase and cylinder-head casting would take two heat pulses in quick succession, and on a shared induction manifold the second

cylinder would be trying to draw its charge through a passage the first had just emptied — the effect called induction robbing, which leaves that cylinder running lean. On a four-cylinder opposed engine that separation cannot be achieved in full. The four cylinders are carried as two facing pairs, and the two cylinders of a pair reach top dead centre together, so they fire a full revolution apart rather than one after the other; with a power stroke every 180 degrees the sequence must therefore alternate from one pair to the other. Each pair has one cylinder on each side of the engine, so two of the four transitions are bound to fall between the two cylinders on the same side, whichever of the orders listed above for that engine is used, and no reordering can avoid it.

- **Crankshaft balance.** The order has to work with the crank throw arrangement, not against it, so that the firing impulses do not reinforce the out-of-balance couples.

The even-spacing constraint is the one that can be calculated. In a four-stroke engine the cycle occupies 720 degrees of crankshaft rotation and each cylinder contributes one power stroke to it, so the interval between firings is 720 degrees divided by the number of cylinders. A **two-stroke** engine completes its cycle in 360 degrees, so the same reasoning gives an interval of 360 divided by the number of cylinders — another reminder that a firing interval quoted without the cycle it belongs to is meaningless.

Firing intervals for common engines (four-stroke):

- Six-cylinder: 720 ÷ 6 = **120 degrees** between firings
- Five-cylinder radial: 720 ÷ 5 = **144 degrees**
- Nine-cylinder radial: 720 ÷ 9 = **80 degrees**
- Fourteen-cylinder twin-row radial: 720 ÷ 14 = about **51.4 degrees**

There is a threshold worth spotting in those numbers. A power stroke occupies 180 degrees of crankshaft rotation, so at a firing interval of 180 degrees a four-cylinder four-stroke engine has exactly one cylinder producing torque at any instant, with no overlap at all. At 120 degrees a six-cylinder engine always has one and a half; at about 51 degrees a fourteen-cylinder radial has three and a half. Fewer than four cylinders and the interval exceeds 180 degrees, leaving gaps in which no cylinder is delivering anything and the rotating inertia has to carry the engine through unaided.

Following a six-cylinder opposed sequence through the cycle. Taking the order 1-4-5-2-3-6 listed above and applying the 120 degree interval, the cylinders fire at:

- 0 degrees — cylinder 1
- 120 degrees — cylinder 4
- 240 degrees — cylinder 5
- 360 degrees — cylinder 2
- 480 degrees — cylinder 3
- 600 degrees — cylinder 6
- 720 degrees — cylinder 1 again, and the cycle repeats

Six firings, evenly spaced, with every cylinder used exactly once in the two revolutions. Any valid firing order must pass this test; a sequence that repeats a cylinder or leaves one out inside 720 degrees is simply wrong.

Firing Order in the Hangar

Firing order is one of the few pieces of pure theory in this subject that an engineer uses directly, with a spanner in hand.

- **Ignition harness connection.** The distributor block of a magneto has its terminals numbered in the order in which the magneto distributes the sparks, **not** by cylinder number. The leads therefore run from terminal 1, 2, 3 and so on to the cylinders taken in firing order, not in numerical order. Connecting the harness "logically" — terminal 1 to cylinder 1, terminal 2 to cylinder 2 — produces an engine that may still start and run, roughly and with a distinctive misfire pattern, which makes the fault harder to recognise than an outright failure would be.
- **Magneto timing.** The engine is timed with a nominated reference cylinder, conventionally number 1, brought to the specified number of degrees before TDC **on its compression stroke**. The qualification matters: the crankshaft passes TDC twice in the cycle, and only one of those is the firing position for that cylinder. Confirming that both valves are closed, or feeling compression at the plug hole, is how the correct one is identified.
- **Compression checks.** A differential compression check requires each cylinder in turn to be brought to TDC on its compression stroke. Working through the cylinders in firing order means the propeller is turned by only the firing interval between each test rather than being rotated back and forth, which is quicker and safer.
- **Fault interpretation.** When an engine runs rough, knowing which cylinders fire consecutively and which share an induction passage turns a vague symptom into a short list of suspects.

Never infer a firing order from the cylinder count. Two engines with the same number of cylinders in the same configuration, from different manufacturers, may use different firing orders **and** different cylinder numbering. Getting either of them from the wrong document, or from memory, will cross-connect the ignition leads. The firing order and the cylinder numbering diagram must both come from the maintenance manual for the specific engine model and dash number in front of you, and the harness should be checked lead by lead against that diagram before the cowling goes back on.

Module 17 — Propeller

17.1 — Fundamentals

Introduction

A propeller is an **aerodynamic device** that converts the rotary power (torque) produced by an engine into **thrust** — a forward-acting force that propels the aircraft through the air. Although jet engines dominate large transport aircraft, propellers remain essential on turboprops, piston-engined aircraft, and many modern regional airliners. Every EASA Part 66 engineer must thoroughly understand propeller theory, because the propeller is the final link in the powerplant chain — if it is inefficient, damaged, or incorrectly set, no amount of engine power will produce adequate performance.

A propeller blade is, in every meaningful sense, a **rotating wing**. It has an aerofoil cross-section, generates lift (which we call thrust when directed forward), produces drag, and is subject to the same aerodynamic principles that govern a wing. The key difference is that while a wing translates through the air, a propeller blade **rotates and translates simultaneously**, giving each blade element a unique combination of rotational velocity and forward velocity.

Blade Element Theory

Blade element theory is the fundamental method for analysing propeller performance. Rather than treating the blade as a single unit, blade element theory divides each blade into a series of small **independent sections** (elements), each at a different distance from the hub. Each element is then analysed as a tiny aerofoil operating in its own local airflow.

Why is this necessary? Because conditions vary dramatically along the blade:

- **Near the hub** — the rotational speed is low (small radius × RPM), so the element "sees" a relatively slow airflow
- **Near the tip** — the rotational speed is very high (large radius × RPM), so the element "sees" a much faster airflow
- At each station, the blade angle, angle of attack, lift, drag, and local efficiency are all different

Rotational velocity at any blade station:

$$V_{rot} = 2\pi r n$$

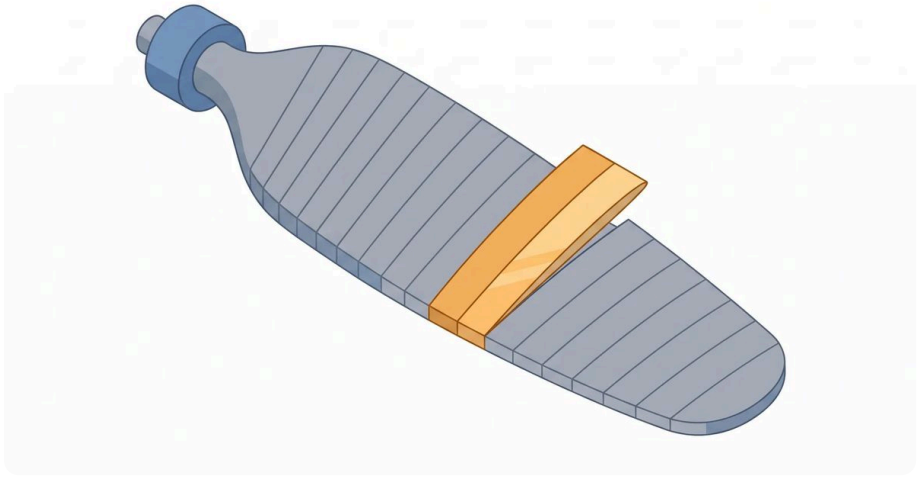
Where: r = distance from hub centre (m), n = rotational speed (rev/s)

At the tip of a 2 m diameter propeller spinning at 2,400 RPM (40 rev/s):

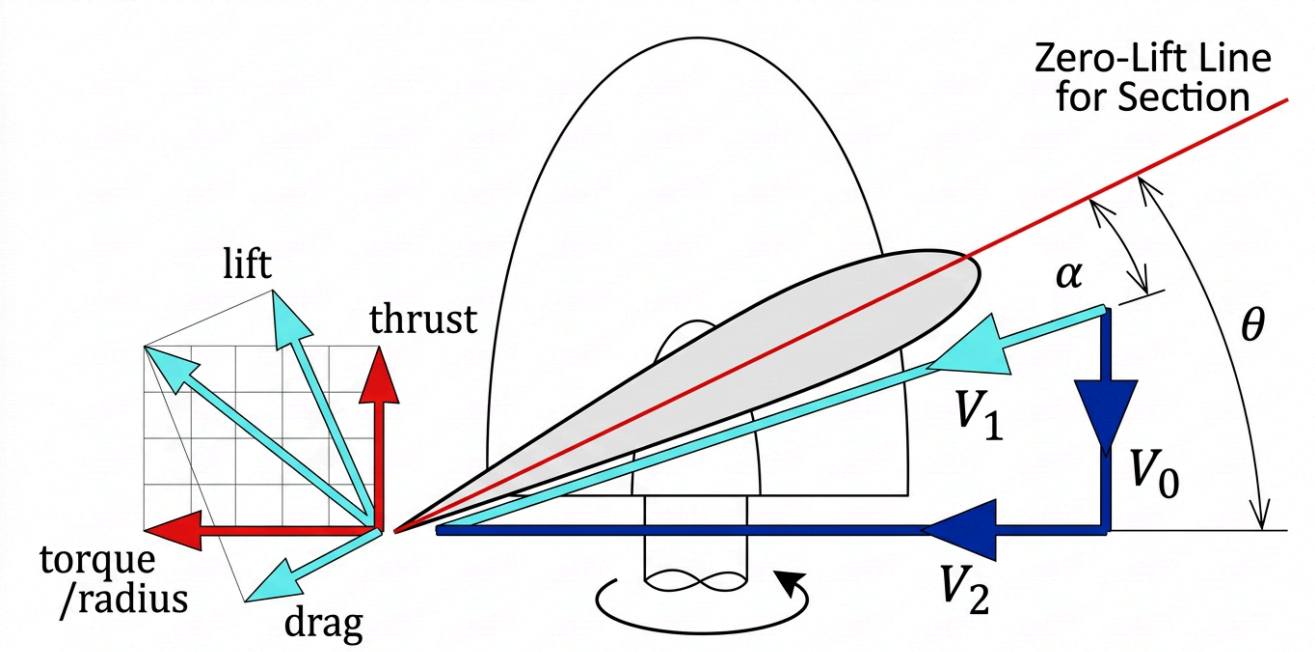
$$V_{rot} = 2\pi \times 1.0 \times 40 = 251 \text{ m/s} \approx 905 \text{ km/h}$$

This is approaching the speed of sound — which is why propeller tips can go supersonic, causing noise and efficiency loss.

Blade element theory sums (integrates) the thrust and torque contributions of every element along the blade span to calculate the total propeller thrust and the power required to turn it. The key insight is that **each element must operate at an efficient angle of attack**, which is why propeller blades are **twisted** — the blade angle is large near the hub and small near the tip.



Blade element theory slices each blade into narrow strips, each treated as a tiny aerofoil, then sums them.



Key Relationship: β = ϕ + α

The diagram above shows the fundamental angular relationship for any blade element:

Blade angle = Helix angle + Angle of attack

$$\beta = \varphi + \alpha$$

- **β (beta) — Blade angle**: the angle between the blade chord line and the plane of rotation. This is a fixed geometric property of the blade at any given station (unless the propeller has variable pitch).
- **ϕ (phi) — Helix angle**: the angle between the relative airflow (RAF) and the plane of rotation. This changes with forward speed and rotational speed.
- **α (alpha) — Angle of attack**: the angle between the chord line and the relative airflow. This determines how much lift (thrust) the blade element produces.

Reading the Velocity Diagram

The velocity diagram in this section is the working drawing of the whole theory, and it repays being read symbol by symbol. It shows one blade section, cut at some radius *r*, viewed edge-on from the blade tip so that the plane of rotation runs left to right across the page. Three velocities meet at that section:

- **V_0** — the axial inflow, or advance, velocity: the speed at which the air approaches the disc along the shaft axis. On an aircraft in flight this is essentially the true airspeed, increased slightly by the air the propeller

has already begun to draw in.

- **V_2** — the rotational velocity of the section, equal to the angular velocity multiplied by the radius. It lies in the plane of rotation and it is the component that grows steadily from hub to tip.
- **V_1** — the relative inflow velocity: the vector sum of the other two, and the airflow the section actually experiences. It is what the aerofoil "feels", and both the lift and the drag of the section are defined with respect to it.

Two angles are then marked against those velocities. The **angle of attack α** is the angle between the relative inflow V_1 and the section's zero-lift line. The **blade pitch angle θ** is the angle between the plane of rotation and that same zero-lift line. The gap between the two — the angle between V_1 and the plane of rotation — is the inflow or helix angle, and it is the angle that changes continuously as the aircraft speeds up, slows down or changes RPM.

Two angular datums, and why they differ

The velocity diagram labels the blade pitch angle **θ** (theta) where the relationship stated above uses **β** (beta). Propeller, rotor and wind-turbine texts use both symbols for the pitch of the blade, and you will meet each of them; the symbol is not the point. The *datum* is. That diagram measures both the pitch angle and the angle of attack from the section's **zero-lift line** — the direction from which the air would have to approach for the section to develop no lift at all. The forces diagram later in this note, and every blade-angle measurement you will ever take on an aircraft, use the **chord line** instead.

The two datums are separated by the aerofoil's zero-lift angle. For a **symmetrical** section the zero-lift line and the chord line are the same line and the distinction disappears. For a **cambered** section — which is what a propeller blade uses — the section still produces lift at a small negative chord-line angle, so the zero-lift line sits a few degrees away from the chord, and both the pitch angle and the angle of attack come out **numerically larger** when quoted from the zero-lift line than when quoted from the chord. Aerodynamicists prefer the zero-lift line because it makes the section's lift proportional to the angle from the very first degree; engineers use the chord line because it is a physical edge-to-edge line you can lay a protractor against. Always check which datum a figure or a data sheet is using before you compare two numbers.

From Section Lift and Drag to Thrust and Torque

Once the relative inflow at a station is known, the section is treated exactly as a two-dimensional aerofoil: it produces a lift force perpendicular to the relative inflow and a drag force parallel to it. Neither of those is directly useful to the aircraft, because both are tilted relative to the flight path by the helix angle. The diagram therefore resolves the pair into the two components an engineer cares about — a component along the shaft axis, which is **thrust**, and a component in the plane of rotation, which multiplied by the radius is the **torque** the engine has to supply.

Elemental thrust and torque

For one element of span *dr* producing elemental lift *dL* and elemental drag *dD* at helix angle ϕ:

$$dT = dL\cos\varphi - dD\sin\varphi$$

$$dQ = r\left(dL\sin\varphi + dD\cos\varphi\right)$$

Read the signs. Drag **subtracts** from thrust and **adds** to torque — it costs you the useful force and increases the bill you pay for it, which is why every scratch, dent, insect deposit and paint run on a blade is worth removing. Lift contributes to both: it is the source of the thrust, but its in-plane component is also part of the load the engine must overcome.

Both expressions contain the helix angle, so both change the moment the aircraft's speed or the propeller's RPM changes, even though the blade itself has not moved. That single fact is the reason a propeller behaves so differently on the take-off run and in the cruise, and it is the reason a pitch-change mechanism was worth inventing.

Worked Example: How the Inflow Angle Varies Along One Blade

A four-blade propeller in the cruise

Diameter 2.5 m (tip radius 1.25 m), 1,200 RPM, true airspeed 100 m/s. Convert the RPM first: $n = 1200 \div 60 = 20$ rev/s. At each station the rotational velocity is $2\pi r n$, and the helix angle follows from $\tan\varphi = V/2\pi r n$.

Station	Radius r	Rotational velocity	Helix angle ϕ	Blade angle for α = 3°
20% (inner blade)	0.25 m	31.4 m/s	72.6°	75.6°
50%	0.625 m	78.5 m/s	51.9°	54.9°
75% (reference)	0.9375 m	117.8 m/s	40.3°	43.3°
100% (tip)	1.25 m	157.1 m/s	32.5°	35.5°

Check one line by hand: at the 75% station $2\pi \times 0.9375 \times 20 = 117.8$ m/s, so $\tan\varphi = 100/117.8 = 0.849$ and $\varphi = 40.3^\circ$. Adding the 3° angle of attack gives a blade angle of 43.3° at that station.

The last column is the twist distribution, produced from nothing but geometry: **75.6° at the inner blade falling to 35.5° at the tip**, a change of forty degrees across one blade, purely so that every station works at the same 3°. This is why the blade angle at the root is greater than the blade angle at the tip, and it is why any blade-angle figure is meaningless unless the station it was measured at is quoted with it.

What an Untwisted Blade Would Do

The clearest way to see why the twist is not a refinement but a necessity is to remove it. Take the same propeller at the same 1,200 RPM and 100 m/s, but set every station to a single blade angle of 45°. The helix angles are unchanged — they depend on the airflow, not on the blade — so the angle of attack at each station is simply 45° minus the figure in the table above:

- At the tip: 45° – 32.5° = **+12.5°** — far past the stall for a propeller section, producing turbulent separated flow, very high drag and very little thrust.
- At 75% radius: 45° – 40.3° = **+4.7°** — roughly right, and the only part of the blade doing sensible work.
- At 50% radius: 45° – 51.9° = **–6.9°** — a negative angle of attack, so this part of the blade is producing *rearward* force and being driven by the airflow rather than driving it.
- At 20% radius: 45° – 72.6° = **–27.6°** — a large negative angle, a substantial drag penalty and a torque the engine must pay for while receiving nothing in return.

An untwisted blade is therefore not merely inefficient, it works against itself: the inner half fights the outer half. Twisting the blade so that the angle falls from root to tip removes the conflict and lets the whole span pull in the same direction, which is the same thing as saying it evens out the distribution of thrust along the blade.

Where the Thrust Actually Comes From

Even on a well-twisted blade the thrust is not spread evenly. Summing the elemental thrust across the span gives a distribution that starts at essentially zero at the hub, rises to a broad maximum in the outer third, and then falls back to zero at the tip. The peak sits at roughly **70% to 80% of the blade length**, which is why that region is treated as the working part of the blade and why the reference station used for quoting blade angle is normally placed within it.

Three separate effects push the distribution into that shape:

- **The root produces little because it cannot.** The rotational velocity there is small, so the dynamic pressure at the section is small. Worse, the inner blade has to be thick and often circular in cross-section where it enters the hub, because that is where the entire centrifugal load of the blade is carried; a round shank is a structural member, not an aerofoil, and it makes no thrust at all.
- **The tip loses what it makes.** The pressure difference between the two faces of the blade cannot be sustained around the tip, so air spills from the high-pressure face to the low-pressure face and rolls up into a trailing vortex. Lift falls away to zero over the last few per cent of span. The blade area is also smallest there.
- **The outer third has the best of both.** High local velocity, a true aerofoil section, full chord, and still far enough inboard for the tip loss not to have bitten. It is doing most of the work.

⚠ Why tip damage matters out of all proportion

Because thrust peaks in the outer third and the local airspeed there is the highest anywhere on the blade, damage to the outer span costs performance far more than the same damage further in. The outer blade is also where the bending stress produced by thrust and by any vibratory loading is being reacted through the greatest lever arm. That is the reasoning behind the tighter dressing and blending limits that propeller maintenance manuals apply near the tip, and behind the rule that a nick which is acceptable at mid-span may be rejectable at the same depth further out. Never carry a limit across from one blade station to another, and never carry it across from one propeller type to another — the figures are specific to the propeller and are only valid from its own manual.

The Aerofoil Changes Along the Blade Too

A propeller blade is not one aerofoil twisted; it is a family of different aerofoils blended into one another. The thickness-to-chord ratio falls steadily from root to tip, because the two ends of the blade are solving different problems.

- **Near the root** the section is thick and deeply cambered. Thickness gives the bending and tensile strength needed to carry the whole blade's centrifugal load, and the low local velocity there means the drag penalty of a thick section is modest.
- **Near the tip** the section is thin and only slightly cambered — a **high-speed aerofoil**. The tip is travelling faster through the air than any other part of the aircraft, so its local Mach number is the highest anywhere on the airframe. A thin section keeps the local flow acceleration over the upper surface low, which delays the formation of shock waves and postpones the drag rise to a higher tip Mach number.

Modern blades add planform shaping to the same end: sweeping the outer blade back, thinning the tip and, on some designs, curving the tip forward or aft so that the shock does not form along the whole span at once. All of it is aimed at the same target — keeping the tip out of trouble while the rest of the blade gets on with producing thrust.

Blade Stations, the Reference Station, and Solidity

Because every property of a blade varies along its length, the blade is divided into numbered **stations**, measured as a distance from the centre of the hub. Manufacturers quote chord, thickness, twist and blade angle station by station, and every drawing, repair scheme and inspection limit is tied to a station number.

One of those stations is nominated as the **reference station**, also called the master station. It is the single station at which the propeller's blade angle is defined and against which the whole blade is set. When a manual says the blade angle is a particular value, or a pitch stop is set to a particular value, it means at that

station and nowhere else. The reference station is commonly placed in the working outer part of the blade, and the 75% radius is a widely used choice, but the actual station is specified by the propeller manufacturer and must be taken from the manual — it is a type-specific figure, not a universal one.

The reference station on the aircraft

Blade angle is checked with a **propeller protractor** (a bubble or digital inclinometer laid on the blade face) or with a **master blade gauge** — a shaped template made for that propeller type which locates on the blade at the reference station. Two conditions have to be satisfied before any reading means anything: the propeller must be positioned so that the reference blade is at the datum position the manual specifies, and the aircraft or the propeller must be levelled so the protractor has a true reference. Setting a blade angle at the wrong station, or with the reference surface unlevelled, produces a number that looks plausible and is wrong, and the error carries straight through into the low-pitch stop, the feather angle and the reverse angle.

The last geometric property worth naming here is **solidity**: the ratio of the total blade planform area to the area of the disc the propeller sweeps. It is a measure of how much of the disc is actually blade, and it governs how much power the propeller can absorb for a given diameter. Solidity is raised in only two ways — by making the blades **wider in chord**, or by fitting **more of them**. It is not affected by blade angle, which changes the pitch but not the planform area, and it is not affected by section thickness, which changes the aerofoil but again not the area seen from the front.

Solidity worked through

Four blades of mean chord 0.25 m on a 2.5 m diameter propeller. Blade planform area is approximately $4 \times 0.25 \times 1.25 = 1.25$ m²; disc area is $\pi \times 1.25^2 = 4.91$ m². Solidity is $1.25 / 4.91 = 0.25$.

Widen the mean chord to 0.30 m and the blade area becomes 1.50 m^2, so solidity rises to $1.50 / 4.91 = 0.31$. The disc has not changed size; more of it is now blade.

Solidity is the designer's answer to the problem of absorbing a lot of power without a large diameter. When ground clearance or tip speed forbids a bigger disc — which on a modern turboprop is nearly always — the power has to be absorbed by widening the blades, adding blades, or both. That is why high-powered turboprops carry five, six or even eight relatively broad blades on a modest diameter, and why older, lower-powered aircraft manage with two slender ones. A closely related figure quoted in propeller data, the **activity factor**, expresses the same idea per blade, weighting the chord by how far out along the blade it is placed, because chord near the tip absorbs far more power than the same chord near the root.

What the Theory Assumes, and Where It Breaks Down

Blade element theory in its simplest form makes three assumptions, and knowing them is the difference between using the theory and believing it:

- **Each element is independent.** The theory treats a section as a two-dimensional aerofoil with no awareness of its neighbours. Real blades carry a spanwise flow outwards along the blade, driven by centrifugal effects in the boundary layer, and that flow couples the elements together.
- **The inflow is the free-stream.** Taking the axial inflow as simply the aircraft's speed ignores the fact that the propeller has already accelerated the air before it reaches the disc. The true inflow is higher, so the true helix angle is larger and the true angle of attack is smaller than the simple calculation suggests. Combining blade element theory with momentum theory — the standard **blade element momentum** method — closes that gap by solving for the induced velocity and the blade forces together.
- **The section behaves linearly.** Lift is taken as proportional to angle of attack. That holds until the section stalls or until the local Mach number climbs into the compressibility region, at which point the two-dimensional data the theory relies on stops applying.

None of that makes the theory less useful. It remains the framework that explains blade twist, the shape of the thrust distribution, the effect of forward speed on angle of attack, the reason for a pitch-change mechanism and the reason the tip is designed differently from the root. Every one of those is a conclusion an engineer uses on the line, and every one of them falls straight out of the velocity triangle.

⚙ Blade Angle — High, Low, and Reverse

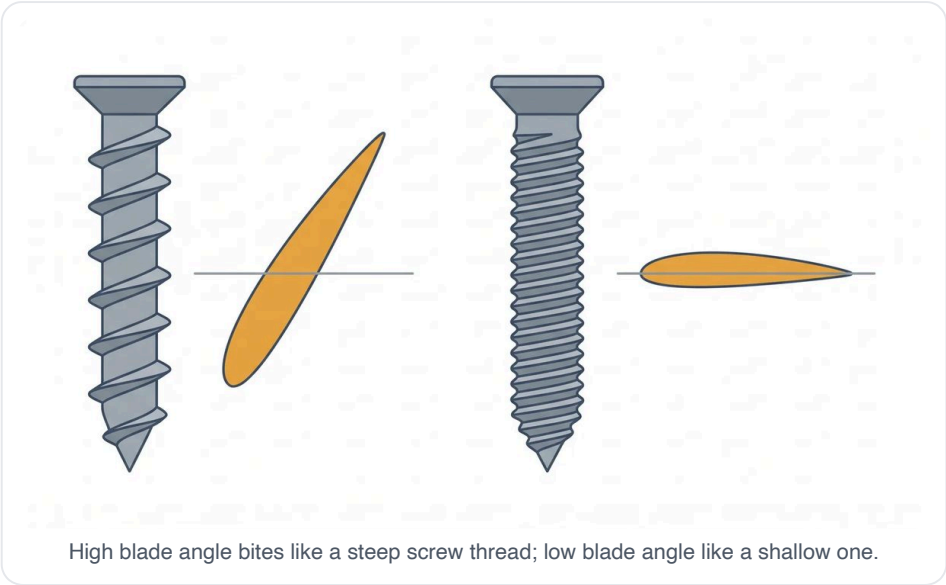
High Blade Angle (Coarse Pitch)

A **high blade angle** (also called **coarse pitch**) means the blade chord is angled steeply relative to the plane of rotation. The blade takes a bigger "bite" of air per revolution — like a screw with a steep thread. This is analogous to a high gear in a car.

- Used during **cruise and high-speed flight**
- Each revolution moves the aircraft a greater distance through the air
- Requires more engine torque (more load on the engine)
- Lower RPM for a given power setting → better fuel efficiency at speed
- Inefficient at low speed because the angle of attack becomes too large, risking blade stall

Low Blade Angle (Fine Pitch)

A **low blade angle** (also called **fine pitch**) means the blade chord is nearly flat relative to the plane of rotation. The blade takes a small bite of air per revolution — like a shallow screw thread. This is analogous to a low gear in a car.



High blade angle bites like a steep screw thread; low blade angle like a shallow one.

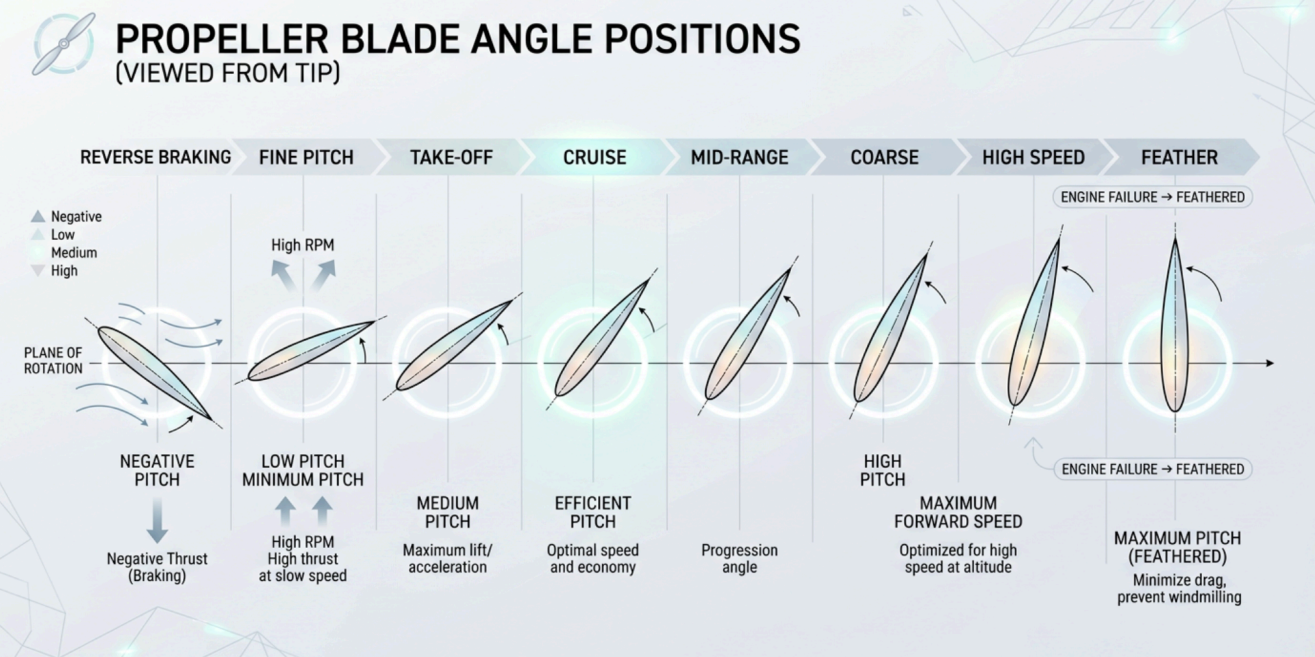
- Used during **take-off, climb, and low-speed flight**
- Allows the engine to develop **high RPM** without excessive load
- Produces maximum thrust at low forward speeds
- Inefficient at high speed because the blade merely "slices" the air with little forward advance per revolution

Reverse Angle (Reverse Pitch)

In **reverse pitch**, the blade angle is rotated past the flat (zero-thrust) position so that the leading edge now faces aft. The propeller produces **reverse thrust** — a braking force directed forward (opposing the aircraft's motion). This is used:

- On the ground after landing to **shorten the landing roll**
- For **ground manoeuvring** — allows the aircraft to taxi backwards or turn sharply
- In some turboprop aircraft, as a **primary braking method** on short runways

Exam caution: Do not confuse *reverse pitch* with *feathering*. Reverse pitch produces reverse thrust for braking. Feathering aligns the blade with the airflow (≈ 90° blade angle) to **minimise drag** after an engine failure. They are opposite ends of the blade angle range.



The Definition, and the Two Angles It Is Constantly Confused With

Blade angle is a purely **geometric** quantity: it is measured between the blade section's **chord line** — the straight line joining the leading edge to the trailing edge — and the propeller's **plane of rotation**, the flat disc swept by the blades at right angles to the shaft axis. Nothing about the airflow enters into it. You can measure a blade angle on a propeller lying on a bench in a hangar, and it will read exactly the same figure it reads at 250 kt.

That independence from the airflow is the single fact that separates blade angle from the two angles students most often mistake it for:

Angle	Measured between	Changes with	Other names
Blade angle	Chord line and plane of rotation	Only a pitch-change command; nothing aerodynamic	Pitch angle, blade pitch, β
Helix angle	Resultant relative airflow and plane of rotation	Forward speed, RPM and blade station	Angle of advance, advance angle, φ
Angle of attack	Chord line and resultant relative airflow	Everything above — it is the difference of the other two	α

All three share the same two reference lines between them — chord, plane of rotation, relative airflow — which is exactly why they are easy to swap by accident. The safe habit in an exam is to read the question for the *pair* of lines it names, not for the words "propeller" and "angle". If the plane of rotation and the chord are named, it is blade angle. If the plane of rotation and the airflow are named, it is the helix angle. If the chord and the airflow are named, it is angle of attack.

Blade Angle and Pitch Are the Same Idea in Different Units

"Pitch" borrows its meaning from a screw thread: the distance advanced per turn. A propeller's geometric pitch at a station follows directly from the blade angle there, because the blade sweeps out a helix whose circumference is $2\pi r$ and whose rise per turn is fixed by the angle:

Blade angle to geometric pitch

$$GP = 2\pi r \tan \beta$$

Because $\tan \beta$ increases as β increases, **raising the blade angle raises the geometric pitch**. This is the whole content of the statement that increasing blade angle makes the pitch coarser and decreasing it makes the pitch finer. Coarse pitch and high blade angle are two descriptions of one setting; fine pitch and low blade angle are two descriptions of the other.

The relationship also explains the blade's twist from the other direction. If the designer wants the whole blade to describe one clean helical surface — the same advance per revolution at every station, like a single thread — then $r \tan \beta$ must be the same number at every station. Since *r* grows steadily towards the tip, $\tan \beta$ must fall in the same proportion, so the blade angle must be large at the root and small at the tip. Two entirely different arguments, one aerodynamic and one geometric, arrive at the same twist.

Checking that the twist really does hold the pitch constant

Using the four blade angles calculated earlier in this note for a 2.5 m propeller at 1,200 RPM and 100 m/s, the geometric pitch at each station is $2\pi r \tan \beta$:

- 20% station: $2\pi \times 0.25 \times \tan 75.6^\circ = 1.571 \times 3.895 = 6.12$ m
- 50% station: $2\pi \times 0.625 \times \tan 54.9^\circ = 3.93 \times 1.42 = 5.58$ m
- 75% station: $2\pi \times 0.9375 \times \tan 43.3^\circ = 5.89 \times 0.94 = 5.56$ m
- Tip: $2\pi \times 1.25 \times \tan 35.5^\circ = 7.85 \times 0.71 = 5.60$ m

Over the outer part of the blade three very different blade angles give three almost identical pitches — 5.58 m, 5.56 m and 5.60 m. The inner station is the odd one out at 6.12 m, about 10% coarser than the rest, so the agreement is close over the working outer blade rather than right down to the root. Across that outer span the twist that holds the angle of attack constant is very nearly the same twist that holds the geometric pitch constant, which is why a propeller can be quoted as, for example, a "2.5 m by 5.6 m" propeller and have that single pitch figure mean something for the part of the blade that does most of the work.

The Full Range of Blade Positions

The diagram of blade positions above shows the same blade section drawn eight times, viewed from the tip, at progressively larger angles to the plane of rotation. Taken together they span everything a propeller blade can be asked to do, from braking the aircraft on the runway to standing edge-on to the airflow with the engine stopped.

Position (increasing blade angle)	What the diagram states about it	Where it is used
Reverse braking	Negative pitch giving negative thrust — a braking effect	After touchdown, and for manoeuvring on the ground
Fine pitch	Low, minimum blade angle; high RPM and high thrust at slow speed	Ground running, engine start, low-speed operation
Take-off	Medium blade angle for maximum acceleration	Take-off run and initial climb
Cruise	An efficient angle giving optimum speed and economy	Level cruise
Mid-range	An intermediate angle in the progression towards coarse	Continuous adjustment between cruise settings
Coarse	A high blade angle	High-speed and high-altitude cruise
High speed	Maximum forward-speed setting, optimised for high speed at altitude	Maximum cruise speed
Feather	Maximum blade angle, to minimise drag and prevent windmilling after an engine failure	Engine shutdown in flight; parked aircraft on some types

Read the sequence as a continuum rather than as a set of detents. On a constant-speed propeller the blade does not jump between named positions; the governor moves it smoothly and continuously through the whole flight range in response to load, and the names simply label the regions the blade passes through. Only the two extremes — the low-pitch stop at one end and the feather stop at the other — are physical, mechanical positions.

Reading it that way also reconciles two ways of speaking that both appear in this note. The diagram separates a minimum *fine pitch* position from a slightly larger *take-off* position, because it is drawing individual angles. In ordinary usage, and in the description of low blade angle above, "fine pitch" names the whole low-angle end of the range — the region a propeller works in for take-off, climb and low-speed flight. Both descriptions point at the same end of the same range: one names a specific angle, the other names a region.

The Two Operating Ranges: Alpha and Beta

On a turboprop the blade angle range is split into two operating regimes, and understanding which one is active explains almost every apparently strange propeller behaviour on the ground:

- **The alpha, or governing, range** covers everything from the flight low-pitch stop up to feather. Here the pilot selects an *RPM*, and the governor sets whatever blade angle is needed to hold it. The power lever controls fuel flow and therefore power; the blade angle is the governor's business, not the pilot's.
- **The beta range** lies below the flight low-pitch stop, and covers ground fine and reverse. Here the relationship is turned round: the power lever is mechanically or electronically linked directly to the *blade angle*, and the pilot is commanding pitch rather than RPM. This is what allows a fine, low-drag blade angle for taxiing and a negative angle for braking.

The transition between the two is protected. A flight low-pitch stop — variously a mechanical latch, a hydraulic stop or an electrically controlled solenoid depending on type — prevents the blade angle falling below the flight minimum while the aircraft is airborne, and is only released by a weight-on-wheels signal or an equivalent ground-mode interlock. The reason is stark: a blade angle that is fine enough to be safe on a runway is fine enough at flight speed to allow an immediate and destructive overspeed, and to produce so much drag that the aircraft may become uncontrollable.

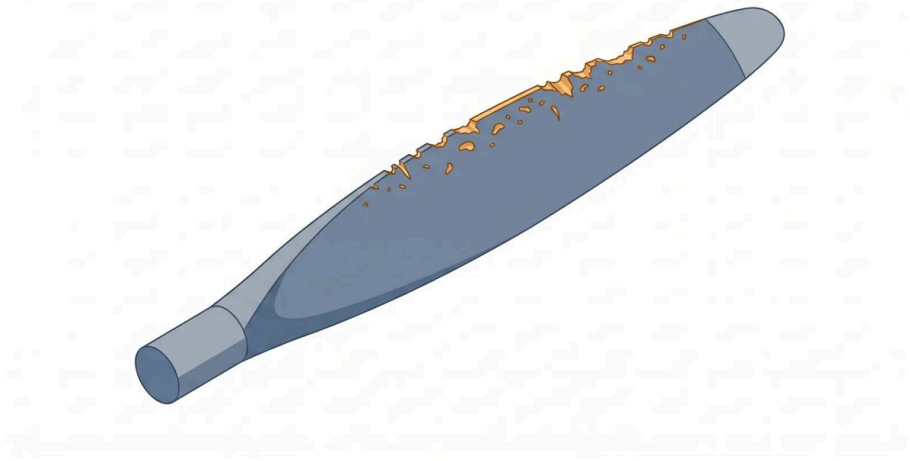
⚠ Why the low-pitch stop is a flight-safety item

If a blade goes below the flight fine stop in flight, two things happen at once and both are bad. The propeller loses its ability to absorb power, so the engine and propeller accelerate towards an overspeed limited only by the governor's remaining authority. At the same time the blade, now at a very fine angle in a fast airflow, presents an enormous flat-plate drag to the airstream — far more than a feathered or even a windmilling blade at a normal angle. Rigging, functional checking and leak-checking the low-pitch stop and its interlock is therefore never routine work, and the ground/flight interlock must be proved after any disturbance of the landing-gear proximity system as well as after work on the propeller itself.

Reverse in More Detail

Reverse is reached by driving the blade angle down through the flat, zero-thrust position and out the other side, to a small negative angle at the reference station. The propeller keeps turning in its normal direction and the leading edge still leads; what changes is that the chord is now inclined so that the leading edge lies *aft* of the plane of rotation instead of ahead of it. The disc therefore accelerates air **forwards**, ahead of the aircraft, and by reaction the aircraft feels a force acting rearwards along its flight path, opposing its motion down the runway. That is the braking effect.

Two consequences of the direction of that airflow matter operationally:



Debris thrown forward in reverse is flown into by the propeller, eroding the blade leading edges.

- **Debris is thrown forward.** On an unpaved, gravel, sandy or contaminated surface, reverse throws whatever is on the runway ahead of the aircraft, where the propeller and the engine intakes then fly into it. That is why many operators limit or prohibit reverse on such surfaces, why blade leading-edge erosion is markedly worse on operators who use full reverse routinely, and why reverse is normally cancelled at a specified low groundspeed before the aircraft's own re-ingestion becomes the problem.
- **The airflow over the tail is disrupted.** The normal propeller slipstream over the wing and tail disappears and is replaced by a disturbed, forward-moving flow. Aerodynamic control effectiveness at low speed falls away just when directional control on a wet or icy runway matters most, which is why reverse thrust is reduced or cancelled as the aircraft slows.

Reverse is also a genuine engine and propeller stress case, not a free extra. The blades are working at a large negative angle of attack in disturbed air; the loading is heavy, unsteady and quite unlike anything in the flight range. Manuals therefore place time limits on continuous reverse operation and specify minimum speeds for its use, and repeated use of maximum reverse is a recognised driver of blade erosion and of gearbox and reduction-gear wear.

Feather at the Other End of the Range

Feathering drives the blade to its maximum angle, so that the chord lies as nearly as possible along the direction of flight and presents its thinnest edge to the airflow. The blade angle needed to do that at the reference station is close to a right angle to the plane of rotation. It is worth being precise about what feathering achieves and what it does not:

- It reduces the propeller's drag to the minimum the installation can achieve, which on a twin directly improves single-engine climb performance and reduces the asymmetric yaw the pilot must trim out.
- It stops the propeller rotating, because at that angle the airflow can no longer produce a turning moment. That protects a damaged or oil-starved engine from being driven by the airstream, and it removes the vibration and noise of a windmilling propeller.
- It does **not** produce any braking force. That is the province of the opposite end of the range.

Because feathering usually has to work after the engine that supplies the propeller's oil pressure has failed, the feathering energy comes from somewhere that does not depend on the engine still turning: a spring pack in the hub, the blades' own counterweights, a nitrogen or air charge in the dome, an electrically driven feathering pump, or a combination of them. Unfeathering, by contrast, needs the engine oil pressure to be restored, or an accumulator or unfeathering pump to supply a burst of pressure to move the blades off the feather stop and let the airflow start the propeller turning again.

Fixed, Two-Position and Variable Pitch

Everything above assumes the blade angle can be changed. The reason variable pitch exists is best seen by looking at what a fixed-pitch propeller has to give up.

Type	Blade angle behaviour	Consequence
Fixed pitch	Set once at manufacture and never changes	Efficient at one combination of speed and RPM only. RPM varies with airspeed and throttle.
Ground adjustable	Blades can be reset in the hub on the ground	The operator can bias the propeller towards climb or cruise for a season or a role, but not in flight.
Two position	Selectable fine or coarse, nothing between	Fine for take-off and climb, coarse for cruise. A large improvement over fixed pitch for very little complexity.
Constant speed	Continuously varied by a governor to hold a selected RPM	Near-optimum angle of attack across the whole speed range; the propeller absorbs the power the engine is producing at all times.
Full range with reverse and feather	Constant speed, plus a beta range below flight fine and a feather position above coarse	Adds ground braking and manoeuvring at one end and engine-out drag reduction at the other.

A fixed-pitch propeller must therefore be chosen for a role. A **climb propeller** is given a finer blade angle: it lets the engine reach its rated RPM at low airspeed, so full power is available on take-off and in the climb, at the cost of the engine reaching its RPM limit early and the propeller running out of efficiency in the cruise. A **cruise propeller** is given a coarser angle: it is efficient at cruising speed but holds the engine RPM down on take-off, lengthening the ground roll and reducing climb performance. Nothing is available that is good at both, which is the entire commercial case for a constant-speed unit.

What "high-speed propeller" means in an exam

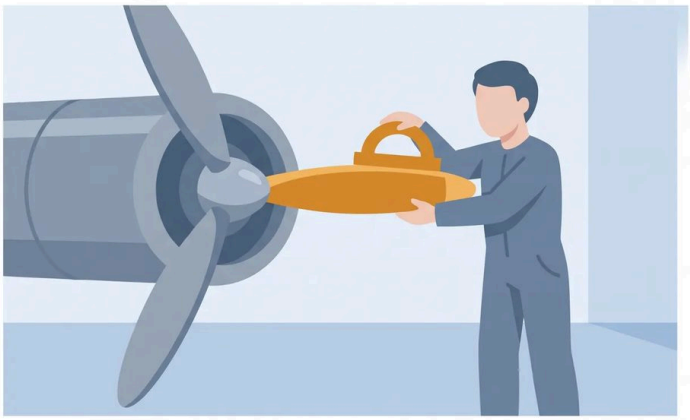
A propeller described as a high-speed design is optimised to be efficient at **high forward airspeeds** — not to turn at high RPM, and certainly not to run supersonic tips. The two things a high-speed design does are to use a blade angle distribution suited to the large helix angles that a fast airflow produces, and to keep the tip's helical Mach number down with thin, often swept tip sections. Since forward speed and rotational speed both feed the tip's total speed, a design aimed at high forward speeds generally has to accept a *lower* rotational speed, not a higher one, to keep the tip subsonic. Tip speed is something a propeller designer spends effort avoiding, never a design goal.

Blade Angle in Maintenance Practice

Almost everything an engineer physically does to a propeller comes back to blade angle. The pitch stops define the ends of the range, and each is a separate rigging task with its own procedure and its own check:

- **The flight low-pitch stop** sets the finest angle available in the governing range, and therefore sets the RPM the engine will reach at a given power in flight. Setting it too fine invites overspeed; setting it too coarse loses take-off performance.
- **The ground fine and reverse stops** define the beta range, and are set with the power lever rigging so that the commanded angle matches the lever position through the whole travel.
- **The feather stop** defines the maximum angle, and is checked because too little angle leaves residual drag and rotation, while too much can jam the blade against the stop or prevent unfeathering.

Every one of those angles is quoted at the propeller's reference station, and every one of the figures is type-specific — take them from the propeller maintenance manual for the exact part number and modification standard fitted, never from another type or from memory. The blade angles are checked with a protractor or the type's master gauge, and a functional check of the resulting RPM and torque against the manual's ground-running figures is the proof that the rigging is right.



Blade angle is checked with a protractor or master gauge at the propeller's reference station.

One more thing follows from blade angle being a per-blade property: the blades on a hub must agree with one another. A blade set even slightly differently from its neighbours produces a different thrust at the same instant, which is an aerodynamic imbalance that the propeller then advertises as a once-per-revolution vibration. Blade angle discrepancy between blades on the same hub is checked as part of the same procedure, and it is one of the first things to look at when a propeller comes back with a vibration complaint that dynamic balancing will not cure.

Finally, blades are controlled parts. Each carries an identification marking — manufacturer, part or drawing number, serial number and modification status — and each has its own life record, because blades accumulate hours and cycles individually and are often interchanged between hubs at overhaul. A blade may be life-limited, it may be subject to a service bulletin or an airworthiness directive calling for a repetitive inspection or a terminating modification, and the propeller as an assembly will carry a calendar and hours-based overhaul interval. None of those figures can be generalised: the applicable life, interval and inspection standard is whatever the propeller's own manual, the airworthiness directives applicable to it and the operator's approved maintenance programme say it is.

Angle of Attack

The **angle of attack (α)** of a propeller blade element is the angle between the element's chord line and the **relative airflow (RAF)** — the resultant of the rotational velocity and the aircraft's forward velocity. This is exactly the same concept as the angle of attack of a wing, applied to each small section of the blade.

The angle of attack determines:

- How much **lift** (thrust) the blade element produces
- How much **drag** (torque resistance) it creates

- Whether the element operates **efficiently** or is stalled

As forward speed increases (at constant RPM and blade angle), the helix angle φ increases, which **reduces** the angle of attack (since β = φ + α and β is fixed). Less lift is produced per element → the propeller becomes less efficient. Conversely, as forward speed decreases, α increases and could lead to blade element stall at very low speeds.

Key insight: This is precisely why variable-pitch propellers exist — by adjusting β (blade angle) to match the changing flight conditions, the propeller can maintain an optimal angle of attack across a wide range of airspeeds. A fixed-pitch propeller can only be optimised for ONE speed.

What Counts as a Good Angle of Attack

A propeller section is not trying to produce the most lift it can. It is trying to produce lift as cheaply as possible, because every unit of drag on the section is torque the engine has to pay for and thrust the aircraft does not get. The figure of merit is therefore the section's **lift-to-drag ratio**, and that peaks at a surprisingly small angle.

For the cambered aerofoil sections used on propeller blades, best lift-to-drag falls in the region of **2° to 4°** angle of attack. Below that, lift falls away faster than drag does and the section is under-worked. Above it, the drag rise begins to outpace the lift gain, and by the time the angle reaches double figures the section is heading for the stall. An angle of the order of 15° is not a working propeller angle at all: a section held there has separated, and it costs the engine a great deal of torque while returning very little thrust.

Why the fixed-pitch designer aims at 2° to 4°

A fixed-pitch propeller can put its blade at the best angle of attack in exactly **one** combination of airspeed and RPM. The designer therefore picks the condition the aircraft will spend most of its useful life in, works out the helix angle at every station for that condition, and adds the optimum angle of attack of a few degrees to produce the blade angle distribution. Everywhere else in the flight envelope the propeller is off-design, and the further the aircraft strays from the design point the worse the angle of attack becomes: too large at low speed and high power, too small at high speed. Recognising that a small positive angle of a few degrees is the target is what makes the whole design logic of a fixed-pitch propeller fall into place.

The Four Regimes a Blade Section Can Be In

The angle of attack of a propeller section is not confined to small positive values. Across the whole of a flight it can pass through four quite distinct regimes, and each has its own characteristic thrust, drag and handling

consequence:

Regime	Angle of attack	Thrust	Who drives whom	When it happens
Normal powered	Small positive, ideally 2°–4°	Positive (forward)	Engine drives propeller	All normal flight
Stalled	Large positive	Reduced, with very high torque	Engine drives propeller, badly	Static run-up, very low speed at high power, inner blade stations
Zero thrust	Very small; the value at which the axial force passes through zero	Zero	Neither — the balance point	Passing through flat pitch; the boundary between drive and windmill
Windmilling	Small negative	Negative (a rearward drag force)	Airflow drives propeller	Engine failure or shutdown with the blades unfeathered

Notice how the propeller passes through the third row on its way from the first to the fourth. There is a single blade angle at any given airspeed and RPM at which the section produces no net axial force at all; that is the flat, zero-thrust condition, and it is the dividing line between a propeller that is being driven and one that is doing the driving.

Blade Stall

A propeller section stalls for exactly the reason a wing does: the angle of attack has become too large for the flow to stay attached to the upper surface. What differs is when it happens and where on the blade.

The angle of attack at a station is the blade angle minus the helix angle, and the helix angle shrinks towards zero as forward speed goes to zero. The worst case is therefore a **static, full-power run**: the aircraft is not moving, so the helix angle at every station is essentially zero, and the angle of attack at every station is essentially the whole blade angle. On a fixed-pitch propeller optimised for the cruise, and on a constant-speed propeller sitting on a coarse blade angle when the power is applied, large parts of the blade — usually the inner stations, where the blade angle is highest — are stalled.

The consequences are all recognisable on the line:

- **Static thrust is disproportionately poor** compared with what the same propeller produces once the aircraft is rolling and the helix angle has grown.
- **Torque is high**, because a stalled section has very high drag, so the engine is loaded heavily without getting thrust in return. On a fixed-pitch installation this shows as an inability to reach the expected static RPM.

- **The propeller is noisy and rough**, because separated flow is unsteady and the separation is not identical on every blade.

On a constant-speed propeller the cure is automatic: the governor senses the low RPM, drives the blade angle finer, and the angle of attack falls back into the working range. That is exactly why a constant-speed propeller sits on its fine-pitch stop for take-off and why the fine stop's setting determines the static RPM you will see on a ground run.

Negative Angle of Attack: Windmilling

Windmilling is the case where the sign flips. If the engine stops driving the propeller — a failure, a deliberate shutdown, a fuel interruption — the propeller decelerates. As the RPM falls, the rotational velocity at each station falls with it, so the helix angle grows. As soon as the helix angle exceeds the blade angle, the angle of attack passes through zero and becomes **negative**: the airflow now meets the forward-facing surface of the blade instead of the rearward-facing one.

With a negative angle of attack the aerodynamic force on the section reverses. Two things follow at once, and both are examined:

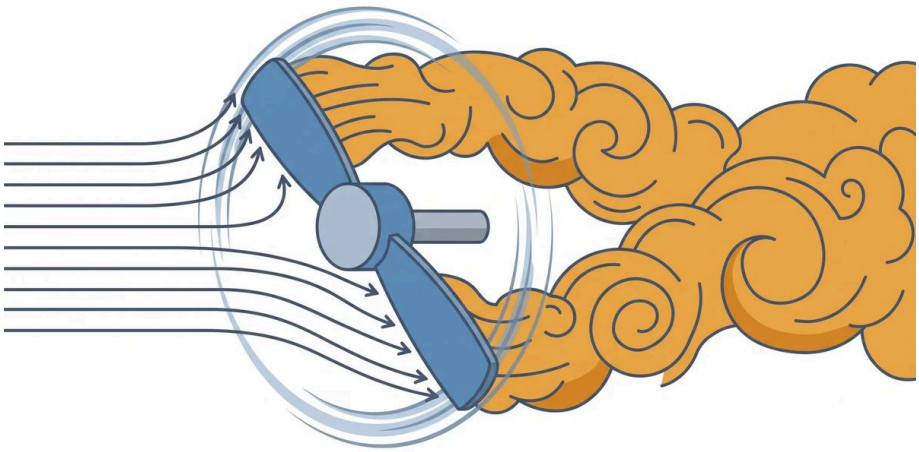
- **The axial component now acts rearwards.** Thrust is negative — the propeller is producing drag, not thrust. A windmilling propeller does not merely stop helping; it actively holds the aircraft back.
- **The in-plane component now drives the rotation instead of opposing it.** That is what keeps the propeller turning with no engine power at all: the airstream is doing work on the blades, and that work is absorbed by the engine's internal friction, its compression and the accessories it is still turning.

So the answer to the classic question is that a windmilling propeller has a small *negative* angle of attack and a *negative* thrust. Both signs are negative, and they are negative for the same single reason.

⚠ Windmilling drag is worst at fine pitch

The drag a windmilling propeller produces depends strongly on the blade angle it is sitting at. At a **fine** blade angle the blades present their broad faces almost square-on to the oncoming airflow and work at a large negative angle of attack, so the disc behaves like a rotating flat plate and the drag is at its maximum. At a **coarse** angle the blades are more nearly aligned with the flow and the drag is much less. **Feathered**, the blade is edge-on to the airflow, the drag is at its minimum and the rotation stops altogether.

This is why the failure case that matters is a propeller windmilling at fine pitch, why an unfeatherable propeller on a twin is such a serious performance problem, and why the drill after an engine failure is to feather without delay. The maximum-drag case and the minimum-drag case are the two ends of the same blade angle range.



At fine pitch a windmilling propeller meets the airflow almost square-on, like a rotating flat plate.

What Sets the Windmilling RPM

A windmilling propeller settles at a definite speed rather than accelerating without limit, and it is worth working out why, because the answer explains which airspeed it depends on.

Equilibrium is reached when the driving moment the airflow applies to the blades exactly balances the resisting moment of the engine and accessories. The driving moment depends on the angle of attack, and the angle of attack at a fixed blade angle depends only on the helix angle. The helix angle in turn depends on the ratio of the forward speed to the rotational speed. So equilibrium corresponds to one particular *ratio* of forward speed to rotational speed — and therefore, for a given propeller at a given blade angle, to one particular value of the advance ratio.

Why windmilling RPM follows TAS

At the windmilling equilibrium the advance ratio is fixed by the blade angle, so:

$$J = \frac{V}{nD} = \text{constant} \quad \Rightarrow \quad n \propto V$$

Double the speed of the air past the blades and you double the windmilling RPM. The speed that appears in that expression is the **real speed of the aircraft through the air mass** — the true airspeed. Indicated airspeed is a pressure measurement that has not been corrected for density, and equivalent airspeed corrects only for compressibility, so neither of them is the velocity the blade actually sees. At altitude, where the true airspeed is well above the indicated airspeed, a windmilling propeller therefore turns faster than the indicated reading would suggest.

There is a practical consequence. A windmilling propeller that cannot be feathered is turning an engine that may have no oil pressure, and the RPM it turns at is governed by how fast the aircraft is flying. Reducing

airspeed is therefore the only control the crew has over the windmilling speed of an unfeatherable propeller, and it is why flight manuals for such a failure specify a speed to fly. Conversely, the airflow can be used deliberately: a **windmill start** uses forward speed to spin the engine up to a self-sustaining speed after an in-flight shutdown, and the manual will quote a minimum airspeed and an altitude band for it precisely because the windmilling RPM depends on true airspeed.

Angle of Attack Is Not the Same All the Way Round the Disc

Everything so far has assumed the airflow arrives squarely down the shaft axis. It usually does not. Whenever the aircraft is flying at an angle of attack of its own — on the climb-out, in a slow approach, in a turn — the propeller disc is inclined to the airflow, and each blade's local angle of attack then varies as it goes round.

Consider a nose-up attitude. The oncoming air no longer arrives along the shaft axis; part of it lies in the plane of the disc. On the side where the blade is sweeping **downwards**, that in-plane component adds to the blade's own tangential motion, so the resultant tangential velocity is higher, the helix angle is smaller and the angle of attack is therefore **larger**. On the opposite side, where the blade is sweeping **upwards**, the in-plane component subtracts, the helix angle is larger and the angle of attack is **smaller**. The descending blade consequently produces more thrust than the ascending one. This is the asymmetric blade effect, usually called **P-factor**, and it has two quite different consequences:

- **An out-of-balance thrust** across the disc, which acts through a point offset from the shaft centreline and therefore yaws the aircraft. This is one of the reasons a single-engine aeroplane needs rudder as it rotates and climbs away.
- **A cyclic load on every blade**, which rises and falls once per revolution as the blade goes round. That is the mechanism behind the once-per-revolution vibratory force discussed later in this note, and it is a fatigue-loading case the blade and hub are specifically designed to survive.

The same effect occurs in yaw and sideslip, in the upwash ahead of a wing, and behind an obstruction such as a nacelle or a fuselage. A propeller is almost never operating in perfectly uniform inflow.

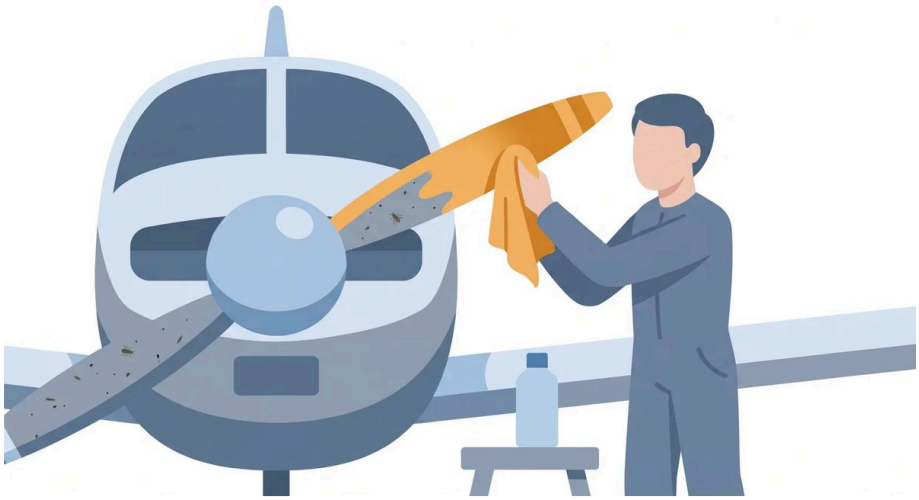
What Changes Angle of Attack in Service

From an engineer's point of view the angle of attack is not something you can adjust — but several things you *can* adjust or damage change it, and the symptoms are distinctive:

- **A mis-set blade angle** shifts the angle of attack at every station by the same amount. Too coarse and the engine cannot reach its RPM; too fine and it overspeeds or reaches its RPM at reduced power.
- **A blade set differently from its neighbours** gives that blade a different angle of attack and therefore a different thrust once per revolution, producing an aerodynamic imbalance that no amount of mass

balancing will remove.

- **Leading-edge erosion, deep dressing or a poor repair** changes the section's shape and therefore both the angle at which it produces best lift-to-drag and the angle at which it stalls. A blade with a badly eroded leading edge is not the aerofoil the designer drew.
- **Contamination** — insect debris, dried salt, thick or peeling paint, a rough or blistered de-icing boot — thickens the boundary layer, raises drag and reduces the lift the section produces at a given angle. Clean blades are a performance item, not a cosmetic one, and blades must be cleaned with the materials the propeller manual approves, since some cleaners attack the blade's protective finish or the de-icing boots and their adhesive.
- **A governor set to the wrong RPM** changes the rotational velocity at every station and therefore the helix angle, which changes the angle of attack even though nobody has touched the blades.



Contamination such as insect debris raises drag, so blades are cleaned with manual-approved materials.

How the Section Responds to Angle of Attack

A propeller section behaves like any other aerofoil, and its lift coefficient rises almost linearly with angle of attack up to the stall. Three features of that response drive propeller behaviour:

- **Lift is roughly proportional to angle of attack** over the working range, so a small change in blade angle or in helix angle produces a proportionate change in thrust. This is why a propeller responds so promptly to a pitch change.
- **Drag rises much faster than linearly.** Profile drag has a component that grows with the square of the lift coefficient, so doubling the angle of attack more than doubles the drag. That asymmetry is why the best lift-to-drag angle is small and why over-pitching is expensive.
- **The stall angle is not a fixed number.** It depends on the section shape, on its surface condition, and on the local Mach number: as the tip's Mach number rises, the angle at which the flow separates falls. A tip

section can therefore be closer to its stall than the same section at the root, even at a smaller angle of attack.

A propeller blade is also, unlike a wing, operating at a wide range of Reynolds numbers along its own span, because both the local velocity and the local chord change from root to tip. That is one more reason the aerofoil sections are varied along the blade rather than scaled.

Negative Angle of Attack Under Power: Reverse

Windmilling is not the only way to reach a negative angle of attack. Reverse pitch does it deliberately and with the engine driving hard, and comparing the two cases side by side removes a common confusion.

	Windmilling	Reverse pitch
Blade angle	Positive, but smaller than the helix angle	Negative — driven past the flat position
Angle of attack	Negative, usually small	Negative, and large
Thrust	Negative — a drag force	Negative — a braking force
Direction of power flow	The airflow drives the propeller, which turns the engine	The engine drives the propeller, absorbing power in the normal sense
When	Engine failure or shutdown, blades unfeathered	Deliberately selected on the ground, below the flight fine stop

Reading the table row by row shows why the two are so easily confused and why the confusion matters. The signs of the angle of attack and of the thrust are identical, so a question that asks only about those two cannot distinguish the cases. Everything else about them is different, and in particular the direction of power flow is opposite: one is a failure condition costing the aircraft performance, the other is a commanded condition doing useful work.

Rotational Speed

Rotational speed (measured in RPM — revolutions per minute, or rev/s) is how fast the propeller turns. It is directly controlled by the engine throttle (on fixed-pitch propellers) or by the propeller governor (on constant-speed propellers).

Rotational speed affects propeller performance in several critical ways:

Effect	Detail
Thrust	Higher RPM = more air accelerated = more thrust (up to a point)

Effect	Detail
Tip speed	Higher RPM increases tip speed. If tip speed approaches or exceeds Mach 1, compressibility effects cause severe efficiency loss, noise, and vibration.
Blade AoA	At constant forward speed and fixed blade angle: increasing RPM decreases the helix angle ϕ, which increases AoA α. Too high → blade stall.
Centrifugal loads	Centrifugal force on each blade increases with the square of RPM. Structural limits set the maximum permissible RPM.
Engine stress	Exceeding maximum RPM (overspeed) can cause catastrophic engine or gearbox failure.

Getting the Units Right Before Anything Else

Almost every arithmetic error in propeller work is a units error, and almost all of them are the same one: putting revolutions per minute into a formula that wants revolutions per second. Three quantities are in circulation and they must be kept apart.

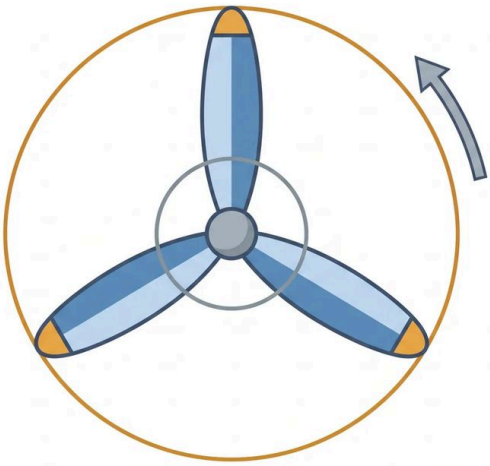
Three ways of stating the same speed

- **N** in revolutions per minute — what the cockpit gauge shows.
- **n** in revolutions per second: $n = N/60$. This is what the pitch, slip, advance-ratio and power formulae use.
- **ω** in radians per second: $\omega = 2\pi n = 2\pi N/60$. This is what the rotational-velocity and centrifugal-force formulae use.

For 1,200 RPM: $n = 20$ rev/s and $\omega = 2\pi \times 20 = 125.7$ rad/s. For 2,400 RPM: $n = 40$ rev/s and $\omega = 251.3$ rad/s.

Tip Speed Is the Real Limit

Rotational speed matters to a propeller designer almost entirely through what it does to the tip. The rotational velocity of the tip is the largest velocity anywhere on the aircraft, and the tip does not experience it alone: the tip's true path through the air is a helix, so its actual speed is the resultant of its rotational velocity and the aircraft's forward velocity.



In each turn the blade tip sweeps the largest circle, so it moves faster than anything else on the aircraft.

Helical tip speed and tip Mach number

$$V_{helical} = \sqrt{(2\pi Rn)^2 + V_{TAS}^2}$$

$$M_{tip} = \frac{V_{helical}}{a} \quad \text{where} \quad a = 20.05\ \sqrt{T}$$

with the local air temperature T in kelvin and the speed of sound a in m/s. The forward-speed term is squared and added, so at cruising speeds it makes a real difference to the tip's Mach number even though it is much smaller than the rotational term.

The same propeller, two conditions

Take the 2.5 m propeller at 1,200 RPM. Its tip rotational velocity is $2\pi \times 1.25 \times 20 = 157.1$ m/s.

- **Cruising at 100 m/s TAS:** $V_{helical} = \sqrt{157.1^2 + 100^2} = \sqrt{24675 + 10000} = 186.2$ m/s.
- **At sea level on a standard day**, $T = 288$ K so $a = 340$ m/s and the tip Mach number is $186.2/340 = 0.55$.
- **At altitude where the outside air is –50 °C**, $T = 223$ K so $a = 299.5$ m/s, and the very same tip speed now gives $186.2/299.5 = 0.62$.

Nothing about the propeller changed. The air got colder, the speed of sound fell, and the tip Mach number rose by 0.07 on its own. This is why a propeller's critical condition is normally a cold, high-altitude, high-speed cruise rather than a warm day at sea level, and why propeller RPM is often reduced in the cruise on types that operate high.

The formula-box example earlier in this note showed the other end of the same problem: a 2 m propeller at 2,400 RPM has a tip rotational velocity of 251 m/s, which at sea level is already Mach 0.74 *before* any forward

speed is added. Add 60 m/s of forward speed and the helical speed becomes 258 m/s, or Mach 0.76. A small, fast-turning propeller runs out of tip-speed margin very quickly indeed.

What Actually Goes Wrong at High Tip Mach Number

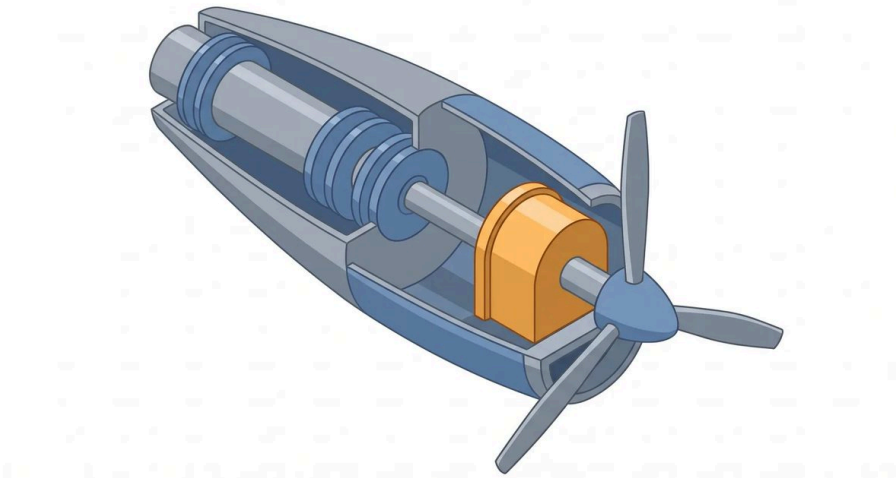
As the tip's local Mach number rises towards unity, the flow accelerating over the curved upper surface of the section reaches sonic speed before the free stream does, and a shock wave forms on the blade. Four things then happen together:

- **Wave drag appears** and rises steeply. The section's drag can multiply several times over a small increase in Mach number, and every bit of that drag is extra torque the engine must supply.
- **Lift falls away** as the shock thickens the boundary layer behind it and eventually separates the flow. The tip stops contributing thrust just when it should be contributing most.
- **Noise rises sharply.** Propeller noise is dominated by tip speed, and the increase near and beyond the critical Mach number is dramatic. This is a certification issue as much as a comfort one, and it is one of the principal reasons modern turboprops turn their propellers relatively slowly.
- **Vibration and flutter risk increase**, because the shock position on the blade is unsteady and moves in response to small changes in angle of attack, feeding an oscillating load into a structure that is already highly stressed.

The designer's defences are all visible on a modern blade: thin, high-speed sections at the tip, sweep on the outer blade, careful control of the maximum diameter, and above all a low rotational speed.

Reduction Gearing: How a Fast Engine Drives a Slow Propeller

Engines and propellers want opposite things. A gas turbine is efficient and compact only when it turns very fast; a propeller is efficient and quiet only when it turns slowly enough to keep its tips subsonic. A reduction gearbox resolves the conflict by letting each run at the speed that suits it.



A reduction gearbox sits between the fast-turning gas turbine and the slow-turning propeller.

The gearbox does two things at once

A power turbine turning at 30,000 RPM drives a propeller at 1,200 RPM through a reduction gearbox. The gear ratio is $30000/1200 = 25:1$.

Consider 1,500 kW passing through it. At the turbine, $n = 500$ rev/s, so $Q = P/2\pi n = 1{,}500{,}000/(2\pi \times 500) = 478$ N·m. At the propeller shaft, $n = 20$ rev/s, so $Q = 1{,}500{,}000/(2\pi \times 20) = 11{,}937$ N·m.

The speed has been divided by 25 and the torque multiplied by 25 — the same power, redistributed. The tip that would have been hopelessly supersonic at 30,000 RPM is comfortably subsonic at 1,200 RPM, and the price is a gearbox that has to carry nearly twelve thousand newton metres.

That price is worth naming, because it is the reason the reduction gearbox is a major maintenance item in its own right. It carries the propeller's full torque, its thrust, and its gyroscopic and out-of-balance loads; it is monitored by magnetic chip detectors and by oil analysis, and the condition of its oil is one of the primary indicators of the health of the whole propeller installation. Piston engines face the same trade in miniature: a directly driven propeller is limited to the engine's crankshaft speed, so a high-revving engine either accepts a small-diameter propeller or is geared down.

Rotational Speed and Structural Load

The centrifugal force on a blade is proportional to the **square** of the angular velocity, which means the structural cost of rotational speed rises far faster than the speed itself. It is worth doing the arithmetic once to feel the size of it.

✣ One small piece of one blade

Take a single element of blade weighing just 0.5 kg, sitting 1.0 m out from the axis.

- At **1,200 RPM**: $\omega = 125.7$ rad/s, so $F_c = 0.5 \times 125.7^2 \times 1.0 = 7{,}896$ N — about 0.8 tonnes-force.
- At **2,400 RPM**: $\omega = 251.3$ rad/s, so $F_c = 0.5 \times 251.3^2 \times 1.0 = 31{,}583$ N — about 3.2 tonnes-force.

Doubling the RPM multiplied the load by four, exactly as the square law says it must. And that is half a kilogram. Summing the same calculation over every element of a full-size turboprop blade is how the whole-blade loads discussed later in this note are arrived at.

Because the relationship is a square law, a modest overspeed is not a modest overload. An engine and propeller taken 20% above their maximum permitted RPM see roughly 44% more centrifugal load, and the blade retention, hub and shaft are sized for the certified limit with a defined margin, not for an open-ended one. This is why a propeller overspeed is a reportable event with a mandatory inspection attached, why the inspection required depends on how far and for how long the limit was exceeded, and why the applicable action must be taken from the propeller and engine manuals rather than assumed.

How Rotational Speed Is Actually Controlled

On a **fixed-pitch** propeller nothing controls RPM directly. The propeller settles wherever the engine's torque output and the propeller's torque demand happen to balance, so the RPM drifts with flight condition. Open the throttle and RPM rises. Increase airspeed at a constant throttle setting and the helix angle grows, the angle of attack at every station falls, the propeller's torque demand drops, and the RPM **rises** — which is why a fixed-pitch aeroplane gains RPM in a dive and loses it in a climb without the pilot touching anything.

On a **constant-speed** propeller the RPM is held by a governor that uses blade angle as its control. The logic is a single loop, and getting its direction right is worth more than memorising any diagram:

Condition sensed	Governor action on blade angle	Effect on propeller torque demand	Resulting RPM
Overspeed — RPM above the selected value	Increase blade angle (coarsen)	Increases — the blade takes a bigger bite and loads the engine more	Falls back to the selected value
Underspeed — RPM below the selected value	Decrease blade angle (fine off)	Decreases — less load on the engine	Rises back to the selected value
On speed	Holds the current angle	Matched to engine torque output	Steady

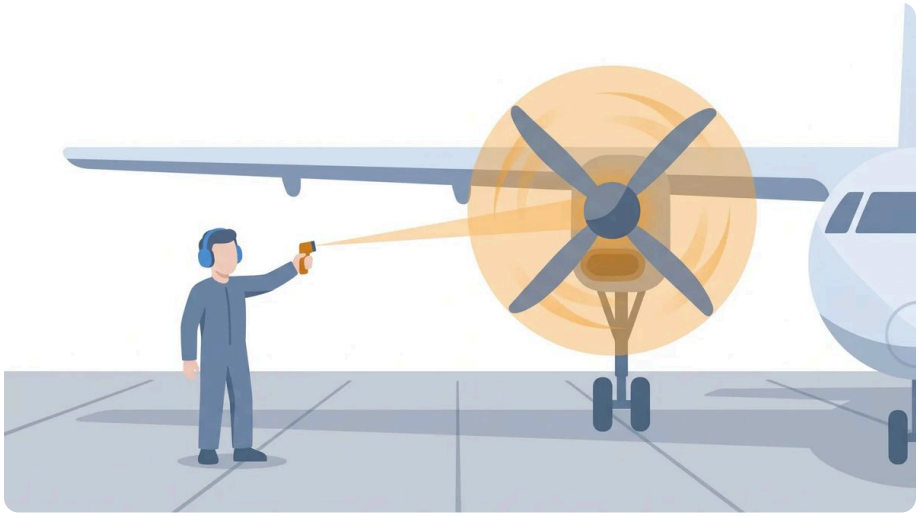
The direction is worth checking against a concrete case, because it is easy to invert. Suppose the aircraft enters a descent: the airspeed rises, the helix angle grows, the angle of attack and hence the propeller's torque demand fall, and the propeller starts to accelerate. The governor senses that overspeed and **coarsens**

the blade, restoring the load and pulling the RPM back to the selected figure. The engine's power setting has not been touched; the propeller has simply changed gear.

One consequence of this is worth carrying into the hangar: on a constant-speed installation the RPM gauge tells you very little about power. The propeller holds the selected RPM across a wide range of power settings, so power has to be read from the manifold pressure on a piston engine or the torque indication on a turboprop. A constant-speed propeller that *does* allow the RPM to wander is telling you something is wrong — a governor fault, low or aerated oil pressure to the pitch-change mechanism, a sticking pitch-change piston, or blade angles that are not what they should be.

Measuring and Verifying Rotational Speed

Because so many limits and checks are expressed in RPM, the accuracy of the indication itself matters. Propeller speed is normally sensed by a magnetic or variable-reluctance pick-up counting the passage of gear teeth or a phonic wheel, or on older installations by a mechanical tachometer generator driven from the gearbox. The indicated figure can be verified independently with a **strobe or optical tachometer** during a ground run, and that check is part of the standard investigation whenever reported RPM behaviour does not match what the propeller is physically doing.



During a ground run, a strobe or optical tachometer independently checks the indicated propeller RPM.

Off the aircraft, a propeller or its governor may be run on a **propeller test stand** or whirl rig, which drives the unit at controlled speeds so that governing, pitch-change rates, feathering and unfeathering times, low-pitch stop behaviour and leakage can be verified against the overhaul manual's acceptance figures before the propeller is released. The stand is also where a newly overhauled or repaired propeller is dynamically balanced. All the acceptance figures for such a run — speeds, times, pressures, permitted leakage — come from the propeller's own manual and are specific to the type and part number.

Finally, remember that the whole permitted RPM range may not be continuously usable. Where a propeller and its installation have a resonance inside the operating range, the flight manual placards a restricted band that must be transited rather than dwelt in; the reasoning behind those placards is dealt with later in this note.

Two Turboprop Architectures, Two Meanings of "RPM"

On a turboprop there is more than one rotating speed, and which of them the propeller is tied to depends on how the engine is built. The distinction changes the whole character of the installation's handling and its ground-running procedure.

	Fixed-shaft (direct-drive) turboprop	Free-turbine turboprop
Arrangement	The propeller is geared to the same shaft as the compressor and its turbine	A separate power turbine, mechanically independent of the gas generator, drives the propeller through the gearbox
Consequence for RPM	Propeller speed and gas generator speed are locked together by the gear ratio	Propeller speed and gas generator speed are independent; the two are linked only by the gas flow between them
Starting	The starter must turn the propeller as well as the compressor, so the blades are set to a fine, low-drag angle for starting	The starter turns only the gas generator; the propeller does not move until there is gas flow to the power turbine
Power control	RPM is normally held near constant and power is varied with fuel flow and blade angle together	Propeller RPM is set by the governor; power is set by fuel flow and read as torque

The free-turbine arrangement is the one most commonly met, and its practical consequence on the ground is that the propeller can be held stationary or at low speed while the gas generator runs, which makes ground handling and starting much simpler. It also means the propeller RPM indication tells you nothing about the gas generator's condition, and vice versa — both have to be read, along with the torque and the turbine temperature, before any judgement about the installation can be made.

Protecting Against Overspeed

Because centrifugal load follows a square law, and because a fine blade angle at speed can drive an enormous acceleration, an overspeed is one of the few propeller failures that can become catastrophic within seconds. Installations therefore carry more than one line of defence, and it is worth knowing them as a layered system rather than as a list:

- **The constant-speed governor itself** is the first defence. In normal operation it coarsens the blade whenever the RPM rises above the selected value, and that action alone handles every ordinary disturbance.

- **An overspeed governor** is a second, independent unit set to a higher speed than any normal selection. It does nothing at all until the propeller exceeds that speed, at which point it dumps oil from the pitch-change mechanism to drive the blade coarse regardless of what the main governor is doing.
- **Fuel topping or power limiting** attacks the problem from the engine side: on detecting an overspeed the fuel control reduces fuel flow, cutting the power going into the propeller.
- **Autofeather and negative-torque sensing** deal with the related case in which the propeller starts driving the engine instead of the other way round. Negative torque sensing detects that reversal of power flow and commands the blade towards coarse or feather, which removes the drag before the crew have had time to identify the failure.

All of these are functionally checked at the intervals the maintenance programme specifies, because every one of them is a system that does nothing whatsoever until the day it is needed. A latent failure in an overspeed governor is invisible in normal operation by design.

Rotational Speed and Noise

The steep growth of noise with tip Mach number, listed above among the consequences of high tip speed, has shaped modern propeller design as strongly as efficiency has — because noise is a certification requirement and, at many airports, an operating restriction with a financial consequence attached.

The design response is the same low-tip-speed philosophy already described: more blades, wider blades, a modest diameter, swept and thinned tips, and a low rotational speed made possible by reduction gearing. The operational response is a reduced propeller RPM for cruise and for noise-sensitive procedures, which is available on any constant-speed installation simply by selecting a lower governed speed. From a maintenance standpoint the relevant point is that the RPM limits and the governed speeds in the flight manual are part of the aircraft's certification basis, including its noise certification, and adjusting a governor outside its specified setting is not a minor rigging matter.

Propeller Slip

Propeller slip is the difference between the *geometric pitch distance* (the theoretical distance the propeller would advance in one revolution if it were screwing through a solid) and the *effective pitch distance* (the actual distance it advances through the air).

Slip:

$$\text{Slip} = \text{Geometric pitch} - \text{Effective pitch}$$

$$\text{Slip (\%)} = \frac{\text{Geometric pitch} - \text{Effective pitch}}{\text{Geometric pitch}} \times 100$$

Slip occurs because air is a **fluid, not a solid**. A screw in wood advances its full pitch distance each revolution because wood is rigid. Air, however, yields and accelerates — some of the propeller's energy goes into accelerating the air rearward rather than advancing the aircraft forward. Typical propeller slip ranges from **10% to 30%** depending on conditions.

Example: A propeller has a geometric pitch of 2.0 m. At a given airspeed and RPM, the aircraft advances 1.5 m per revolution.

$$\text{Slip} = 2.0 - 1.5 = 0.5 \text{ m}$$

$$\text{Slip } \% = \frac{0.5}{2.0} \times 100 = 25\%$$

Slip is **not wasted energy** — it is the mechanism by which thrust is produced. Without slip, there would be no acceleration of air and no thrust (just as a wing at zero angle of attack produces no lift). However, excessive slip indicates low propeller efficiency.

Getting the Two Pitches Straight

Slip is a difference between two distances, so the definitions of both distances have to be exact before the difference means anything.

Quantity	Definition	Depends on	How it is found
Geometric pitch	The pitch of the helical surface the blade sweeps — its advance per revolution with no slip at all	Blade angle and radius only. Nothing aerodynamic.	Calculated: $2\pi r \tan \beta$ at the reference station
Effective pitch	The distance the aircraft actually advances in one revolution	True airspeed and rotational speed	Measured or calculated: V/n
Slip	The shortfall of the second against the first	Both of the above	Subtraction, expressed as a distance or a percentage

Two exam traps live in that table. The first is the wording of geometric pitch: it is the advance **without** slip, not the advance with it, and not simply "the distance moved in one revolution" — that phrase belongs to effective pitch. The second is the sign: effective pitch is geometric pitch **minus** slip, never plus. If a distractor offers "geometric pitch plus slip", it has the subtraction the wrong way round, and the giveaway is that the answer it produces would have the aircraft travelling further than the theoretical maximum.

Effective pitch from airspeed and RPM

$$EP = \frac{V}{n}$$

with V the true airspeed in metres per second and n the rotational speed in **revolutions per second**. Using RPM here without dividing by 60 gives an answer sixty times too small, and it is the single most common slip in a slip calculation.

Worked Examples: How Slip Moves With Airspeed

One propeller, one blade angle, two speeds

A propeller has a geometric pitch of 5.6 m at its cruise blade angle, and is turning at 1,200 RPM, so $n = 20$ rev/s. The blade angle and RPM are held constant while the aircraft slows down.

- **At 100 m/s TAS:** $EP = 100/20 = 5.0$ m. Slip $= 5.6 - 5.0 = 0.6$ m, which is $(0.6/5.6) \times 100 = 10.7\%$.
- **At 70 m/s TAS:** $EP = 70/20 = 3.5$ m. Slip $= 5.6 - 3.5 = 2.1$ m, which is $(2.1/5.6) \times 100 = 37.5\%$.

Nothing was done to the propeller, and the slip more than tripled. At a fixed blade angle and a fixed RPM, **slowing down increases slip and speeding up reduces it** — which is the same statement, in different units, as the earlier result that slowing down increases the angle of attack.

That equivalence is worth holding onto, because slip and angle of attack are two descriptions of one thing. Large slip means the propeller is advancing much less than its geometry would allow, which means the airflow is arriving at a large angle to the chord, which is a large angle of attack. Small slip means the propeller is nearly keeping up with its own geometry, so the airflow is nearly along the chord and the angle of attack is small. Push that to the limit and the blade reaches the condition where the angle of attack goes negative — the propeller advancing *more* than its geometric pitch — and that is precisely the windmilling condition described earlier.

Slip and Advance Ratio Are the Same Measurement

The non-dimensional parameter that propeller performance is normally plotted against, the advance ratio, is built out of exactly the same numbers:

Advance ratio is effective pitch expressed in diameters

$$J = \frac{V}{nD} = \frac{EP}{D}$$

Using the cruise case above: $J = 100/(20 \times 2.5) = 2.0$, and equally $EP/D = 5.0/2.5 = 2.0$. At 70 m/s: $J = 70/50 = 1.4$, and $3.5/2.5 = 1.4$. The advance ratio simply states how many propeller diameters the aircraft travels for each revolution.

This is why the same physical situation can be described three ways in three different questions — as a slip percentage, as an advance ratio, or as an angle of attack — and why an engineer who can move between them is not really learning three topics.

Why Slip Has to Exist: The Momentum Argument

The reason air yields where wood does not is worth stating in the language of Newton's laws rather than by analogy, because it leads directly to several examinable results.

A propeller produces thrust by accelerating air rearwards. Thrust is the rate of change of momentum of that air, so it is the mass flow through the disc multiplied by the increase in velocity the disc gives it. If the air were not accelerated, its momentum would not change and there would be no thrust at all. Two conclusions follow immediately:

- **The slipstream must be moving faster than the surrounding air.** The propeller has added energy and momentum to the air passing through it, so the flow leaving the disc is faster than the ambient air it is travelling through. It is also narrower — a faster flow carrying the same mass occupies a smaller cross-section, so the slipstream **contracts** behind the disc.
- **The air's rearward acceleration is the slip.** The extra distance the air moves backwards is exactly the distance the propeller fails to advance. Slip is not a separate loss sitting alongside thrust production; it is the visible face of thrust production.

Ideal propulsive efficiency, and why a big disc wins

If the disc adds a velocity v to the air passing through it — the induced velocity — and the aircraft is flying at V, then the ideal propulsive efficiency, ignoring all blade drag, is:

$$\eta_i = \frac{V}{V + v}$$

The message is unambiguous. To make efficiency high, v must be small compared with V — that is, the air must be given only a gentle push. But thrust needs momentum change, so if each kilogram of air is only pushed gently, a great many kilograms per second must be handled. That is the whole argument for a **large mass of air accelerated to a low velocity** rather than a small mass accelerated to a high velocity.

Halving the disc for the same thrust

A 2.5 m propeller has a disc area of $\pi \times 1.25^2 = 4.91$ m^2. Flying at 100 m/s in air of density 1.225 kg/m^3 and giving the flow an induced velocity of 10 m/s, momentum theory gives a thrust of about 13,200 N, and an ideal efficiency of $100/(100 + 10) = 90.9\%$.

Now demand the same 13,200 N from half the disc area, 2.45 m^2. Less air passes through, so each kilogram of it must be pushed harder: the induced velocity rises to about 18.6 m/s. The ideal efficiency falls to $100/(100 + 18.6) = 84.3\%$.

Six and a half points of efficiency lost, with no change in thrust and no change in blade quality — purely because the same job was done with a smaller disc and a harder shove. This is the fundamental reason a propeller beats a jet at low speed, and the reason propeller diameter is made as large as ground clearance and tip speed will permit.

The Slipstream Is Not Only Faster — It Also Rotates

The blades apply a torque to the air as well as an axial force, so the air leaves the disc with a swirl in the direction of propeller rotation. That rotational kinetic energy is put into the air by the engine and never returns as thrust, so it is a genuine loss. It grows with the torque being applied and shrinks as the aircraft speeds up, so it bites hardest during take-off and the climb — one of the reasons static and low-speed efficiency is poor.

The swirl has effects an engineer meets directly. It wraps around the fuselage as a spiralling flow and strikes one side of the fin, producing a yawing moment that changes with power setting. It disturbs the airflow over the parts of the wing behind the disc. And it is the reason contra-rotating propellers exist: a second propeller turning the opposite way on the same axis recovers the swirl energy the first one put in, straightening the flow and improving efficiency, at the cost of considerable mechanical complexity.

Slip in the Hangar

Slip is not something an engineer measures directly, but it is behind several checks that are done routinely:

- **The static RPM check.** With the aircraft stationary, effective pitch is zero and slip is total; the propeller's torque demand is then set almost entirely by its blade angle. That makes static RPM on a fixed-pitch installation, and static RPM against the fine-pitch stop on a constant-speed one, a direct and sensitive test of whether the blade angle is right. An engine that will not reach its static RPM figure is being over-pitched; one that exceeds it is under-pitched.
- **The cruise performance complaint.** A crew report of a shortfall in cruise speed or an inability to reach cruise RPM, with the engine otherwise healthy, points at blade angle or at blade surface condition, both of which change the relationship between geometric and effective pitch.
- **Reading the propeller's own markings.** Light-aircraft fixed-pitch propellers are commonly identified by a diameter-and-pitch designation, both in inches, stamped or etched on the hub with the manufacturer's part and serial number. The pitch figure in that designation is the nominal geometric pitch at the propeller's reference station, so two propellers of the same diameter but different pitch figures are the climb and cruise variants of the same design. Always confirm the propeller fitted against the aircraft's approved type data before assuming which one is on the aeroplane.

The Same Calculation in Imperial Units

Light-aircraft propellers are almost always described in inches, and slip questions are often posed that way, so the conversion is worth doing once in full. The trap is that the units of the two pitches must match before they can be subtracted, and airspeed in knots is a distance per hour while RPM is a count per minute.

A 74 by 58 propeller at 100 kt

The designation means a diameter of 74 in and a nominal geometric pitch of 58 in at the reference station. The aircraft is doing 100 kt true and the engine is turning 2,400 RPM.

- Convert the speed to feet per minute: $100 \times 6076 = 607{,}600$ ft/h, and $607{,}600 \div 60 = 10{,}127$ ft/min.
- Effective pitch is the distance per revolution: $10{,}127 \div 2400 = 4.22$ ft, which is $4.22 \times 12 = 50.6$ in.
- Slip is $58 - 50.6 = 7.4$ in, or $(7.4/58) \times 100 = 12.7\%$.

Note that the diameter played no part in the calculation. Diameter enters the advance ratio, but slip depends only on the two pitches, and the geometric one comes from the blade angle at the reference station.

Changing RPM Instead of Speed

The worked examples above held the RPM and varied the airspeed. The other half of the relationship is just as important and is the one the pilot actually commands with the propeller lever.

Effective pitch is airspeed divided by rotational speed, so at a constant airspeed, **raising the RPM lowers the effective pitch** — the aircraft covers the same ground while the propeller makes more revolutions doing it, so each revolution accounts for less distance. Take the same 5.6 m geometric pitch at 100 m/s. At 1,200 RPM the effective pitch is 5.0 m and the slip is 10.7%. At 1,600 RPM the propeller turns at 26.7 rev/s, the effective pitch falls to $100/26.7 = 3.75$ m, and the slip climbs to $(5.6 - 3.75)/5.6 = 33\%$.

In practice a constant-speed propeller does not behave that way for long, because coarsening or fining the blade also changes the geometric pitch. Selecting a higher RPM at a constant airspeed causes the governor to fine the blade off, which reduces the geometric pitch at the same time as the effective pitch is falling. Both numbers move, and only the calculation tells you which way the slip ends up. What is unambiguous is the underlying relationship: **slip rises whenever the propeller is turning fast relative to the aircraft's progress, and falls whenever the aircraft is making good progress for each revolution.**

Slip Cannot Go Below Zero While the Propeller Is Pulling

Because slip is the shortfall of actual advance against theoretical advance, zero slip would mean the propeller advancing the full geometric pitch every revolution. At that point the air is arriving along the chord line at the reference station, no air is being accelerated rearward, and the propeller has stopped producing useful thrust. A propeller that is pulling therefore always has positive slip, and the figure quoted in performance work — commonly in the region of 10% to 30%, as noted above — reflects how heavily the propeller is loaded at that condition rather than how well or badly it is built.

Push past that point and the propeller advances *further* than its geometric pitch each revolution. The angle of attack goes negative, the thrust goes negative, and the propeller is windmilling — the condition described earlier in this note. Read as a slip figure, windmilling is the negative-slip region, which is a useful way to remember that slip, angle of attack and thrust all change sign together.

The Actuator Disc: What the Air Does as It Passes Through

Momentum theory treats the propeller as an infinitely thin disc that adds energy to the air passing through it, and following one streamline through that disc explains several observations that are otherwise just facts to memorise.

- **Ahead of the disc** the air is already accelerating, drawn in by the low pressure just ahead of the blades. Its static pressure falls as it speeds up, and by the time it reaches the disc it has picked up **half** of the total

velocity increase it will eventually receive.

- **Across the disc** the velocity is continuous — the air cannot change speed instantaneously — but the static pressure jumps upward. That pressure jump, multiplied by the disc area, **is** the thrust.
- **Behind the disc** the air expands back down to ambient static pressure, and in doing so it picks up the second half of its velocity increase. The final increment in the far wake is therefore twice the increment at the disc itself.
- **The streamtube contracts** throughout, because the same mass flow is travelling steadily faster.

How big is that pressure jump?

The 2.5 m propeller producing 13,200 N through a disc area of 4.91 m^2 creates a static pressure rise of $13{,}200 / 4.91 = 2{,}690$ Pa across the disc. Sea-level static pressure is about 101,300 Pa, so the propeller is raising the pressure by under three per cent.

That is the whole of it. A propeller is a very low pressure-ratio machine working on an enormous mass flow, which is exactly the opposite of the compressor a few feet behind it. It also explains the arithmetic identity that thrust divided by disc area — the **disc loading** — is numerically the same thing as the pressure jump. Disc loading is the parameter that decides how hard the air is being worked, and it is why two propellers of the same diameter absorbing very different powers feel so different at low speed.

Forces Acting on a Propeller Blade

Every rotating propeller blade is subject to several simultaneous forces. Understanding these is critical for appreciating blade design, pitch control, and structural requirements.

1. Centrifugal Force (Centrifugal Twisting Moment — CTM)

Centrifugal force acts on every particle of the blade, pulling it radially outward from the axis of rotation. This is the dominant structural load — it tries to pull the blade out of the hub. On a large turboprop at full RPM, the centrifugal load on each blade can exceed **20 tonnes**.

Centrifugal force also creates the **Centrifugal Twisting Moment (CTM)**. Because the centre of mass of each blade element does not coincide with the pitch-change axis, centrifugal force creates a torque that tends to twist the blade toward **fine pitch** (low blade angle). This is because centrifugal force tries to align all the mass in the plane of rotation, which means reducing the blade angle.

Centrifugal force on a blade element:

$$F_c = m\omega^2 r$$

Where: m = mass of the element, ω = angular velocity (rad/s), r = radius from the axis

Since $\omega = 2\pi n$, centrifugal force increases with the **square** of RPM.

2. Aerodynamic Force (Aerodynamic Twisting Moment — ATM)

The **aerodynamic force** is the resultant of lift and drag acting on each blade element. This force acts through the **centre of pressure** of the blade section. Because the centre of pressure is typically ahead of (forward of) the pitch-change axis, the aerodynamic force creates the **Aerodynamic Twisting Moment (ATM)**, which tends to twist the blade toward **coarse pitch** (high blade angle).

The ATM opposes the CTM. In most propeller designs:

- **CTM > ATM** — the net tendency is toward fine pitch
- The pitch control system must therefore work **against** the CTM to increase blade angle (toward coarse/feather)
- This is an important safety consideration — if the pitch control fails, the blade will tend toward fine pitch, which at high speed means engine overspeed

3. Thrust Bending Force

The thrust produced by each blade element acts in the forward direction (parallel to the aircraft's flight path). Because the blade is attached at the hub and the thrust loads are distributed along its span, a **bending moment** is created that tries to bend the blade forward. The blade root must be strong enough to resist this bending.

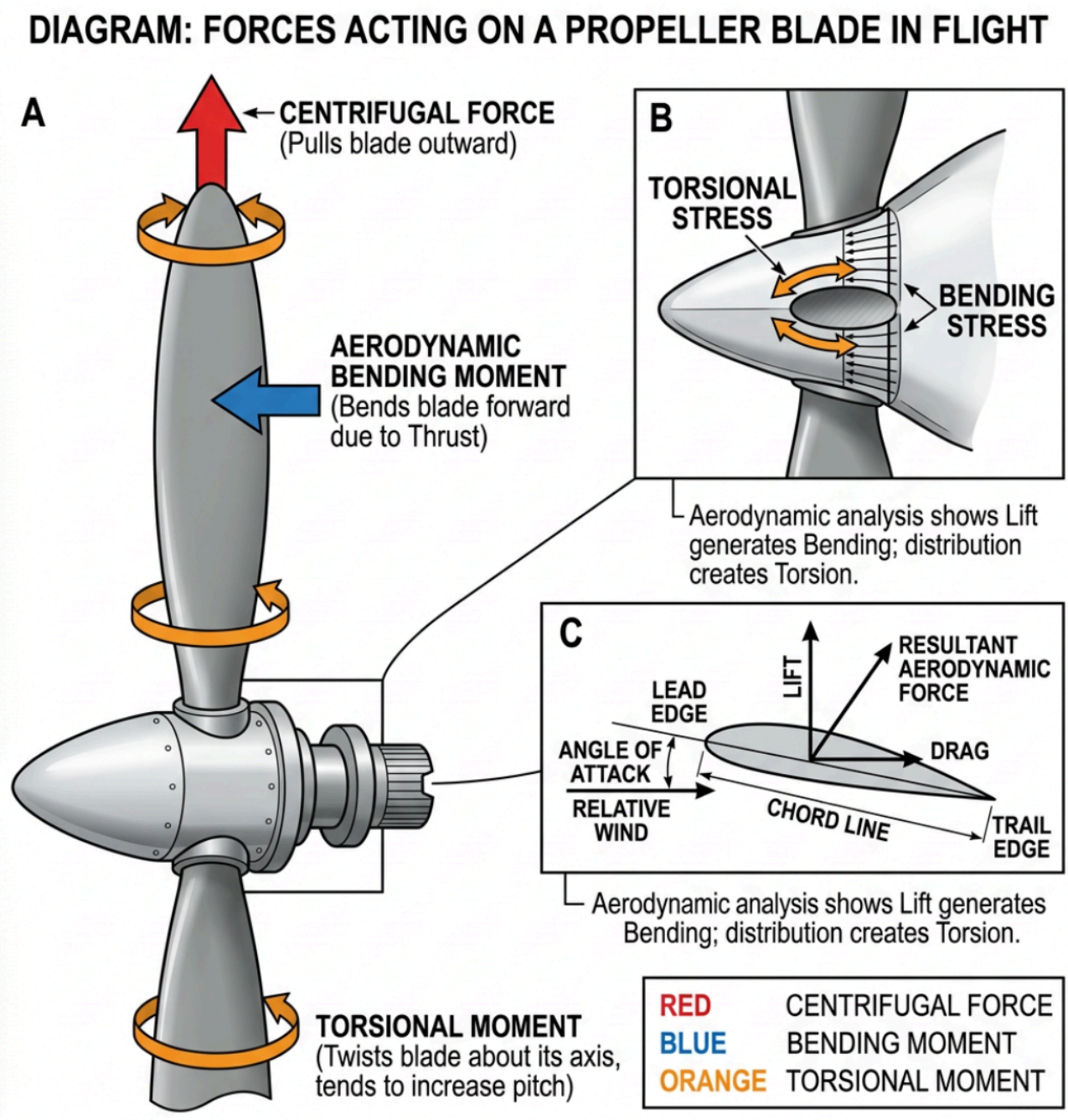
4. Torque Bending Force

As the engine drives the propeller, aerodynamic drag on each blade element resists the rotation. This creates a **torque bending force** that tries to bend the blade backward (against the direction of rotation). This opposes the thrust bending force but acts in a different plane.

5. Vibratory Forces

Cyclic variations in airflow (e.g., due to the aircraft's angle of attack causing the descending blade to see a different angle than the ascending blade) produce **once-per-revolution (1P) vibratory forces**. Higher-order

vibrations (2P, 3P, etc.) also occur. These vibratory forces can cause fatigue cracking if not controlled.



Reading the Forces Diagram

The diagram of forces above draws three of these loads on a single blade in flight, colour-coded, and it repays being read carefully because two of the arrows are easy to confuse with one another.

- **The red arrows are centrifugal force**, drawn pulling the blade radially outward, away from the hub. This is the tension load, and it is by far the largest force on the blade.

- **The blue arrow is the aerodynamic bending moment**, drawn bending the blade forward under the action of thrust. It acts out of the plane of rotation, along the flight direction.
- **The orange arrow is the torsional moment**, drawn twisting the blade about its own spanwise axis, and annotated as tending to increase the pitch.

⚠ Which twisting moment is which

The torsional moment drawn in that diagram, the one that **increases** the pitch, is the **aerodynamic** twisting moment. It arises from the aerodynamic resultant acting through a centre of pressure that lies ahead of the pitch-change axis, exactly as described above, and **in normal powered flight** it drives the blade towards a *higher* blade angle.

The **centrifugal twisting moment** — the moment produced by the blade's own mass distribution about its pitch-change axis — drives the blade the other way, towards a *lower* blade angle, and it does so because the mass of each blade element wants to swing into the plane of rotation. That direction is fixed: the centrifugal twisting moment acts towards fine pitch in every condition. It does not follow that centrifugal force can never coarsen a blade. **Counterweights** are offset masses fitted at the blade root precisely so that the centrifugal force acting on *them* produces a moment towards coarse pitch; they are a deliberate engineering use of centrifugal force, not an exception to what the centrifugal twisting moment does, and they are dealt with further on.

So in **normal powered flight** the two twisting moments are opposite in direction, opposite in origin and unequal in size: a twisting moment described as increasing the pitch is the aerodynamic one, and one described as reducing it is the centrifugal one. That pairing does not survive into the **windmilling** case. There the aerodynamic force on the blade has reversed, so the aerodynamic twisting moment reverses with it and acts towards fine as well, and the two *assist* each other instead of opposing — a case examined in full later in this note. Mixing the two moments up is the single most reliable way to get a propeller question wrong, and applying the powered-flight pairing to a windmilling question is the second.

The diagram also carries an inset showing lift, drag and the resultant aerodynamic force on one blade section, with the angle of attack measured between the relative wind and the **chord line**. That is the engineer's datum, and it is the datum used everywhere else in this note; be aware that the velocity diagram earlier in this note used the section's zero-lift line instead, and the two differ by the section's zero-lift angle.

Centrifugal Force in Numbers

The tension a blade carries is easy to underestimate until it is calculated, so it is worth building up from a single element. A blade is treated as a stack of elements, each contributing $m\omega^2 r$, and the total is the sum of all of them. The root of the blade carries the whole of that sum, because everything outboard of it is hanging on.

✣ Adding a blade up, slice by slice

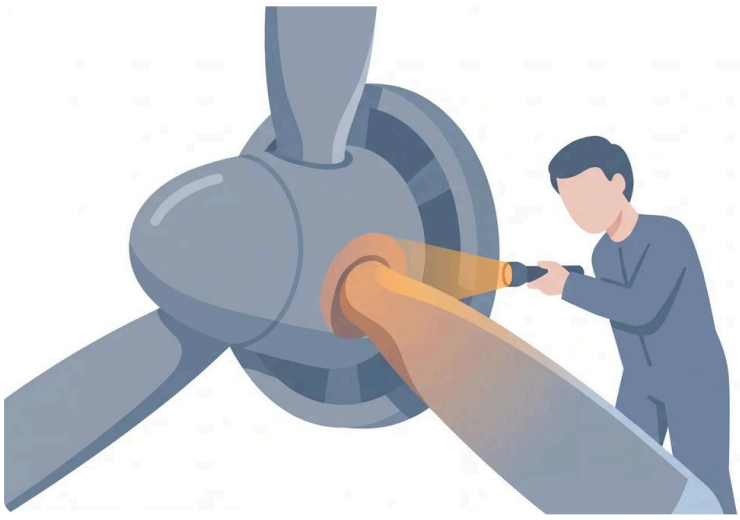
Model a 10 kg blade on the 2.5 m propeller as five 2 kg slices, at radii of 0.25, 0.50, 0.75, 1.00 and 1.25 m, with the propeller at 1,200 RPM so that $\omega^2 = 15{,}791$ (rad/s)^2. Each slice contributes $2 \times 15{,}791 \times r$:

- 0.25 m: 7,896 N · 0.50 m: 15,791 N · 0.75 m: 23,687 N
- 1.00 m: 31,583 N · 1.25 m: 39,478 N
- **Total carried by the blade root: 118,435 N** — about 12 tonnes-force.

Twelve tonnes from a ten-kilogram blade at a very moderate speed. Scale the blade up in mass and in radius, as a large turboprop does, and the whole-blade figures quoted above — loads of the order of tens of tonnes — are the entirely ordinary consequence of a square law applied to a heavy object at a large radius. Notice also how steeply the contributions grow towards the tip: mass added near the tip costs several times as much root stress as the same mass added near the hub, which is one more reason blades are made thin and light in their outer span.

Three design and maintenance consequences flow from that magnitude:

- **The blade retention is the most critical joint on the propeller.** Whether the blade is held by a ball or roller bearing race in a metal hub, by a clamped ferrule, or by a tension-torsion strap arrangement, that joint carries the full centrifugal load while also having to allow the blade to rotate freely in pitch. Its bearings, races and retaining components are inspected and lubricated to a schedule, and the lubrication is not optional — a dry or contaminated blade bearing raises the pitch-change friction, which shows up as sluggish or hunting governing long before it shows up as a mechanical failure.
- **The blade root and shank are the highest-stressed part of the blade.** Damage there is treated far more seriously than damage of the same size further out, and blending or dressing in the shank area is usually severely restricted or prohibited outright.
- **The load is present whenever the propeller turns**, including during ground running. It is not a flight-only case.



The blade root and shank carry the highest stress, so damage there is treated far more seriously.

Centrifugal Force Also Straightens the Blade

Centrifugal force does not only pull the blade out of the hub. Because it acts along a radial line from the axis of rotation, it also tries to pull the blade **straight** into the plane of rotation. Any deflection of the blade out of that plane — and thrust bending produces exactly such a deflection, forwards — therefore has a centrifugal restoring moment working against it.

This is genuinely useful. The bending stress the blade actually experiences is the thrust bending moment *minus* the centrifugal relief, and on a fast-turning propeller the relief is substantial. It also means the blade's real bending stress is not proportional to thrust alone: a blade producing high thrust at low RPM is worse off than one producing the same thrust at high RPM, because the second has more centrifugal straightening working for it. Some designs exploit the effect deliberately by building a small amount of forward rake into the blade, so that the blade's centre of mass sits slightly ahead of the plane of rotation and the centrifugal moment opposes thrust bending from the outset.

Living With the Twisting Moments

Because the centrifugal twisting moment normally exceeds the aerodynamic one, the blade's natural tendency in powered flight is towards fine pitch, and every pitch-change system has to be designed around that fact. Designers have three levers, and most propellers use more than one:

- **Counterweights** bolted to the blade root, set at an angle to the blade chord so that the centrifugal force acting on *them* produces a moment towards coarse pitch. They use the same square law as the moment they are cancelling, so their opposing effect grows with RPM at the same rate — which is exactly what is needed, since that is when the centrifugal twisting moment is largest.

- **Springs or a gas charge** in the pitch-change dome, providing a coarse-pitch or feathering force that does not depend on engine oil pressure and is therefore still available after the engine has failed.
- **Hydraulic pressure** acting on a pitch-change piston, which is what the governor actually modulates.

⚠ Never assume which way the oil drives the blade

The direction in which oil pressure moves the blade is a design choice and it genuinely differs between propeller types. On some designs the governor supplies oil to drive the blade towards **coarse**, with the centrifugal twisting moment and a spring returning it towards fine when the pressure is released. On others the oil drives the blade towards **fine**, and counterweights, springs or a gas charge drive it towards coarse and feather when pressure is lost.

The consequence of a loss of oil pressure is therefore opposite on the two arrangements, and so is the correct troubleshooting logic for a propeller that has run to one end of its range. Read the propeller maintenance manual for the type you are working on and establish which arrangement it uses before you form any theory about a fault. A safe assumption carried over from a different propeller is not a safe assumption.



Before troubleshooting a propeller at one end of its range, check the manual for which way the oil drives it.

One case reverses the usual relationship between the two moments, and it is examined. In powered flight the aerodynamic twisting moment works towards coarse and therefore opposes the centrifugal one. In the **windmilling** condition the blade is at a negative angle of attack and the aerodynamic force on it has reversed, so the aerodynamic twisting moment reverses with it and now acts towards fine pitch — the same direction as the centrifugal twisting moment. In windmilling the two **assist** each other rather than opposing, so the total moment driving the blade towards fine pitch is greater than at any point in normal flight. That is the least helpful moment for it to happen, because fine pitch is also the maximum-drag condition, and it is one of the reasons a feathering system has to be powerful and independent of the engine.

All Five Loads Side by Side

Load	Origin	Direction it acts	Grows with	What resists it
Centrifugal tension	Blade mass rotating at radius	Radially outward from the hub	Square of RPM; blade mass; radius	Blade shank, retention bearing and hub
Centrifugal twisting moment	Element centres of mass offset from the pitch axis	Twists the blade towards fine pitch	Square of RPM	Pitch-change mechanism, counterweights, springs
Aerodynamic twisting moment	Centre of pressure ahead of the pitch axis	Twists the blade towards coarse pitch in powered flight; towards fine when windmilling	Aerodynamic loading — airspeed, blade angle, power	Pitch-change mechanism
Thrust bending	Thrust distributed along the span	Bends the blade forward, greatest deflection at the tip	Thrust; falls with centrifugal relief	Blade section stiffness and root strength
Torque bending	Blade drag resisting rotation	Bends the blade backward, opposite to the direction of rotation, in the plane of rotation	Blade drag — power absorbed, RPM, blade angle	Blade edgewise stiffness and root strength

Grouping them makes the standard exam phrasing easy to see. The categories a propeller blade is designed against are **centrifugal, twisting and bending** — centrifugal tension pulling it out, twisting moments trying to change its angle, and bending in two perpendicular planes from thrust and from torque. Thrust and torque themselves are the propeller's useful aerodynamic *outputs*; they appear in the list only through the bending they cause, which is why an answer offering "thrust and torque" as the forces acting on a blade is describing what the propeller produces rather than what the blade has to survive.

Vibratory Loading and Fatigue

The five loads above are steady loads, and a blade would live indefinitely under steady loads. What actually limits a blade's life is the vibratory loading superimposed on them, because a stress that oscillates will eventually crack a component that a constant stress of the same size would never trouble.

Metals show a broadly consistent behaviour under repeated loading: the larger the alternating stress, the fewer cycles the part survives, and the relationship is steep. A propeller running at 1,200 RPM accumulates 72,000 once-per-revolution cycles in an hour, so even a small alternating stress accumulates enormous cycle counts in ordinary service. Two things then determine whether the blade survives:

- **The size of the alternating stress**, which is set by the excitation — asymmetric inflow, engine firing impulses, an out-of-balance condition, operation in a resonant band.

- **The presence of a stress raiser**, which multiplies the local stress at one spot. A sharp-bottomed nick, a scratch, a corrosion pit, a machining mark or a poorly blended repair all concentrate stress far above the nominal level in the surrounding material.

This is the whole reasoning behind the way propeller blade damage is treated. A nick is not removed because it looks untidy or because the missing metal weakens the blade — the amount of metal involved is trivial. It is removed because its sharp root is a crack starter, and it is dressed out to a **smooth, generously radiused blend** so that no sharp bottom remains. The same reasoning explains why leading-edge damage is treated more severely than damage elsewhere, why damage on the cambered face is judged differently from damage on the flat face, why corrosion must be removed and the area re-protected rather than simply cleaned, and why the finished blend must be re-protected and, on many types, re-shot-peened to restore the compressive surface stress that resists crack initiation.



A nick is dressed out to a smooth, generously radiused blend so no sharp, crack-starting root remains.

⚠ Damage limits are never transferable

The permitted depth, width, area and location of blade damage, the maximum amount of material that may be removed, the number of repairs allowed and the blend radius required are all specific to the propeller type, the blade part number and often the blade station. They differ between aluminium, steel and composite blades, and they differ between manufacturers. There is no general figure, and quoting one from another type is not a conservative guess — it is an unairworthy assessment in either direction. Take the limits from the propeller maintenance manual for the blade in front of you, and if the damage is outside them, the blade goes to an approved overhaul facility.

Finding the Damage Before It Finds You

Because the consequence of a blade failure is severe — the loss of a blade produces an instantaneous and enormous out-of-balance force that can tear an engine from its mounts — propeller inspection is layered:

- **Visual and tactile inspection** at every scheduled check, and after any event that could have damaged the blades: a suspected bird or ground strike, ingestion of stones or debris, operation from a contaminated surface, or a report of unusual vibration. Running a gloved hand along the leading edge finds nicks the eye misses.
- **Dye penetrant inspection** for surface-breaking cracks in non-ferrous blades, and **magnetic particle inspection** for ferrous components such as steel blades, hub parts and retaining hardware.
- **Eddy current** for surface and near-surface cracks without stripping the finish, and **ultrasonic** for sub-surface defects and for checking bond integrity and disbonds in composite blades.
- **Tap testing and specialised techniques** for composite blades, where the failure modes — delamination, disbond of the erosion shield, water ingress into the core — are quite different from the cracking that dominates metal blades.

Alongside the blades themselves, the installation is monitored for the consequences of these loads: magnetic chip detectors and oil filters in the reduction gearbox catch the debris that bearing and gear distress produces long before the distress becomes a failure, and any finding on a chip detector is investigated to establish what the material is and where it came from rather than simply cleaned off. A propeller strike — any contact between a turning blade and the ground or an obstruction — is a special case: it applies a violent, instantaneous torque reversal through the blade, hub, shaft and gearbox, and it triggers a mandatory inspection of the whole drive train whose scope is defined by the engine and propeller manufacturers, not by how the blade happens to look afterwards.

Blade Materials and the Loads They Are Chosen For

The five loads above are the design case a blade material has to answer, and the materials in service answer them in noticeably different ways. Understanding the differences is what makes the inspection and repair regimes of each type make sense rather than seeming arbitrary.

Material	Strengths against the load set	Characteristic problems
Aluminium alloy	Good strength for its weight, easy to form into an accurate aerofoil, and light enough to keep centrifugal loads moderate	Notch-sensitive, so nicks and pits become crack starters; subject to corrosion; limited in how much material may be removed in a repair
Steel	Very high strength, and tolerant of erosion and impact; used for hollow blades on some large propellers	Heavy, so the centrifugal loads it must itself carry are larger; corrodes readily if the protective finish is broken

Material	Strengths against the load set	Characteristic problems
Wood and laminated wood	Light, well damped, and forgiving of vibration; still found on light and vintage aircraft	Absorbs moisture and can distort or delaminate; erodes quickly without a leading-edge sheath; sensitive to storage position and to over-torqued mounting bolts
Composite	Very light for its stiffness, which allows wide, swept, low-tip-speed blades; internal damping is good and the shape can be tailored	Damage can be internal and invisible — delamination, disbond, water ingress into the core; needs its own inspection methods and specialised repair

Two features are common to all of them. Every blade needs its **leading edge protected**, because that edge meets rain, hail, grit and insects at the highest relative velocity on the aircraft; the protection may be a bonded metal erosion shield, a plated or sprayed coating, or a replaceable strip, and its condition and bond are inspected in their own right. And every blade needs its **surface finish maintained**, because the finish is a corrosion barrier and, on notch-sensitive materials, part of the fatigue protection — which is why a dressed area must be re-protected rather than left bare, and why on many types a repaired area is re-peened to restore the compressive surface stress.

The consequence for an engineer is that no damage assessment can be made without knowing what the blade is made of. A dent that would be dressed out and returned to service on one material may be a scrap condition on another, and a composite blade that looks undamaged on the surface may have a substantial internal disbond. The propeller maintenance manual for the specific blade part number is the only source for that judgement.

Torque

Torque is the rotational force (turning moment) that the engine must supply to keep the propeller turning against aerodynamic drag. It is the "resistance" the propeller presents to the engine.

Relationship between power, torque, and RPM:

$$P = 2\pi n Q$$

Where: P = shaft power (W), n = rotational speed (rev/s), Q = torque (N·m)

Rearranging: $Q = \frac{P}{2\pi n}$

This shows that for a given power, **reducing RPM increases torque** — this is why turboprop engines use reduction gearboxes. A turbine spinning at 30,000 RPM produces relatively low torque; the gearbox converts this to ~1,200 RPM at much higher torque.

Engine torque must match propeller torque at the selected RPM. If engine torque exceeds propeller torque, the propeller accelerates. If propeller torque exceeds engine torque, the propeller decelerates. The constant-

speed propeller governor works to maintain this balance by adjusting blade angle.

Four Things Called Torque, and Keeping Them Apart

The word appears in four distinct senses in propeller work, and questions exploit the overlap. Fixing them in order of where they act removes most of the confusion:

Term	Where it acts	Which way
Propeller torque	On the propeller, from the air	Opposite to the direction of rotation — it is a resistance to being turned
Engine torque	On the propeller shaft, from the engine	In the direction of rotation — it is what overcomes the above
Torque reaction	On the airframe, from the engine mounts	Opposite to the direction of propeller rotation — it tries to roll the aircraft the other way
Torque bending	On the individual blade	Bends the blade rearward, opposite to its own direction of travel — one of the two bending cases described above

The first is the one meant when a question asks what propeller torque is: the aerodynamic **resistance the propeller offers to rotation**. It is not a resistance to changing pitch, and it is not a resistance to feathering — those loads exist, but they are carried by the pitch-change mechanism and have nothing to do with the propeller's torque. Nor does propeller torque act in or at right angles to the plane of rotation as a linear force: it is a moment about the shaft axis, opposing rotation, and the plane-of-rotation and axial directions belong to the blade's bending loads and to thrust respectively.

Where the Torque Comes From on the Blade

Torque is not a separate phenomenon from lift and drag; it is what those two look like when they are resolved into the plane of rotation. As the elemental expressions earlier in this note show, the in-plane component contains the section's drag *plus* a share of its lift tilted into the plane by the helix angle. Two things follow that are not obvious:

- **Even a perfectly drag-free blade would demand torque.** The lift vector is tilted rearward by the helix angle, and that tilted component has to be paid for. This is the induced part of the torque, and it is the price of producing thrust at all.
- **The outer blade dominates the torque as well as the thrust**, because the in-plane force is multiplied by the radius to give the moment. A given force at the tip costs several times the torque of the same force near the hub, which is another reason blade condition matters most in the outer span.

Tying torque, power, thrust and efficiency together

An engine delivers 1,500 kW to a propeller turning at 1,200 RPM ($n = 20$ rev/s) with an efficiency of 80% at a true airspeed of 100 m/s.

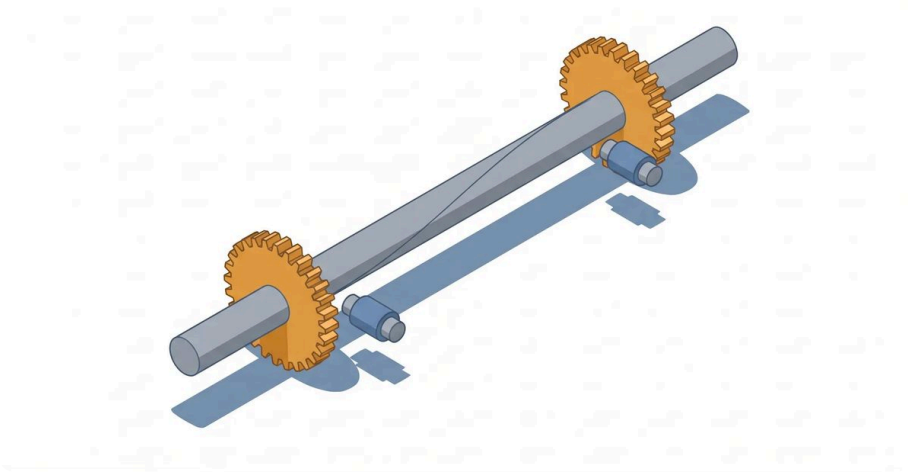
- Shaft torque: $Q = P / 2\pi n = 1{,}500{,}000 / (2\pi \times 20) = 11{,}937$ N·m.
- Thrust power delivered: $0.80 \times 1{,}500{,}000 = 1{,}200{,}000$ W.
- Thrust: $T = 1{,}200{,}000 / 100 = 12{,}000$ N.

Notice what the efficiency figure did: the missing 300 kW went into blade drag, slipstream swirl and the energy left in the accelerated air, and it appears in the shaft torque without appearing in the thrust. The engine paid for it; the aircraft did not receive it.

Measuring Torque: the Torquemeter

On a free-turbine turboprop the pilot has no direct measure of shaft power, and propeller RPM is being held constant by the governor and so tells him nothing about it either. Torque is therefore measured directly and becomes the primary power-setting parameter, because with RPM held constant, torque is proportional to power.

Two families of torquemeter are in common use, and both measure the reaction to the torque rather than the torque itself:



A phase-shift torquemeter compares signals from toothed wheels at each end of a shaft that twists under load.

- **Hydromechanical torquemeters** built into the reduction gearbox. The gear train is arranged so that the torque reacted by a helical gear or a ring gear produces an axial movement against a piston, and the oil

pressure needed to hold that piston in place is proportional to the torque being transmitted. The indication may be presented directly as a pressure or converted into torque units.

- **Torsion-shaft or phase-shift torquemeters**, which measure the small angular twist of a shaft under load by comparing the timing of signals from two toothed wheels at either end of it. The greater the torque, the greater the phase difference between the two signals.

Torque may be indicated as a percentage of a datum, in newton metres, in pound-feet, or as a pressure, depending on the installation, and none of those is interchangeable with another. What matters operationally is that torque is only one of the limits in force at any moment:

Which limit bites first, and why it changes with the weather

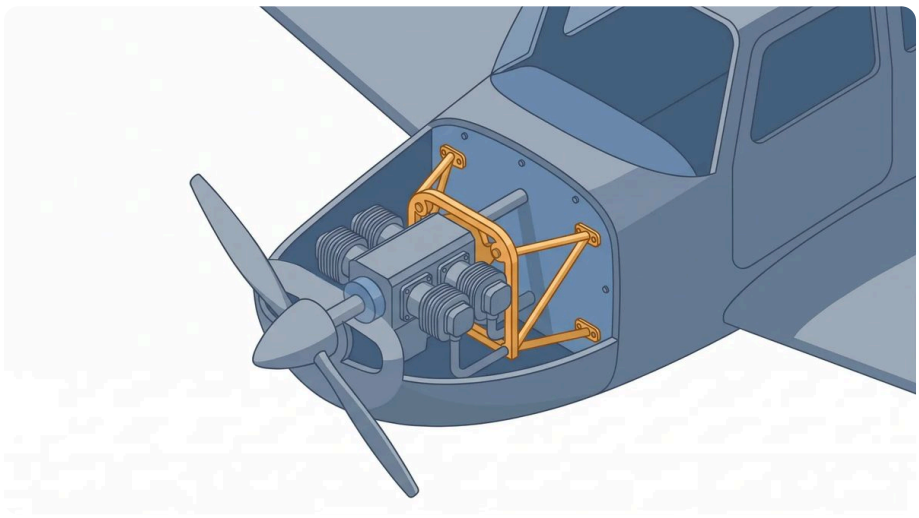
A turboprop is normally limited either by torque or by turbine temperature, and which one is reached first depends on the ambient conditions. On a **cold day** the air is dense, the engine handles a large mass flow at a modest turbine temperature, and the shaft power available is high — so the **torque limit** is reached while temperature is still comfortable and the engine is said to be torque limited. This is the purpose of **flat rating**: the engine is certificated at a power its gas generator could exceed on a cold day, so that the gearbox and propeller are never asked to carry more torque than they were designed for, and so that the same rated power stays available as the ambient temperature rises. On a **hot day** the mass flow falls and the turbine has to run hotter for the same power, so the **temperature limit** is reached before full torque can be developed, and the available power is reduced.

This is why a turboprop's power setting procedure is a table or a chart rather than a single number, and why an engineer investigating a "low power" report must establish which limit the crew were up against before concluding anything is wrong with the engine or the propeller.

From a maintenance point of view, the torque indication is also a diagnostic. Torque and RPM together define power, so an engine reaching its target RPM at a higher-than-expected torque, or a lower-than-expected one, is telling you that the relationship between the propeller's load and the engine's output has shifted — a mis-set blade angle, a governing fault, blade surface deterioration, or an indication error. Ground-running figures for torque at defined conditions are published for exactly this purpose, and a torque indication is verified against them after work on either the engine or the propeller.

Torque Reaction and the Other Propeller Effects on the Airframe

Newton's third law does not stop at the propeller shaft. If the engine applies a torque to the propeller in the direction of rotation, the propeller applies an equal and opposite torque to the engine, and the engine mounts pass it into the airframe. The aircraft therefore experiences a rolling moment **opposite** to the direction of propeller rotation, and it is largest when the torque is largest — high power and low airspeed, which is to say take-off.



The engine mounts pass the propeller's equal and opposite torque into the airframe as a rolling moment.

Torque reaction is one of four propeller effects that act on a single-propeller aeroplane, and it is worth having all four in one place with their directions stated. Take, as the reference case, a propeller that rotates **clockwise when viewed from behind**, which is to say from the cockpit looking forward — the convention used for most Western installations, though the direction of rotation for any particular type must be taken from its own type data.

Effect	Mechanism	Result on the reference aircraft	Worst at
Torque reaction	Equal and opposite reaction to the torque driving the propeller	Rolls the aircraft to the left, which loads the left wheel and adds a left yaw through tyre friction on the ground	High power, low speed
Spiralling slipstream	The swirl the propeller adds to the slipstream wraps around the fuselage and strikes one side of the fin	Yaws the aircraft to the left	High power, low speed
Asymmetric blade effect (P-factor)	At a nose-up attitude the descending blade works at a higher angle of attack and makes more thrust than the ascending blade	Thrust centre moves to the right of the disc, yawing the aircraft to the left	High aircraft angle of attack — climb, slow flight
Gyroscopic effect	The rotating mass of the propeller precesses: a force applied to the disc takes effect 90° further round in the direction of rotation	Pitching the nose down — raising the tail — yaws the aircraft to the left; pitching up yaws it right	Rapid pitch changes at high RPM

Reverse the direction of propeller rotation and every one of those directions reverses with it. That is not a trivia point: it is the reason propeller **handedness** is a controlled characteristic of a propeller assembly, and the reason a left-hand and a right-hand propeller are different part numbers rather than the same propeller fitted

the other way up. Blade twist and section camber both have a handedness built into them, so a blade for a left-turning propeller cannot be made to work on a right-turning one.

Counter-Rotating, Contra-Rotating, Tractor and Pusher

Four terms describe how propellers are arranged, and two pairs of them are routinely confused:

- **Counter-rotating** describes a multi-engine aircraft whose engines turn in *opposite directions* — one propeller clockwise, the other anticlockwise. Each engine's torque reaction, slipstream and P-factor cancels the other's, and, importantly, neither engine is the critical engine.
- **Contra-rotating** describes *two coaxial propellers on one engine*, turning in opposite directions on concentric shafts. The rear propeller recovers the swirl the front one put into the air, improving efficiency and cancelling torque reaction, at the cost of a complex gearbox and a difficult noise signature.
- **Tractor** means the propeller is ahead of its mounting and pulls the aircraft, with the blades working in undisturbed air and the slipstream passing over the airframe behind them. This is the usual arrangement.
- **Pusher** means the propeller is behind its mounting and pushes. The airframe is not immersed in the slipstream, which is quieter in the cabin and leaves the wing in clean air, but the propeller itself works in the wake of whatever is ahead of it — and that wake is a once-per-revolution disturbance, so pusher propellers face a harsher vibratory environment and are more exposed to debris thrown up by the wheels.

The critical-engine idea is worth one more sentence, because it is the practical consequence of the whole section. On a twin whose propellers both turn clockwise viewed from behind, P-factor moves each propeller's effective thrust line to the right of its own shaft. If the **left** engine fails, the surviving right engine's thrust line is displaced further from the aircraft centreline, giving it a longer arm and a larger asymmetric yawing moment to control. The left engine is therefore the critical engine on such an aircraft, and its failure is the more demanding case. Counter-rotating propellers remove the asymmetry, and with it the concept of a critical engine.

Torque and Power in Imperial Units

A great deal of propeller and engine data is still published in pound-feet and horsepower, and the imperial version of the power relationship carries a constant that is worth understanding rather than memorising.

Where 5252 comes from

$$P_{hp} = \frac{Q_{lb\cdot ft} \times N_{RPM}}{5252}$$

One horsepower is 550 foot-pounds per second, and power is torque times angular velocity, with the angular velocity in radians per second equal to $2\pi N/60$. Dividing $2\pi/60$ by 550 and inverting gives $60 \times 550/2\pi = 5252$. The constant is nothing more than a unit conversion, and it is exactly the same physics as $P = 2\pi n Q$ in SI units.

The same shaft, both systems

1,500 kW at 1,200 RPM. In SI the shaft torque is 11,937 N·m, as calculated earlier in this note.

- Convert the power: $1{,}500{,}000/745.7 = 2{,}012$ hp.
- $Q = 2012 \times 5252/1200 = 8{,}804$ lb·ft.
- Cross-check by converting the SI answer directly: $11{,}937/1.356 = 8{,}804$ lb·ft.

The two routes agree, which is the only real proof that a unit conversion has been done correctly. Always finish an imperial-to-SI calculation by working it the other way round.

How the Torque Actually Reaches the Propeller

Between the engine and the propeller sits a gearbox that has to reduce the speed by a large factor while carrying the whole of the multiplied torque. Two architectures are common, and they place the loads differently:

- **Offset spur or layshaft gearing**, in which the propeller shaft is displaced from the engine centreline. The offset can be used to raise the propeller for ground clearance, which is why the propeller shaft of many turboprops sits noticeably above the engine axis. The load is carried through a small number of gear meshes, so the teeth are heavily loaded and the bearing reactions are large.
- **Planetary, or epicyclic, gearing**, in which a central sun gear drives several planet gears inside an annulus. The propeller shaft stays on the engine centreline, and the torque is shared between several meshes at once, which allows a large reduction in a compact and relatively light unit. Multiple stages can be arranged in series where a very large reduction is needed.

Whichever is fitted, the gearbox is a torque path with a defined life, and its condition is monitored through the same oil system that lubricates it. It also carries the propeller's thrust into the airframe through a thrust bearing, and its casing reacts the torque into the engine mounts. From an engineer's point of view the practical points are consistent across types: the oil is a condition-monitoring medium as much as a lubricant, chip

detector findings are investigated rather than wiped, and the gearbox is part of the mandatory inspection following any propeller strike.

Torque Is Not Steady

The formulae above treat torque as a constant, and on a turbine driving a propeller through a gearbox it very nearly is. On a piston engine it is not: each firing stroke delivers a pulse, and between pulses the torque falls away and can even reverse during compression. The propeller, the crankshaft and the gearing form a torsionally flexible system with natural frequencies of its own, and the firing pulses excite it in exactly the way described later in this note.

The consequences appear as torsional vibration rather than as a change in mean torque, and they are managed by the measures covered in the vibration section — tuned dampers on the crankshaft, careful choice of the propeller's own frequencies, and where necessary a restricted RPM band. It is worth being clear that the mean torque an engine delivers and the alternating torque superimposed on it are separate quantities: an installation can be perfectly within its torque limit and still be destroying itself torsionally, which is why RPM restrictions exist in aircraft that are nowhere near any power limit.

Relative Airflow on Blade Angle of Attack

The **relative airflow (RAF)** striking any blade element is the **vector sum** of two perpendicular velocities:

1. **Rotational velocity** — tangential to the circle of rotation, in the plane of rotation. This is $V_{rot} = 2\pi r n$, where r is the blade station radius.
2. **Forward velocity (TAS)** — the aircraft's true airspeed, perpendicular to the plane of rotation.

The resultant RAF approaches the blade element from below and ahead. The angle this resultant makes with the plane of rotation is the **helix angle (ϕ)**:

$$\tan\varphi = \frac{V_{TAS}}{2\pi r n}$$

The helix angle is **small near the tip** (because rotational velocity is large) and **large near the hub** (because rotational velocity is small).

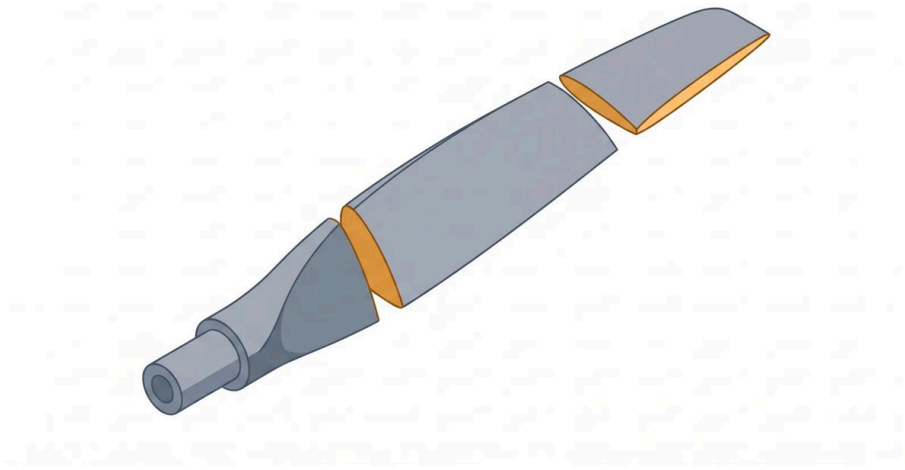
Since the blade angle β is fixed at each station (for a given pitch setting), and β = ϕ + α, the angle of attack α at each element depends on the helix angle:

$$\alpha = \beta - \varphi$$

This means:

- **Increasing forward speed** (at constant RPM) → larger ϕ → smaller α → less thrust
- **Decreasing forward speed** (at constant RPM) → smaller ϕ → larger α → more thrust (up to stall)
- **Increasing RPM** (at constant TAS) → smaller ϕ → larger α → more thrust

Why propeller blades are twisted: Because rotational velocity increases with radius, the helix angle ϕ is large at the root and small at the tip. To maintain a roughly constant (efficient) angle of attack α along the entire blade, the blade angle β must be **large at the root** and **small at the tip**. This built-in twist is called **geometric twist** or **washout**.



Blade angle is large at the root and small at the tip, keeping angle of attack roughly constant.

One Station Through a Whole Flight

Earlier in this note the helix angle was calculated at four stations along the blade at a single flight condition. The other way of cutting the same problem is more useful to an engineer: take one station and follow it through the flight, so that the demand the governor has to satisfy becomes visible.

✈ The 75% station, constant 1,200 RPM, from chocks to descent

The reference station of the 2.5 m propeller sits at $r = 0.9375$ m, and at 1,200 RPM its rotational velocity is $2\pi \times 0.9375 \times 20 = 117.8$ m/s. That figure does not change, because the governor is holding the RPM. Only the forward speed changes.

Condition	TAS	Helix angle ϕ	Blade angle needed for α = 3°
Static run-up	0	0°	3°
Take-off	50 m/s	23.0°	26.0°
Climb	70 m/s	30.7°	33.7°
Cruise	100 m/s	40.3°	43.3°
High-speed cruise	130 m/s	47.8°	50.8°
Descent	150 m/s	51.9°	54.9°

Fifty-two degrees of blade angle movement at one station, demanded by nothing more than the aircraft accelerating from a standstill to its descent speed. That is the size of the job a constant-speed unit does, and it is the reason a fixed-pitch propeller can only ever be right in one row of that table.

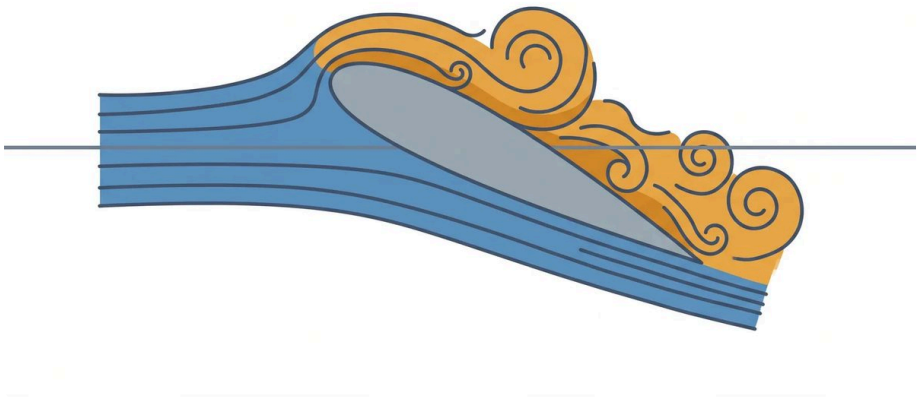
The Static Case Deserves Its Own Line

The first row of that table is worth pausing on, because it is the condition an engineer meets most often. With the aircraft stationary the forward velocity is zero, so the relative airflow is purely rotational, it lies entirely in the plane of rotation, and the helix angle is zero. The relationship therefore collapses to

🖩 With no forward speed, the whole blade angle is angle of attack

$$V = 0 \;\Rightarrow\; \varphi = 0 \;\Rightarrow\; \alpha = \beta$$

Whatever blade angle the propeller is sitting at during a ground run is the angle of attack every section is working at. A blade angle of 15° at the fine stop means an angle of attack of 15°, which is far past the few degrees the section wants, so a substantial part of the blade is stalled on every static run.



With no forward speed, air meets the blade only in the plane of rotation, so a steep blade stalls.

Two practical consequences follow, and both are things engineers see:

- **A ground run is not a test of cruise behaviour.** The propeller is operating at a condition it never sees in flight, at a high angle of attack, with heavily disturbed inflow. Static figures are perfectly good as a repeatable comparison against the manual, but they are not evidence about how the propeller performs in the cruise.
- **The finer the blade angle, the better the static case.** This is exactly why the flight fine stop determines the static RPM, why a constant-speed propeller sits against that stop for take-off, and why an over-coarse fine stop shows up as an engine that cannot reach its static RPM figure.

The Inflow Is Faster Than the Aircraft

The simple vector sum takes the axial velocity at the disc as the aircraft's true airspeed. It is not quite. The propeller has already begun to accelerate the air before that air reaches the disc, so the axial velocity through the disc is the true airspeed **plus** the induced velocity. This is the axial inflow velocity labelled on the velocity diagram earlier in this note, and it is the reason that diagram distinguishes an advance velocity from the aircraft's own speed.

How much difference the induced velocity makes

Take the cruise row above — 100 m/s TAS, 117.8 m/s rotational velocity at the reference station — and add an induced velocity of 10 m/s, so the true axial inflow is 110 m/s.

- Simple calculation: $\tan\varphi = 100/117.8$, giving $\varphi = 40.3°$.
- With induced velocity included: $\tan\varphi = 110/117.8$, giving $\varphi = 43.0°$.

The real helix angle is nearly three degrees larger, so the real angle of attack at a given blade angle is nearly three degrees **smaller** than the simple calculation predicts. At working angles of attack of only a few degrees, that is not a rounding error — it is most of the angle. The blade angle actually needed to achieve 3° at this condition is about 46°, not the 43.3° the simple sum gives.

The correction always goes the same way: the induced velocity always adds to the axial inflow, so it always increases the helix angle and always reduces the angle of attack below the simple estimate. It is largest when the propeller is heavily loaded at low forward speed — take-off and climb — and smallest in a fast, lightly loaded cruise. None of this changes any of the qualitative conclusions in this note; it is the reason those conclusions are qualitative, and the reason real propeller design uses the combined blade-element momentum method rather than the simple triangle.

Density, Altitude and the Difference Between Airspeeds

The velocity that enters the vector sum is the **true** airspeed, because it is the actual rate at which the aircraft is moving through the air mass. That has consequences that are easy to state and easy to get backwards, so work them through in order for an aircraft climbing at a constant indicated airspeed:

- Air density falls with altitude, so for a **constant indicated airspeed the true airspeed rises**.
- A higher true airspeed at constant RPM means a **larger helix angle**.
- A larger helix angle at a given blade angle means a **smaller angle of attack**, and therefore less thrust per blade.
- Separately, the falling density reduces the dynamic pressure at every blade section, which reduces the force produced at any given angle of attack.

Both effects push the same way, so propeller thrust falls with altitude for two independent reasons. A constant-speed propeller compensates for the first by coarsening the blade — the governor holds the RPM and the blade angle simply increases as the climb progresses — but nothing can compensate for the second, which is a property of the air rather than of the propeller.

The Inflow Is Not Uniform Across the Disc

Every calculation so far has assumed the air arrives evenly across the whole disc. In an installed propeller it never does, and the departures are the origin of most propeller vibration:

- **Aircraft angle of attack and sideslip** incline the disc to the oncoming flow, so each blade's local angle of attack rises and falls once per revolution. This is the asymmetric blade effect discussed earlier in this note, and it is present to some degree in almost every phase of flight.
- **Upwash ahead of the wing** bends the flow upward before it reaches a wing-mounted propeller, adding a further inclination that varies with the aircraft's own lift coefficient — so it is largest when the aircraft is slow and heavy.
- **Blockage and wake from what is behind or ahead of the disc.** A tractor propeller's inflow is disturbed by the nacelle, spinner and cowling behind it; a pusher propeller works directly in the wake of the wing, pylon or fuselage ahead of it, which is a far stronger and sharper disturbance.
- **Proximity to the ground** during take-off and landing distorts the inflow and, with reverse selected, can allow the propeller to draw its own disturbed slipstream back through the disc.

Each of these makes the relative airflow, and therefore the blade loading, a function of the blade's position around the disc as well as of its radius. That is precisely why a propeller blade is designed as a fatigue-critical component and why its loading is described in terms of harmonics of the rotational speed rather than as a steady figure.

What the Engineer Controls

The relative airflow itself is not adjustable. What is adjustable is the blade's relationship to it, and there are only two things to get right:

- **The blade angle must be correct at the reference station**, because that sets the angle of attack at every station for any given flight condition.
- **Every blade on the hub must agree with every other blade**, because they all meet the same relative airflow at the same instant. This is the reason the blade-to-blade angle comparison described earlier in this note is part of the same procedure as setting the angle itself.

The related physical check is **blade tracking**: with the propeller stationary and the engine safely isolated, the blades are rotated one at a time past a fixed reference and the position of each tip is compared with the others. A blade that runs out of track is not following the same circular path as its neighbours, so it is not meeting the airflow in the same place, and the disc is aerodynamically as well as geometrically uneven. Out-of-track can come from a bent blade, a damaged or incorrectly assembled hub, a distorted mounting flange or

an incorrectly torqued installation. The permitted track tolerance is a type-specific figure from the propeller or aircraft maintenance manual.



Blades are turned one at a time past a fixed reference so each tip position can be compared.

The order of work matters and is worth committing to memory: **track and blade angle first, balance afterwards**. Tracking and blade-angle errors are aerodynamic faults, and no amount of mass balancing can cancel an aerodynamic imbalance — adding a weight to fix an aerodynamic problem simply produces a propeller that is wrong in two ways at once.

The Relative Airflow Has a Size as Well as a Direction

All the angle work above uses the *direction* of the resultant velocity. Its *magnitude* matters just as much, because the aerodynamic force a section produces is proportional to the dynamic pressure, which goes with the square of that magnitude:

Resultant velocity and local dynamic pressure

$$V_R = \sqrt{V_{TAS}^2 + (2\pi r n)^2} \qquad q = \tfrac{1}{2}\rho V_R^2$$

Because the rotational term grows with radius and the whole thing is then squared, the dynamic pressure a section experiences rises very steeply from root to tip.

✈ Root and tip compared

Same propeller, 1,200 RPM, 100 m/s TAS, sea-level density 1.225 kg/m^3:

- **20% station:** $V_R = \sqrt{100^2 + 31.4^2} = 104.8$ m/s, so $q = 0.5 \times 1.225 \times 104.8^2 = 6{,}730$ Pa.
- **Tip:** $V_R = \sqrt{100^2 + 157.1^2} = 186.2$ m/s, so $q = 0.5 \times 1.225 \times 186.2^2 = 21{,}240$ Pa.

Just over three times the dynamic pressure at the tip, at the same instant, on the same blade. Combine that with the fact that the tip also has the longer moment arm for both bending and torque, and the reason the outer blade dominates the propeller's behaviour — and the reason it is treated so carefully in maintenance — needs no further explanation.

Solving the Triangle the Other Way Round

The relationship between the three angles can be run backwards, and doing so answers questions that are otherwise hard to reason about. The criterion is purely geometric:

- Blade angle **greater** than the helix angle: positive angle of attack, positive thrust, engine driving.
- Blade angle **equal to** the helix angle: zero angle of attack at the chord line, and the boundary between the two states.
- Blade angle **less** than the helix angle: negative angle of attack, negative thrust.

🖩 At what RPM does this propeller start to windmill?

The blade is at 45° at the reference station (radius 0.9375 m) and the aircraft is doing 100 m/s. The boundary is where the helix angle also equals 45°, which happens when the rotational velocity equals the forward velocity, because $\tan 45^\circ = 1$.

$$2\pi r n = V \;\Rightarrow\; n = \frac{100}{2\pi \times 0.9375} = 17.0 \text{ rev/s}$$

That is about **1,019 RPM**. Above it the blade angle exceeds the helix angle, the angle of attack is positive and the propeller can produce thrust. Below it the propeller is being driven by the airflow. This is exactly what happens when an engine fails: the RPM decays, crosses that boundary, and the propeller changes from a thrust producer to a drag producer without anybody touching a lever.

One refinement, since this note has already drawn the distinction elsewhere: the point at which the *axial force* passes through zero lies a fraction of a degree away from the point at which the angle of attack does, because the section's own drag has to be resolved as well as its lift. The two are close enough together that the blade-angle-equals-helix-angle crossover is the boundary used both in practice and in examination questions.

Relative Airflow With the Blade in Reverse

Reverse pitch is the one case where the blade angle itself is negative, and the triangle handles it without any special rules — the arithmetic simply carries the sign through.

On the landing roll

The aircraft is rolling at 30 m/s, the propeller is at 1,200 RPM, and the blade has been driven to –10° at the reference station. The rotational velocity there is still 117.8 m/s, so:

$$\tan\varphi = \frac{30}{117.8} = 0.255 \;\Rightarrow\; \varphi = 14.3^\circ$$

$$\alpha = \beta - \varphi = -10^\circ - 14.3^\circ = -24.3^\circ$$

A negative angle of attack of more than twenty degrees — far beyond anything in the flight range, and with the engine driving hard. The flow over the blade is heavily separated, the loading is large and unsteady, and the section is nowhere near any condition it was optimised for. That is the quantitative reason reverse is time-limited, noisy, rough and hard on the propeller, and why the blade angle used for reverse is only modestly negative rather than as negative as the mechanism could reach.

Notice how this case differs from windmilling even though both have a negative angle of attack and a negative thrust. In reverse the **engine is driving the propeller** and torque is being absorbed in the normal sense; in windmilling the **airflow is driving the propeller** and the engine is being turned. Same sign on the angle of attack, same sign on the thrust, opposite direction of power flow — which is why a question asking about angle of attack and thrust can be answered from the signs alone, but a question asking what is driving what cannot.

Vibration and Resonance

Propeller vibration is a critical concern for both structural integrity and crew/passenger comfort. Vibration originates from several sources:

Sources of Propeller Vibration

Source	Description	Frequency
Mass imbalance	Unequal mass distribution between blades (manufacturing tolerance, damage, repair, ice). Creates 1-per-revolution (1P) vibration.	1P (once per rev)
Aerodynamic imbalance	Unequal thrust between blades due to differences in blade angle, aerofoil profile, or surface condition. Creates 1P vibration.	1P

Source	Description	Frequency
Asymmetric airflow	When the aircraft is at an angle of attack, the descending blade has a higher AoA than the ascending blade. Creates cyclic 1P loading.	1P
Engine firing impulses	On piston engines, each cylinder fires in sequence, creating torque pulses that excite the propeller. A 4-cylinder engine at 2,400 RPM creates pulses at 80 Hz.	Varies with engine type
Propeller wake/P-factor	Interaction between the propeller slipstream and aircraft structure (wing, nacelle) creates periodic forces.	Blade-passing frequency

Resonance

Resonance occurs when the frequency of an exciting force matches a **natural frequency** of the propeller blade (or the engine/airframe system). At resonance, vibration amplitude increases dramatically — potentially to the point of **structural failure**.

Every propeller blade has multiple natural frequencies corresponding to different modes of vibration:

- **First flap mode** — the blade bends forward and backward (like a diving board)
- **First edgewise mode** — the blade bends in the plane of rotation (side to side)
- **Torsional mode** — the blade twists about its spanwise axis
- **Higher-order modes** — more complex shapes with multiple nodes

The operating RPM range of the propeller must be chosen to **avoid** critical resonance speeds. If a resonance cannot be avoided, it must pass through quickly (hence "restricted RPM ranges" in some aircraft flight manuals — RPM ranges where you must not operate continuously).

Operational note: Some aircraft have **yellow-arc RPM ranges** on the tachometer that indicate resonance danger zones. These ranges must be transited quickly and never used for continuous operation. Operating in a resonance range can lead to fatigue cracking and ultimately blade failure.

Counting the Excitation in Orders

Propeller vibration is described in **orders** — multiples of the shaft rotational frequency — rather than in absolute frequencies, because that is the form in which the excitation actually arrives. An order is written as a number followed by P: 1P is one cycle per revolution, 2P is two, and so on. Converting an order into a frequency needs only the shaft speed in revolutions per second.

From RPM to frequency

$$f_{nP} = n \times N_{order} \quad \text{with } n \text{ in rev/s}$$

A four-blade propeller at 1,200 RPM turns at $n = 20$ rev/s, so 1P is 20 Hz and 2P is 40 Hz. The **blade-passing frequency** — the rate at which blades pass any fixed point such as a nacelle or a fuselage side — is the number of blades multiplied by the shaft frequency, so $4 \times 20 = 80$ Hz. Add a blade and the blade-passing frequency rises to 100 Hz without the shaft speed changing at all.

Engine excitation is counted the same way but against the engine's own speed. A four-stroke piston engine fires each cylinder once every two revolutions, so a four-cylinder engine produces two firing impulses per revolution and a six-cylinder engine produces three. At 2,400 RPM — 40 rev/s — the six-cylinder engine therefore excites the propeller at 120 Hz. On a geared installation the engine and the propeller run at different speeds, so their order families are different and a resonance can be excited by either.

Knowing the order matters diagnostically, because the order points at the cause:

- **1P** is the signature of anything that is different about one blade or about one part of the disc: a mass imbalance, an aerodynamic imbalance from a mis-set or damaged blade, or the once-per-revolution loading of asymmetric inflow. A vibration that tracks propeller speed at exactly one cycle per revolution is a propeller problem until proved otherwise.
- **Blade-passing frequency** in the cabin or on the fuselage skin is normal propeller noise and pressure loading, not a fault. It changes when the RPM changes and it is the frequency that synchrophasing is designed to manage.
- **An engine order** that does not track propeller speed on a geared installation points at the engine rather than the propeller.

Why a Resonance Is So Much Worse Than a Large Force

A structure driven well away from its natural frequency responds by roughly the amount the force would deflect it statically. Driven *at* a natural frequency, each cycle of the exciting force arrives exactly in step with the motion already present and adds energy to it, so the amplitude builds up cycle after cycle until the energy being put in is balanced by the energy the damping removes. The result is a response many times larger than the same force would produce anywhere else in the frequency range — from a force that may be quite modest.

Two things limit the peak. The first is **damping**, which is small in a stiff metal structure; the second is **how long the condition is held**, because the build-up takes time. That second point is the entire logic behind a restricted RPM band: the resonance cannot always be designed out of the operating range, but it can be

made harmless by ensuring the engine passes through the band quickly rather than dwelling in it. "Transit briskly, never cruise" is not a comfort recommendation; it is a fatigue-life instruction.

How the restricted bands are established

During certification the blade's natural frequencies for each mode are plotted against rotational speed, together with straight lines representing each excitation order — 1P, 2P, 3P and so on. Wherever an order line crosses a natural-frequency line, the excitation and a natural frequency coincide and a resonance exists at that speed. This plot is the engine and propeller manufacturers' standard tool for the job, and it is how they decide which crossings are harmless, which are removed by redesign, and which have to be handled by placarding the speed range they occur in.

One consequence is worth carrying away: the blade's natural frequencies are not fixed numbers. Centrifugal force stiffens a rotating blade, so the natural frequencies **rise** with rotational speed. That is why the natural-frequency lines on the plot are curves rather than horizontal lines, and it is why a blade's frequencies measured statically on a bench are not the frequencies it has in flight.

Two Kinds of Imbalance, Two Different Cures

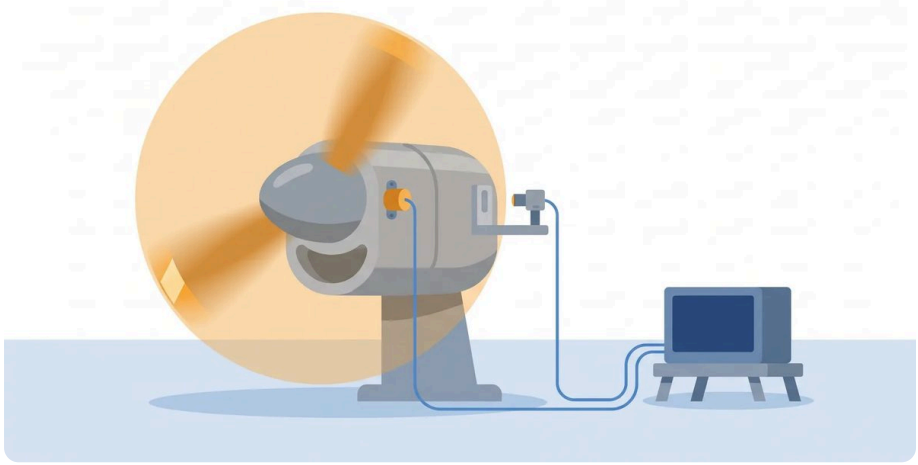
The two sources of once-per-revolution excitation that an engineer can actually do something about are mass imbalance and aerodynamic imbalance, and confusing them wastes a great deal of time on the ramp because they produce an identical symptom.

	Mass imbalance	Aerodynamic imbalance
Cause	Unequal mass or unequal mass distribution between blades — manufacturing tolerance, a repair, a paint difference, water or ice in a blade, a lost balance weight	Unequal thrust between blades — a blade angle differing from its neighbours, out-of-track, differing blade profile, erosion or damage on one blade, a damaged de-icing boot
Present when?	Whenever the propeller rotates, whether or not it is producing thrust	Only when the propeller is producing thrust, and it grows with power
Cure	Balancing — static on the bench, dynamic on the aircraft	Correct the blade angles and tracking, or replace the damaged blade. Balancing cannot fix it.

Static balancing is done off the aircraft, with the propeller assembly supported on knife edges or a balance stand so that it is free to rotate. An out-of-balance assembly always comes to rest with its heavy point downward, and balance weights are added or moved until it shows no preference for any position. This deals with the gross distribution of mass but cannot address anything that only appears when the assembly is turning.

Dynamic balancing is done with the propeller installed and running. An accelerometer on the engine or gearbox measures the vibration amplitude, while an optical or magnetic pick-up watching a reflective target on

one blade establishes where in the revolution the vibration peaks. The analyser converts the two into a magnitude and a clock position, and from these it calculates the mass and the angular location of the correction weight, which is fitted to the spinner bulkhead or to the designated balance provision on the hub. The run is repeated until the residual vibration is inside the manual's acceptance figure, which is normally expressed as a velocity in inches per second.



Dynamic balancing: an accelerometer measures vibration while an optical pick-up watches a target on one blade.

⚠ Balancing is the last step, not the first

Blade tracking, blade angle agreement between blades, blade surface condition, spinner condition and correct hub installation are all checked and corrected **before** a dynamic balance is attempted. Balancing an aircraft that has an aerodynamic imbalance simply adds a mass fault that partly cancels an aerodynamic fault at one power setting, and the cancellation falls apart as soon as the power changes — leaving an aircraft that vibrates differently in every phase of flight and a maintenance record that hides the real defect. If a propeller will not come within limits, or if the required weight is unusually large or keeps changing, stop and look for the underlying fault instead of adding metal.

One more source of a sudden, severe 1P imbalance deserves naming because it is transient and leaves no evidence: **asymmetric ice shedding**. Ice accreting on the blades is rarely identical on each one, and it does not leave in equal amounts either. A blade that sheds while its neighbours have not is instantly and heavily out of balance, and the vibration can be alarming. This is why propeller de-icing systems cycle the blades in a controlled sequence rather than heating everything at once, and why a report of severe, short-lived vibration in icing conditions is investigated as a de-icing system fault rather than as a balance problem.

Two Instabilities Beyond Ordinary Resonance

Resonance needs an external exciting force at the right frequency. Two propeller phenomena need no external force at all, because the motion itself generates the force that sustains it, and both are correspondingly more dangerous:

- **Blade flutter** is a coupling between the blade's bending and torsional motions. As the blade twists, the change in angle of attack produces an aerodynamic force that bends it; the bending changes the twist, and if the phasing is unfavourable each cycle extracts more energy from the airflow than the structure's damping can absorb. Flutter is strongly associated with high tip speeds and with reduced torsional stiffness, which is one of the reasons a blade's permitted repairs and the amount of material that may be removed are so tightly controlled — a blade thinned outside limits is a blade with less torsional stiffness than the one that was flutter-cleared.
- **Whirl flutter** is a coupling between the propeller's gyroscopic behaviour and the flexibility of the engine mounting and nacelle structure. If a disturbance sets the disc precessing on its mounts, the resulting change in inflow across the disc produces forces that can reinforce the precession, and the whole engine can be driven into a divergent whirling motion. It depends on the stiffness of the mounting as much as on the propeller, which is why engine-mount condition, mount bolt torques and the structural integrity of the nacelle are propeller-related airworthiness items and not just engine ones.



Whirl flutter depends on mount stiffness, so mount bolt torques are a propeller airworthiness item.

Designing the Vibration Down

Where a resonance cannot be moved out of the operating range, energy has to be removed from it. The measures an engineer meets on aircraft are:

- **Flexible engine mounts**, usually elastomeric, which isolate the airframe from the engine and propeller and absorb energy as they deflect. They are life-limited and they deteriorate: a hardened, cracked, oil-soaked or collapsed mount transmits far more vibration than a serviceable one, and a mount set replaced as a set is a recognised cure for a general increase in cabin vibration.
- **Dynamic dampers** — including the pendulum-type absorbers fitted to the crankshaft counterweights of some piston engines — which are tuned to a particular excitation order and oscillate out of phase with it, absorbing the torsional pulse before it reaches the propeller.
- **Tuning of the blade itself**, by placing its natural frequencies away from the strong excitation orders, and by material choice: a composite blade has different stiffness, mass distribution and internal damping from an aluminium one, so its modal behaviour is different by design rather than by accident.
- **Synchronising and synchrophasing** on multi-engine aircraft, which holds all propellers at the same RPM and, in the case of synchrophasing, at a fixed relative blade angular position, so that the pressure pulses from different propellers do not beat against one another. It is a noise and comfort system rather than a structural one, but a synchrophaser fault is a common cause of a "vibration" complaint that is really an audible beat.

What Vibration Does If It Is Left Alone

A vibration complaint is not a comfort item to be scheduled at leisure. Persistent vibration is a fatigue-loading condition applied to the whole nacelle, and its consequences appear well away from the propeller:

- Fatigue cracking in the blade, the hub, the blade retention and the propeller shaft.
- Cracked engine mounts, mount brackets, exhaust systems, cooling baffles and cowling attachments.
- Accelerated wear in the reduction gearbox and its bearings, often announced first by findings on a magnetic chip detector or in the oil filter.
- Loosening of fasteners, chafing of pipes, looms and control cables, and failures in instruments and avionics mounted in the affected structure.
- Crew fatigue and, over time, a reduced ability to notice when the vibration signature changes — which is the very cue that a developing failure produces.

The proper response to a vibration report is therefore to characterise it before touching anything: at what RPM and power does it occur, does it track propeller speed or engine speed, is it present in flight only or on the ground as well, does it change with airspeed, and did it appear suddenly or grow gradually? A vibration that appears suddenly after a ground event, or that grows steadily over a few flights, is a different investigation from one that has always been there at one particular RPM.

Measuring Vibration: Which Quantity, and Why

Vibration can be described by how far something moves, how fast it moves, or how hard it is accelerated, and the three are not interchangeable because each emphasises a different part of the frequency range.

- **Displacement** emphasises low frequencies. At a few hertz a large motion produces a large displacement reading, while a damaging high-frequency vibration barely registers.
- **Acceleration** emphasises high frequencies, which makes it the right measurement for bearing and gear defects but a poor one for a once-per-revolution propeller imbalance.
- **Velocity** sits between the two, and over the frequency range that covers rotating-machinery faults it correlates reasonably well with the energy in the vibration and hence with the damage being done.

That is why propeller balance limits are quoted as a velocity, conventionally in inches per second, and why the same analyser that measures the amplitude also measures the phase — the angular position in the revolution at which the peak occurs. Amplitude alone tells you how bad the imbalance is; phase tells you where to put the correction. Both are needed, and a reading taken without a phase reference is of very little use.

The sensor position matters as well. An accelerometer mounted at a different point on the engine or gearbox will read a different amplitude for the same imbalance, because it is measuring the response of the structure at that point rather than the force at the propeller. Manuals therefore specify the mounting location, orientation and, on some types, the sensor type, and readings from different positions are not comparable.

Trend Monitoring, and Why the Record Is the Point

A single balance reading answers only one question: is this propeller inside limits today. The record of readings over time answers a far more useful one: is this propeller getting worse, and how quickly. A propeller that has needed a slightly larger correction weight at each of its last three balances is telling a story that no individual within-limits reading tells.

The information worth capturing at every balance is the same each time: the residual amplitude before and after, the weights fitted and their positions, the conditions the run was carried out at, and the sensor location used. Kept consistently, that record turns balancing from a corrective task into a condition-monitoring one, and it is often the first place a developing blade, hub or mount problem shows itself.

Mode Shapes in a Little More Detail

The modes listed above are the fundamentals, and each has higher-order versions with additional **nodes** — points along the blade that remain still while the material either side of them moves in opposite directions. Two consequences follow that matter in practice:

- **The stress is greatest where the curvature is greatest**, not where the movement is greatest. The tip of a blade in its first flap mode moves the furthest but carries almost no bending stress; the root, which barely moves, carries the most. This is why fatigue cracks appear where the blade looks least active, and why the location of a crack is diagnostic of which mode caused it.
- **Higher modes have higher frequencies**, so they are excited by higher orders. A blade can be perfectly clear of resonance in its first flap mode across the whole RPM range and still meet a second-mode resonance with a higher engine order somewhere in the middle of it.

Edgewise modes deserve a particular mention because they are the least well damped. Motion in the plane of rotation is opposed by very little aerodynamic damping — the blade is moving edge-on to the air — whereas flapping motion is strongly damped because it changes the section's angle of attack and generates an opposing aerodynamic force. An edgewise resonance is therefore more likely to build to a damaging amplitude than a flapping one of the same excitation strength.

Propeller Efficiency

Propeller efficiency (η) is the ratio of useful thrust power to the shaft power delivered by the engine:

$$\eta = \frac{\text{Thrust} \times V_{TAS}}{\text{Shaft Power}} = \frac{T \times V}{P}$$

Typical maximum propeller efficiency: **80–88%** for well-designed propellers at their optimal operating point.

Propeller efficiency varies with **advance ratio** (J), which is the ratio of forward speed to the product of RPM and diameter:

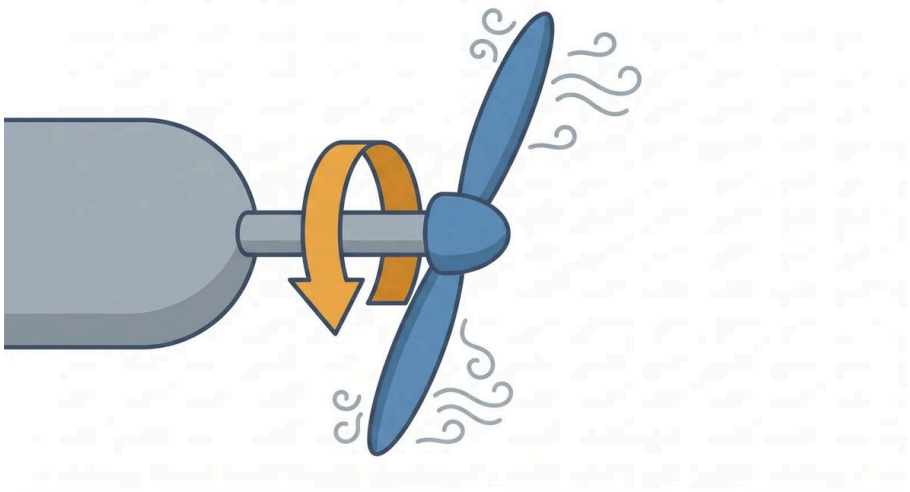
$$J = \frac{V}{nD}$$

At J = 0 (static, zero forward speed), efficiency is zero — thrust is produced but no useful work is done (no distance covered). As J increases, efficiency rises to a peak and then falls as the propeller loses its ability to produce thrust at high forward speeds. A variable-pitch propeller maintains high efficiency over a wider range of J by adjusting blade angle.

What the Ratio Is Actually Comparing

Efficiency is always output divided by input, and for a propeller both terms are rates of doing work. The **input** is the shaft work the engine puts into the propeller — torque multiplied by rotational speed. The **output** is the useful propulsive work the propeller delivers to the aircraft — thrust multiplied by the speed at which the

aircraft is moving. Everything that appears in the first and not in the second has gone into blade drag, into the rotation of the slipstream, and into the kinetic energy left behind in the accelerated air.



The engine's shaft work goes into the propeller; part of it leaks away as blade drag and disturbed air.

Two consequences fall straight out of that definition and are both examined. The ratio is always less than one, because those losses can be minimised but never eliminated. And it is not a comparison of speeds, or of pitches, or of anything measured at the constant-speed unit — it is specifically *useful work done by the propeller* divided by *work done by the engine on the propeller*. That is also the clearest single statement of what a propeller is for: it is a device for converting engine shaft power into thrust, and the efficiency is simply how well it does that one job.

An efficiency calculated from end to end

A propeller absorbs 1,000 kW of shaft power and produces 8,000 N of thrust at a true airspeed of 100 m/s.

$$\eta = \frac{T \times V}{P} = \frac{8000 \times 100}{1{,}000{,}000} = 0.80$$

Thrust power delivered is 800 kW; 200 kW has been lost inside the propeller. Watch the units: thrust in newtons times speed in metres per second gives watts directly, so the only thing that can go wrong is comparing watts with kilowatts.

Reading the Efficiency Curve

Plotting efficiency against advance ratio produces the single most informative picture in propeller theory. For a propeller held at a **fixed blade angle**, the curve starts at zero, climbs to a peak, and then falls away steeply back to zero. Understanding why it does that at each end explains almost every propeller design decision:

- **At J = 0 the efficiency is zero**, and this is not because the propeller is producing no thrust — thrust is actually at its maximum at zero forward speed. It is because efficiency measures useful *work*, and work requires the force to move something through a distance. The aircraft is not moving, so no propulsive work is being done, however hard the propeller pulls.
- **Below the peak** the aircraft is slow for the blade angle set. The helix angle is small, the angle of attack is large, blade drag is high, parts of the blade may be stalled, and the slipstream is being given a large velocity increment — so the induced and swirl losses are at their worst. The propeller is producing plenty of thrust and wasting a large fraction of the power doing it.
- **At the peak** the blade sections across most of the span are working near their best lift-to-drag angle, and the induced losses have fallen because the propeller no longer has to push the air so hard.
- **Above the peak** the aircraft is fast for the blade angle set. The helix angle has grown, the angle of attack has shrunk, and the thrust falls faster than the speed rises. Eventually the angle of attack reaches the value at which the blade produces no net thrust, the numerator of the efficiency goes to zero, and the curve returns to the axis. Past that point the propeller is windmilling.

A fixed-pitch propeller lives on one such curve and can only be at its peak in one flight condition. A constant-speed propeller has a whole family of curves — one for every blade angle — and, by choosing the blade angle, rides along the **envelope** of their peaks. That is the entire performance case for variable pitch, and it is why a constant-speed propeller maintains high efficiency across a wide range of advance ratios while a fixed-pitch one cannot.

Peak efficiency is not the whole story

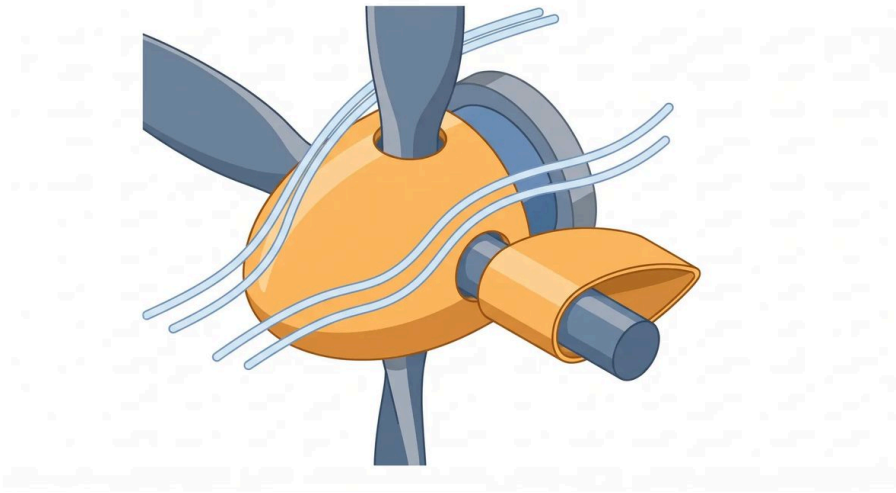
The best peak efficiency a fixed-pitch propeller reaches can be as high as the peak of a constant-speed one — the constant-speed unit's advantage is not a higher maximum but a **flatter** characteristic across the speed range. A well-chosen fixed-pitch propeller in its own design cruise can be very nearly as efficient as a constant-speed propeller at the same condition. Everywhere else it is far worse. When comparing propellers, always ask over what range the figure is being quoted, not just how large it is.

Where the Missing Twenty Per Cent Goes

Loss	Mechanism	Worst when	Reduced by
Profile drag	The blade sections' own drag, resolved into the plane of rotation as torque	Always present; grows with the square of local velocity	Good aerofoils, smooth clean blades, correct angle of attack
Induced (slipstream) loss	Kinetic energy left in the air that has been accelerated rearward	Low forward speed, high disc loading	A larger disc, so more air is moved more gently

Loss	Mechanism	Worst when	Reduced by
Swirl loss	Rotational kinetic energy given to the slipstream by the blade torque	High torque at low forward speed	Lower disc loading; recovered by a contra-rotating second disc
Tip loss	Flow spilling round the tip from the high-pressure face, rolling into a vortex	Heavily loaded tips, few blades	Tip shaping, unloading the tip, more blades sharing the load
Compressibility loss	Shock formation on the outer blade as the helical tip Mach number rises	High RPM, high TAS, cold air	Lower tip speed, thin swept tip sections, reduction gearing
Installation loss	Blockage by spinner, nacelle and cowling; slipstream drag over the airframe; cooling airflow taken from the disc	Bulky nacelles, poor spinner fit, large tractor installations	Well-faired spinner and cowling, correct blade cuffs, clean nacelle lines

The last row is the one an engineer can most directly affect, and it is often overlooked. The spinner is not a decorative fairing: it smooths the flow into the root of the disc, where the blade shank produces drag and no thrust, and it directs cooling air where the installation needs it. A dented, ill-fitting, cracked or incorrectly indexed spinner disturbs that flow, costs performance, and — because it rotates with the propeller — can itself become an imbalance. Blade **cuffs**, the aerofoil-shaped fairings fitted over the round shank of some blades, do the same job further out: they turn a structural cylinder into something that at least resembles an aerofoil and helps duct cooling air.



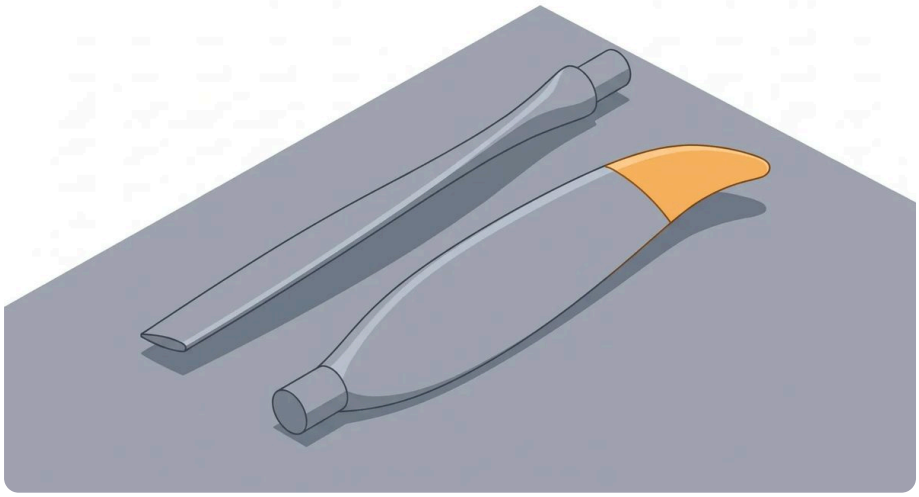
The spinner smooths flow into the root of the disc; a blade cuff wraps the round shank in an aerofoil shape.

Diameter, Blade Count and the Compromises Between Them

Momentum theory says the way to a high efficiency is a large disc gently loaded. Real propellers are not as large as that argument would like, and it is worth being clear about what stops them:

- **Tip speed.** For a given RPM, the tip velocity is directly proportional to the radius, so diameter and rotational speed trade against one another against a fixed Mach limit. A bigger propeller must turn more slowly.
- **Ground clearance** and, on a multi-engine aircraft, fuselage clearance, which sets a hard geometric ceiling.
- **Blade weight and root stress**, both of which grow rapidly with radius under the square law.

When diameter is capped and the power still has to be absorbed, the designer adds blade area — wider blades, more blades, or both. That works, but it is not free. Each blade operates partly in the disturbed flow of the blade ahead of it, so as blade count rises the efficiency of each individual blade falls slightly; the gain is that the total thrust is produced with a lower loading per blade, a lower tip speed for the same power, and markedly less noise. This is the trade behind the appearance of modern turboprops, which carry many relatively broad, slow-turning, swept-tipped blades where an older design of the same power would have used three or four long, slender ones turning faster.



Older designs used long, slender blades; modern turboprops use broader, slow-turning blades with swept tips.

Why the Propeller Wins at Low Speed and Loses at High Speed

The momentum argument earlier in this note explains the first half. A propeller handles a very large mass flow and gives it a small velocity increment, so its ideal propulsive efficiency at low forward speed is high — much higher than a jet, which handles a small mass flow and gives it a very large increment. That is why a propeller-driven aircraft accelerates and climbs so strongly for its installed power, and why turboprops dominate short-sector and short-field operations.

The second half is the tip. As the aircraft's speed rises, the forward-speed contribution to the helical tip velocity rises with it, and the tip Mach number climbs whether or not the RPM has been touched. Somewhere in the high subsonic range the tip drag rise overwhelms the propeller's propulsive advantage, and the aircraft is better served by a fan enclosed in a duct where the incoming flow can be slowed before it meets the blades.

Conventional propellers therefore become progressively less attractive as cruise Mach number rises through the upper subsonic range, which is exactly the region where turbofans take over. It is not a limitation of the propeller principle but of the unshrouded blade tip, which is why open-rotor research concentrates on tip shaping and on the interaction between contra-rotating discs rather than on abandoning the propeller.

Shaft Power, Thrust Power and Equivalent Power

Three power figures circulate around a turboprop installation and they are not the same number:

- **Shaft power** is what the engine delivers to the propeller shaft. It is what the torquemeter and the RPM together measure, and it is the input term in the efficiency ratio.
- **Thrust power** is shaft power multiplied by propeller efficiency — the useful output. It is zero on the ground however much shaft power is being produced.
- **Equivalent shaft power** accounts for the fact that a turboprop's exhaust also produces a small amount of residual jet thrust. It adds the shaft power to a power-equivalent of that jet thrust, using a conversion stated in the engine's own performance data, so that installations with different amounts of residual thrust can be compared on one scale.

Keeping them apart matters when reading performance data or interpreting a power check. A power assurance run measures the engine's ability to produce shaft power at a defined condition; it tells you nothing about how efficiently the propeller is converting that power into thrust. A propeller fault shows itself as an aircraft that cannot achieve its book performance despite an engine that passes its power check — and that combination is one of the more useful diagnostic pairings on a turboprop.

The Non-Dimensional Coefficients

Propeller performance data is published in non-dimensional form, so that a single chart covers a whole family of sizes, speeds and altitudes. Three coefficients do the work, and all of them are built from the same ingredients: air density, rotational speed and diameter.

Thrust coefficient, power coefficient and their relationship to efficiency

$$C_T = \frac{T}{\rho n^2 D^4} \qquad C_P = \frac{P}{\rho n^3 D^5} \qquad J = \frac{V}{nD}$$

$$\eta = \frac{TV}{P} = \frac{C_T \rho n^2 D^4 \times JnD}{C_P \rho n^3 D^5} = J\,\frac{C_T}{C_P}$$

In each case *n* is in revolutions per second and *D* in metres. The whole of propeller performance can then be plotted as C_T, C_P and η against J, with one curve per blade angle.

The coefficients for the worked cruise case

The 2.5 m propeller at 1,200 RPM ($n = 20$ rev/s), 100 m/s, sea level, absorbing 1,500 kW and producing 12,000 N of thrust:

- $\rho n^2 D^4 = 1.225 \times 400 \times 39.06 = 19{,}141$, so $C_T = 12{,}000/19{,}141 = 0.627$.
- $\rho n^3 D^5 = 1.225 \times 8000 \times 97.66 = 957{,}000$, so $C_P = 1{,}500{,}000/957{,}000 = 1.567$.
- $J = 100/(20 \times 2.5) = 2.00$.
- $\eta = J \times C_T/C_P = 2.00 \times 0.627/1.567 = 0.80$.

Eighty per cent, which is the same answer the direct TV/P calculation gives. That agreement is the point of the exercise: the coefficients are not a different theory, they are the same numbers expressed so that they can be carried from one propeller to another.

The practical use of the coefficient form is that it makes the effect of density explicit. Thrust and power both scale directly with air density at a given advance ratio and blade angle, so climbing to altitude reduces both in the same proportion, leaving the efficiency — a ratio — largely unaffected. That is why a propeller's efficiency does not collapse with altitude even though its thrust does.

Efficiency and Fuel

Propeller efficiency multiplies straight into the aircraft's fuel consumption, and this is where the difference between a well-maintained and a neglected propeller becomes a commercial figure rather than an engineering one. The engine's fuel consumption is set by the *shaft* power it is producing; the aircraft's progress is set by the *thrust* power the propeller delivers. Anything that lowers the efficiency raises the shaft power needed for the same cruise speed, and therefore raises the fuel burn for every hour of the aircraft's remaining life.

The items on a propeller that affect this are unglamorous and entirely within an engineer's control: correct blade angles, blades in track, an undamaged and correctly profiled leading edge, a smooth and properly finished blade surface, de-icing boots that are bonded and not lifting, an undamaged and correctly fitted spinner, and clean blades. None of them will ground an aircraft on its own, and all of them are worth money continuously.

Reading a Propeller Performance Chart

A typical chart carries efficiency on the vertical axis, advance ratio along the horizontal, and a fan of curves each labelled with a blade angle at the reference station. Using it follows a fixed sequence:

- Calculate the advance ratio from the true airspeed, the rotational speed in revolutions per second and the diameter.

- Enter the chart at that advance ratio and read up to the curve for the blade angle in use, or — for a constant-speed propeller — to the envelope that touches the peak of each curve.
- Read the efficiency across, and combine it with the shaft power to obtain thrust power, and hence thrust once divided by the true airspeed.

Two habits prevent most errors. First, keep the advance ratio dimensionless by checking that the units cancel: metres per second divided by (revolutions per second times metres) leaves nothing behind. Second, remember that the chart is for the propeller alone; the installed efficiency is lower, because the spinner, nacelle and cowling are not in the chart and the propeller in the wind tunnel had none of them.