

FIRST EDITION · OCTOBER 2026

Naple's Basic Guide to Semiconductors

Physics, mathematics and applications · Economics · The global market

From Grade 10 to varsity and professional level, with worked examples, diagrams, practice questions, a South African case study and links to the best free resources.

GRADE 10 GRADE 11 GRADE 12 VARSITY PROFESSIONAL



Naple's Basic Guide to Semiconductors

First edition, October 2026. Published by Naples Semiconductor, part of Naples Corporation, Durban, KwaZulu-Natal, South Africa.

Editor and founder: Esethu Tshintsholo. Written and illustrated by Naples Semiconductor with AI research and writing assistance (Claude, by Anthropic). All calculations and plots were generated in code for this edition.

About the numbers. Physical constants and material values are textbook values at 300 K unless stated. Industry figures carry their source and date; most semiconductor cost data is proprietary, so public figures are often analyst estimates and are labelled as such. Markets and policies change quickly: check the latest sources before relying on any industry or policy figure.

Links. Online resources were checked on 15 October 2026. Links can change.

Safety. Practical activities in this book use batteries and low voltages only. Never work on mains (230 V) wiring, solar arrays or inverters unless you are qualified and supervised.

© 2026 Naples Semiconductor. This guide may be shared freely for non-commercial educational use with attribution.

How to Use This Guide

This book is built to grow with you. A Grade 10 learner can start on page one; a working engineer can jump straight to the professional sections. Here is how it is organised.

Three parts, one story

PART 1 · PHYSICS & MATHS

Why silicon conducts, how junctions and transistors work, how chips are made and designed, ending with a real South African chip project.

PART 2 · ECONOMICS

Fixed costs, yield, the cost of a good chip, fabs, design costs, break-even and business models.

PART 3 · THE MARKET

Market size, leading companies, supply chains, geopolitics, AI, and where South Africa and you fit in.

Level badges

Every chapter and section has a badge showing the level it is written for. Read what matches your level, and come back to the harder sections later. You will not get lost: each chapter starts simple and ends with a summary in plain words.

Badge	Who it is for	Assumes
G10	Grade 10 and beginners	Basic algebra, atoms, simple circuits
G11	Grade 11	Exponents, logarithms (introduced), chemical bonding
G12	Grade 12	Photoelectric effect, exponential functions, electrodynamics
Varsity	University students and serious self-learners	Calculus, first-year physics
Pro	Engineers, analysts, founders	Device physics or industry background

The boxes

KEY IDEA

The one thing to remember from a section.

WORKED EXAMPLE

A full calculation, step by step, with units.

TRY IT

Something to do with a phone, a simulator or simple parts.

COMMON MISTAKE

A trap that catches many learners, and the fix.

GOING DEEPER

Extra depth for higher levels. Skip it on first reading if you like.

SOUTH AFRICAN CONTEXT

How the idea connects to life and industry here.

ABOUT THESE NUMBERS

Where a figure comes from, when, and how reliable it is.

WATCH & EXPLORE

► Checked videos, simulators and free reading, with printed links.

Each chapter ends with **Check yourself** questions (answers at the back), **Watch & explore** links, an **In plain words** summary and its **Sources**.

Labels on numbers

Industry figures carry a label so you know how much to trust them: **PUBLISHED** (official statistics or company reports), **ANALYST ESTIMATE** or **FORECAST** (an outside estimate or prediction), **COMPANY CLAIM** (a company's own statement or plan, not independently checked), and **DERIVED** (calculated in this book). Physics constants are standard values at 300 K (27 °C) unless stated.

Suggested routes

If you are...	Read
a Grade 10–11 learner	Chapters 1–5, 8 (first half), 12, 15, 17, and the "In plain words" boxes everywhere
a Grade 12 learner	Chapters 1–8, 10–13, 15, 17
a varsity student	Everything; work every Varsity question
a professional or founder	Chapters 9–11, 13–16, and the Pro sections

What is inside

• The big picture: chips are everywhere	6
PART 1 · PHYSICS & MATHS OF SEMICONDUCTORS AND THEIR APPLICATIONS	7
01 What Is a Semiconductor? GRADE 10	8
02 Atoms, Bonds and the Silicon Crystal GRADE 10 GRADE 11	13
03 Electricity essentials GRADE 10 GRADE 11	17
04 Energy Bands and the Band Gap GRADE 11 GRADE 12 VARSITY	28
05 Electrons, Holes and Doping GRADE 11 GRADE 12 VARSITY	33
06 How Charges Move: Drift and Diffusion GRADE 12 VARSITY	38
07 The p–n Junction: Diodes, LEDs and Solar Cells GRADE 12 VARSITY	43
08 The Transistor: From Switch to Logic and Memory GRADE 12 VARSITY	49
09 Inside the MOSFET: Physics, Limits and Scaling VARSITY PROFESSIONAL	54
10 How Chips Are Made and Designed GRADE 12 VARSITY PROFESSIONAL	59
11 Case Study: Raven-01, a Chip to Warn of Sewer Blockages GRADE 12 VARSITY PROFESSIONAL	65
PART 2 · THE ECONOMICS OF SEMICONDUCTORS	71
12 Why Chips Are an Economics Story GRADE 10 GRADE 11 GRADE 12	72
13 The Cost of a Chip: Wafers, Dies and Yield GRADE 12 VARSITY	76
14 Fabs, Design Costs and Business Models VARSITY PROFESSIONAL	80
PART 3 · THE SEMICONDUCTOR MARKET	85
15 The Global Chip Market and Who Leads It GRADE 11 GRADE 12 VARSITY	86
16 Supply Chains, Geopolitics and the AI Boom GRADE 12 VARSITY PROFESSIONAL	90
17 South Africa, Africa and Your Career in Chips ALL LEVELS	94
BACK OF THE BOOK	
• Resource library: videos, courses, books and simulators	99
• Formula sheet	101
• Glossary	103
• Answers to Check Yourself	105
• About Naple Semiconductor	109

The Big Picture: From Sand to Smartphone

Before the detail, here is the whole story on one page: how a grain of sand becomes the brain of your phone, and where each chapter fits.

PART 1 · PHYSICS: WHY IT WORKS

Atoms & bonds

Ch. 1–3
silicon, circuits

Bands & doping

Ch. 4–6
electrons, holes

Junctions

Ch. 7
diodes, LEDs

Transistors

Ch. 8–9
logic, scaling

Make & design

Ch. 10
fabs, Verilog

Case study

Ch. 11
Raven-01

PART 2 · ECONOMICS: WHAT GETS BUILT

Scale & learning

Ch. 12

Yield & cost per chip

Ch. 13

Fabs, NRE, business models

Ch. 14

PART 3 · THE MARKET: WHO, WHERE, WHAT NEXT

Market & companies

Ch. 15

Supply chain, politics, AI

Ch. 16

South Africa & your career

Ch. 17

Physics sets what is possible. Economics decides what is built. The market decides who builds it, and where.

Figure 0.1 The map of this book.

Five numbers to remember

Number	What it is
1.12 eV	Silicon's band gap at 300 K (Ch. 4)
10^{10} cm^{-3}	Free electrons in pure silicon at room temperature (Ch. 5)
~60 mV/decade	Best a normal transistor can switch at room temperature (Ch. 9)
13.5 nm	Wavelength of EUV light that prints today's chips (Ch. 10)
US\$ 792 billion	World chip sales in 2025 (SIA; Ch. 15)

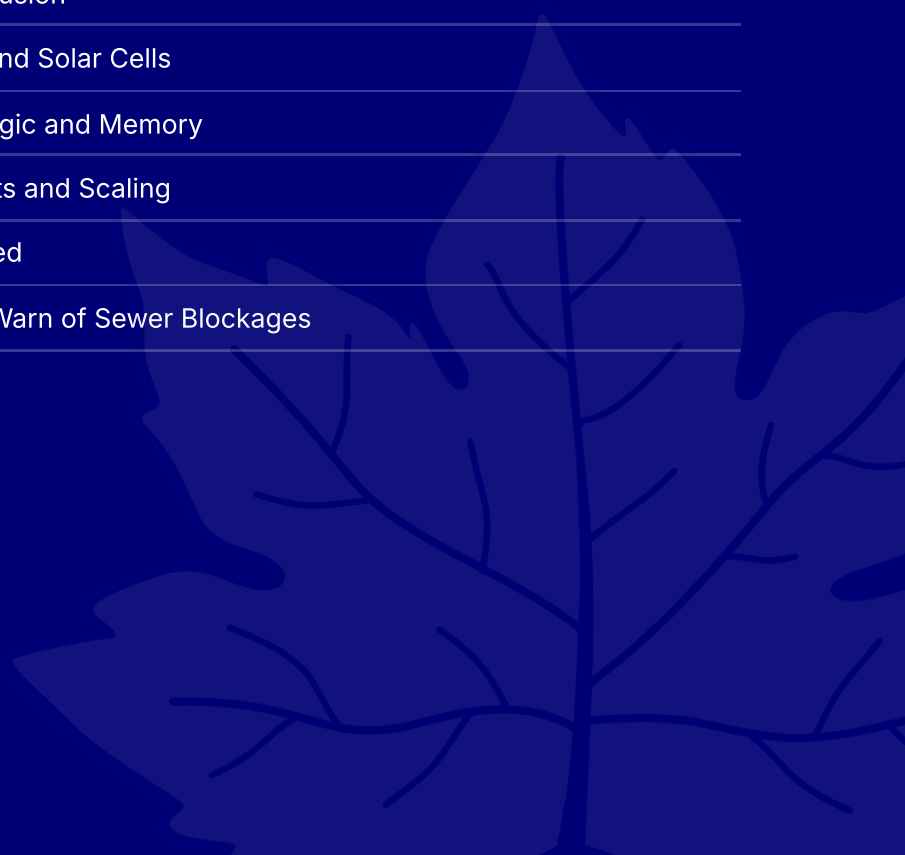
IN PLAIN WORDS

Chips are made from silicon, which comes from sand. Physics explains how they work. Money decides which chips get made. A few big companies and countries lead the industry today, but anyone who learns the skills can take part.

PART 1

Physics & Maths of Semiconductors and Their Applications

From why a piece of silicon conducts at all, to how a transistor switches, how chips are made, and how a real chip is designed. Starts at Grade 10 and climbs to professional level.

- 01 What Is a Semiconductor?
 - 02 Atoms, Bonds and the Silicon Crystal
 - 03 Electricity essentials
 - 04 Energy Bands and the Band Gap
 - 05 Electrons, Holes and Doping
 - 06 How Charges Move: Drift and Diffusion
 - 07 The p-n Junction: Diodes, LEDs and Solar Cells
 - 08 The Transistor: From Switch to Logic and Memory
 - 09 Inside the MOSFET: Physics, Limits and Scaling
 - 10 How Chips Are Made and Designed
 - 11 Case Study: Raven-01, a Chip to Warn of Sewer Blockages
- 

What Is a Semiconductor?

GRADE 10

Your phone holds more switches than there are people on Earth. Every one of them is built from a material that is neither a good conductor nor a good insulator: a semiconductor. This chapter explains what that means and why it matters so much.

IN THIS CHAPTER YOU WILL LEARN

- how conductors, insulators and semiconductors differ
- what resistivity is and how to compare materials
- the four "control knobs" that make semiconductors useful
- why a switch is the most important device ever made
- where chips hide in everyday South African life
- a short history, from crystals to AI chips

Three kinds of material G10

In Grade 10 Physical Sciences you sort materials by how well they let electric charge flow. A **conductor** (copper, aluminium, silver) lets charge flow easily. An **insulator** (glass, rubber, plastic) almost stops it. A **semiconductor** sits in between, and that "in between" is not the interesting part. The interesting part is that **we can control how well a semiconductor conducts**, very precisely and very quickly.

The most important semiconductor is **silicon** (symbol Si), the second most common element in the Earth's crust after oxygen. Sand and quartz are mostly silicon dioxide (SiO₂). Other semiconductors you will meet in this book are germanium (Ge), gallium arsenide (GaAs), gallium nitride (GaN) and silicon carbide (SiC).

Comparing materials fairly: resistivity

A long thin wire has more resistance than a short thick one, even if both are copper. To compare *materials* rather than particular wires, we use **resistivity**, symbol ρ (the Greek letter "rho"). It is defined by

$$R = \rho \frac{L}{A} \quad (1.1)$$

where R is resistance in ohms (Ω), L is the length in metres and A the cross-sectional area in square metres. Rearranged, $\rho = RA/L$, so its unit is the **ohm-metre** ($\Omega \cdot \text{m}$). A low resistivity means a good conductor.

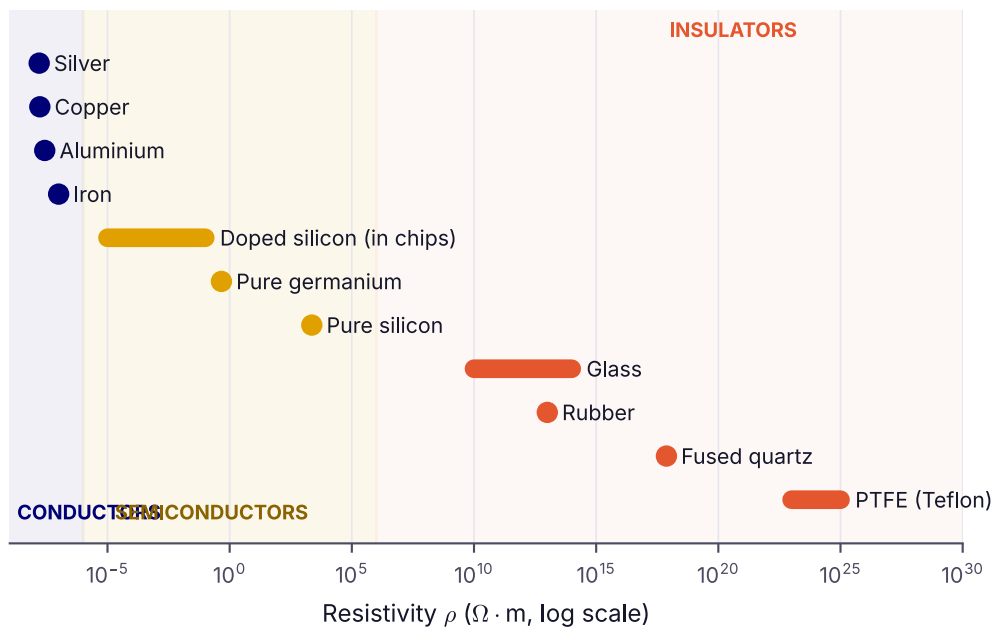


Figure 1.1 Approximate room-temperature resistivities. The axis is logarithmic: every tick is a factor of 10^3 . Metals and the best insulators differ by about 30 powers of ten. Silicon's resistivity can be moved across a huge range just by adding tiny amounts of other atoms ("doping"), which is why one material can be both the wires and the switches of a chip. Values are typical handbook figures and vary with purity and temperature.

WORKED EXAMPLE Copper wire versus a silicon bar of the same size

Question. Find the resistance of a bar 1 cm long with a $1 \text{ mm} \times 1 \text{ mm}$ cross-section, made of (a) copper, $\rho = 1.7 \times 10^{-8} \Omega \cdot m$, and (b) pure silicon, $\rho \approx 2.3 \times 10^3 \Omega \cdot m$.

Step 1. Convert to SI: $L = 0.01 \text{ m}$, $A = (0.001)^2 = 1 \times 10^{-6} \text{ m}^2$, so $L/A = 1 \times 10^4 \text{ m}^{-1}$.

Step 2. Copper: $R = 1.7 \times 10^{-8} \times 10^4 = 1.7 \times 10^{-4} \Omega$, a tiny fraction of an ohm.

Step 3. Pure silicon: $R = 2.3 \times 10^3 \times 10^4 = 2.3 \times 10^7 \Omega$, about 23 million ohms.

Answer. Same shape, but silicon's resistance is about 10^{11} times bigger. Add the right impurity at a level of one atom in a million and silicon's resistivity drops to roughly $0.05 \Omega \cdot m$: the bar becomes about 500Ω .

Four control knobs G10

What makes semiconductors special is that their conductivity responds strongly to four things. Each knob is the basis of a whole family of devices.

Table 1.1 How we control a semiconductor, and what we build with each knob

Knob	What happens	Devices it gives us	Chapter
Doping (adding impurity atoms)	Adds free charges; conductivity can rise more than a million times	Every chip, diode and transistor	5
Electric field (a voltage on a nearby plate)	Pulls charges into or out of a thin layer	The transistor (MOSFET), memory	8, 9
Light	Light can free charges; charges can give off light	Solar cells, LEDs, camera sensors, lasers	7
Temperature	Warmer means more free charges, so lower resistance	Thermistors, temperature sensors	5

The switch: the idea behind every chip G10

A light switch has two states: on and off. Computers store and process information as **binary digits** (bits), 1 and 0, which map neatly onto on and off. If you can build a switch that is **tiny**, **fast**, **cheap**, uses **very little energy** and can be **controlled by another switch**, you can build a computer. The semiconductor **transistor** is exactly that switch, controlled by a voltage instead of a finger.

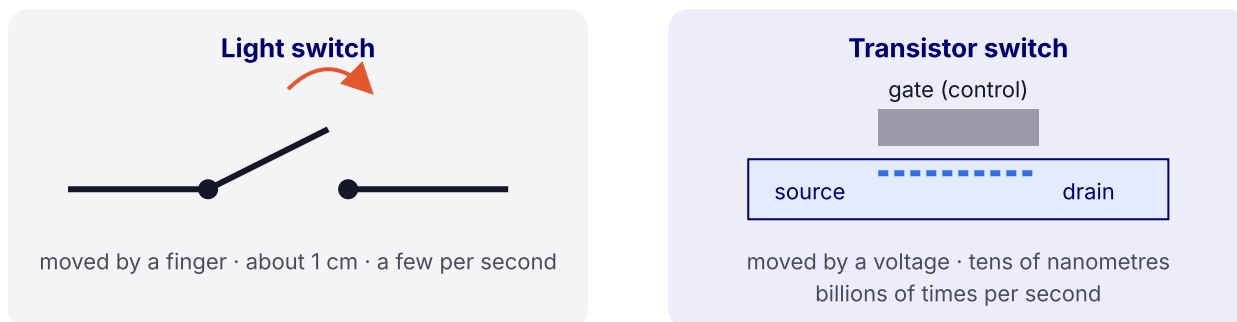


Figure 1.2 The same idea at two scales. A transistor is a switch with no moving parts: a voltage on the gate opens or closes a conducting channel (dashed blue) between source and drain. Chapter 8 explains how.

KEY IDEA

A semiconductor is useful because its conductivity can be **controlled**: by doping, by an electric field, by light and by temperature. Controlling conduction with a voltage gives us the transistor, a switch with no moving parts. Billions of transistors working together make a chip.

Chips in your day G10

A **chip** (or **integrated circuit**, IC) is a small piece of silicon, usually a few millimetres across, carrying many transistors and their wiring, sealed in a plastic or ceramic **package** with metal pins or balls.

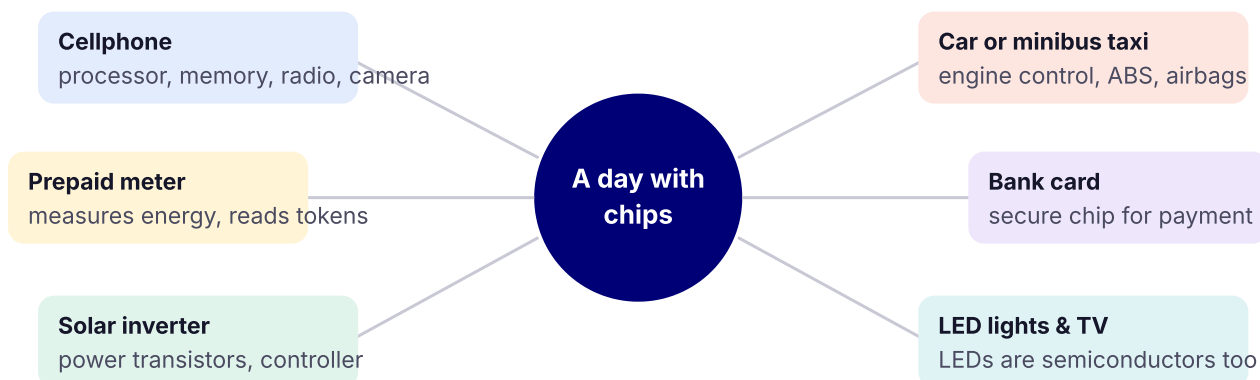


Figure 1.3 Chips are not only in computers. Anything that measures, decides, communicates, stores information or controls power almost certainly contains semiconductors.

SOUTH AFRICAN CONTEXT

During load-shedding, many households bought inverters and solar panels. Both are semiconductor products. A solar panel is a large set of silicon diodes that turn light into current (Chapter 7). An inverter uses power transistors switching thousands of times a second to turn the battery's direct current into the 230 V, 50 Hz alternating current your appliances expect.

A short history G10

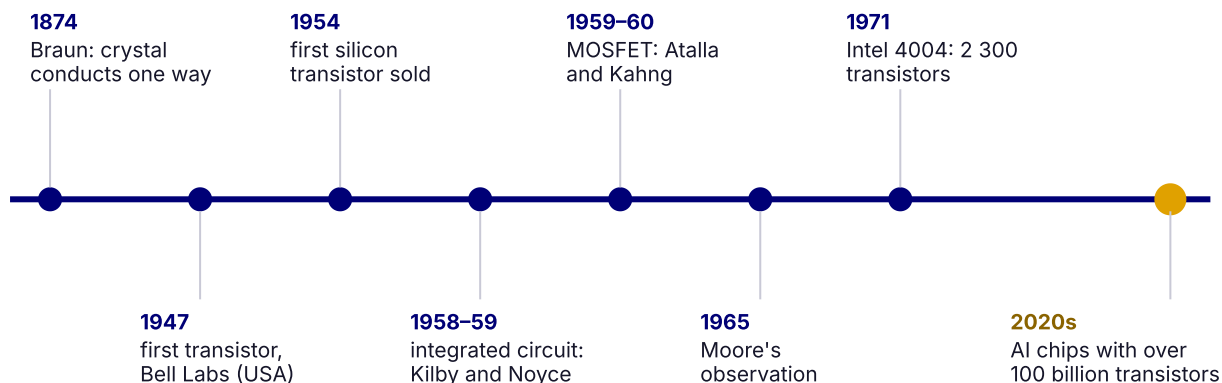


Figure 1.4 Key moments. In about 50 years the number of transistors on one chip grew from thousands to hundreds of billions (Chapter 9 plots this "Moore's law").

The first transistor (1947, by John Bardeen, Walter Brattain and William Shockley at Bell Labs) was made of germanium and was about the size of a thumbnail. The **integrated circuit**, many components made together on one piece of semiconductor, came in 1958–59 from Jack Kilby (Texas Instruments) and Robert Noyce (Fairchild). The **MOSFET**, the type of transistor in almost every chip today, was demonstrated by Mohamed Atalla and Dawon Kahng at Bell Labs in 1959–60.

COMMON MISTAKE

"Semiconductors are just bad conductors." No. A poor conductor such as a thin nichrome wire just has high resistance. A semiconductor's conductivity can be *changed* by doping, light, heat and electric fields. That controllability is the whole point.

CHECK YOURSELF

- G10 Name three conductors, three insulators and two semiconductors.
- G10 What is the SI unit of resistivity? Show how it follows from $R = \rho L/A$.
- G10 A copper wire is 2 m long with cross-section 0.5 mm^2 . Find its resistance ($\rho = 1.7 \times 10^{-8} \Omega \cdot \text{m}$).
- G10 List the four "control knobs" of a semiconductor and one device that uses each.
- G10 Why is a switch enough to build a computer?
- G11 Using Figure 1.1, roughly how many orders of magnitude separate copper from glass?

WATCH & EXPLORE

- [Circuit Construction Kit: DC](#) — PhET (University of Colorado). Build circuits with conductors and insulators, then measure current and voltage. Free, runs in a browser. phet.colorado.edu/en/simulations/circuit-construction-kit-dc **G10**
- [Siyavula Physical Sciences](#) — free CAPS-aligned textbooks and practice for Grades 10–12. Read the Grade 10 chapters on matter and on electric circuits. siyavula.com **G10**
- ▶ [The Ridiculous Engineering of the World's Most Important Machine](#) — Veritasium. The story of the machines that print modern chips; watch it now for the "wow", and again after Chapter 10. youtube.com/watch?v=M1UHjLxm3V0 **G10**
- ▶ [Branch Education](#) — YouTube channel with animated explainers. Search the channel for "How are Microchips Made?". youtube.com/@BranchEducation **G11**

IN PLAIN WORDS

Some materials let electricity through easily (metals). Some block it (glass, rubber). Semiconductors, mainly silicon from sand, are in between. Their big trick is that we can change how well they conduct. We do it by adding a few other atoms, by shining light, by heating them, or by putting a voltage near them. Using a voltage to turn conduction on and off gives us a tiny switch called a transistor. Chips are billions of these switches on a small piece of silicon. They are in phones, meters, cars, solar inverters and bank cards.

SOURCES

1. Resistivity values: typical handbook values at about 20 °C, e.g. D. R. Lide (ed.), *CRC Handbook of Chemistry and Physics*, CRC Press; pure-silicon value from S. M. Sze & K. K. Ng, *Physics of Semiconductor Devices*, 3rd ed., Wiley, 2007.
2. History: Computer History Museum, "The Silicon Engine" timeline (computerhistory.org/siliconengine); Sze & Ng (2007), ch. 1.
3. Intel 4004 transistor count: Intel (company-stated).

Atoms, Bonds and the Silicon Crystal

GRADE 10

GRADE 11

To understand why silicon behaves the way it does, zoom in until you can see atoms. The answer sits in four outer electrons and the way silicon atoms share them.

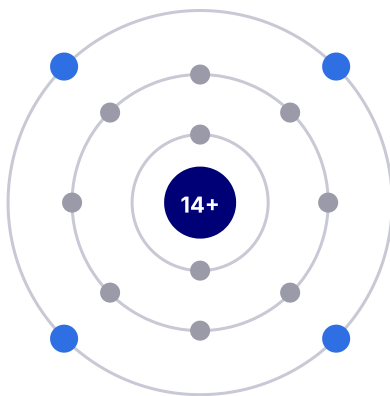
IN THIS CHAPTER YOU WILL LEARN

- the structure of the silicon atom
- why group 14 of the periodic table matters
- how covalent bonds hold a silicon crystal together
- how many atoms are in a cubic centimetre of silicon
- what happens when a bond breaks: electrons and holes
- crystalline, polycrystalline and amorphous silicon

Inside the silicon atom G10

An atom has a small, heavy, positive **nucleus** (protons and neutrons) surrounded by negative **electrons**. Silicon has atomic number 14: 14 protons and, in a neutral atom, 14 electrons. In the simple shell model from Grade 10 the electrons fill shells as 2, 8, 4.

The outer shell holds the **valence electrons**. They decide how an atom bonds and conducts. Silicon has **four** valence electrons. Its outer shell would be "full" with eight, so it is exactly half full. That is why silicon prefers to **share** electrons rather than give them away (like a metal) or grab them (like chlorine).



Silicon, Si • atomic number 14

Shell 1: 2 electrons (grey)

Shell 2: 8 electrons (grey)

Shell 3: 4 valence electrons (blue)

Four outer electrons → four bonds.

Each bond is a shared pair of electrons.

Simple model: real electrons form orbitals.

Figure 2.1 The Bohr-style shell model of silicon. Only the four blue valence electrons take part in bonding and conduction. The inner ten stay tightly bound to the nucleus.

The neighbourhood: groups 13, 14 and 15 G10

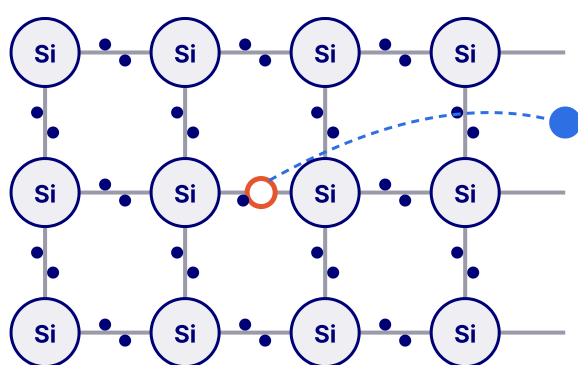
Elements in the same **group** (column) of the periodic table have the same number of valence electrons. Three groups matter most for semiconductors.

Group 13 · 3 valence e ⁻	Group 14 · 4 valence e ⁻	Group 15 · 5 valence e ⁻
B boron	C carbon (diamond)	N nitrogen
Al aluminium	Si silicon	P phosphorus
Ga gallium	Ge germanium	As arsenic
p-type dopants	elemental semiconductors	n-type dopants

Figure 2.2 Silicon's neighbours. Group 14 elements form semiconductor crystals on their own. Adding a group 15 atom gives an extra electron (n-type); adding a group 13 atom leaves a missing electron (p-type), as Chapter 5 explains. Pairing a group 13 with a group 15 element, such as Ga + As or Ga + N, makes "III-V" compound semiconductors used in LEDs, lasers and fast radio chips.

Sharing electrons: covalent bonds and the crystal G10 G11

In Grade 11 you learn that a **covalent bond** is a pair of electrons shared between two atoms. Each silicon atom shares its four valence electrons with four neighbours, and each neighbour shares one back. Every atom then "sees" eight outer electrons, a stable arrangement. Repeating this in three dimensions gives a perfectly regular **crystal**: each atom sits at the centre of a tetrahedron formed by its four neighbours. This is the same **diamond structure** as carbon in a diamond.



2-D bond model

Grey lines: covalent bonds.

Dark dots: the two shared electrons in each bond.

Free electron (blue):

heat or light knocked it out of a bond.
It can now wander and carry current.

Hole (red ring):

the empty place it left behind.
It behaves like a positive charge.

Figure 2.3 The silicon crystal flattened into two dimensions. Every atom has four bonds of two shared electrons. Breaking one bond creates **two** charge carriers at once: a free electron and a hole. This is the starting point of all semiconductor physics.

How many atoms? A first real calculation

The repeating unit of the silicon crystal, the **unit cell**, is a cube of side $a = 0.5431$ nm (the **lattice constant**) that contains 8 atoms. The distance between neighbouring atoms is $\sqrt{3}a/4 \approx 0.235$ nm.

(a) Unit cell of silicon

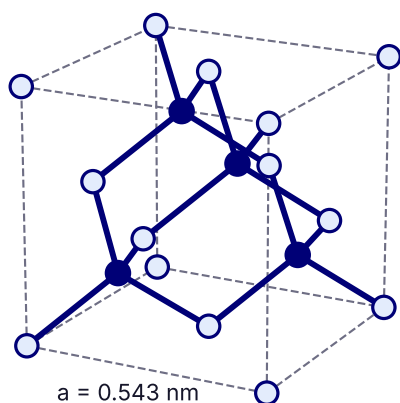
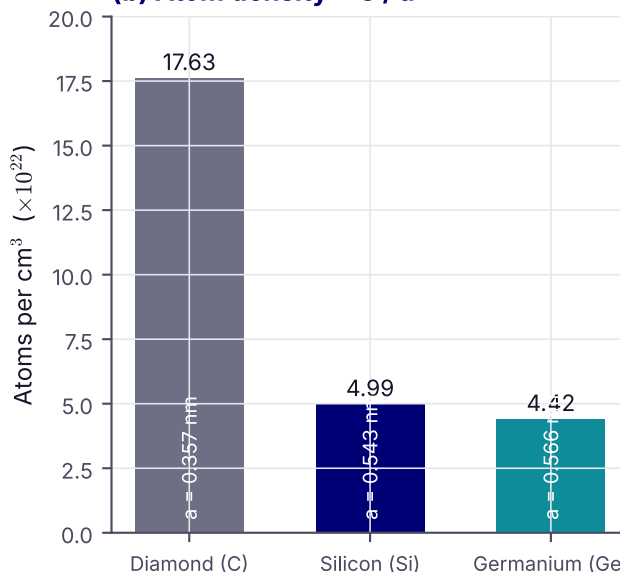
(b) Atom density = $8 / a^3$ 

Figure 2.4 (a) The cubic unit cell of the diamond structure: filled circles are the four atoms fully inside the cell, open circles are shared with neighbouring cells. Together they count as 8 atoms. (b) Atom density $8/a^3$ for the three group-14 crystals. Silicon has about 5×10^{22} atoms per cm³.

WORKED EXAMPLE Atoms in one cubic centimetre of silicon

Step 1. Volume of a unit cell: $a^3 = (0.5431 \times 10^{-7} \text{ cm})^3 = 1.602 \times 10^{-22} \text{ cm}^3$.

Step 2. Atom density: $N = 8/a^3 = 8/1.602 \times 10^{-22} = 4.99 \times 10^{22}$ atoms per cm³.

Check with density. Silicon's density is 2.33 g/cm³ and its molar mass is 28.09 g/mol, so $N = 2.33 \times 6.022 \times 10^{23} / 28.09 = 5.00 \times 10^{22} \text{ cm}^{-3}$. The two routes agree.

Why it matters. When we "dope" silicon with, say, 10^{16} phosphorus atoms per cm³, that is only one phosphorus atom per five million silicon atoms, yet it changes conductivity enormously (Chapter 5).

Crystal, polycrystal, amorphous G11

Table 2.1 Three forms of silicon

Form	Atomic order	Where it is used
Single crystal (monocrystalline)	One perfect lattice across the whole wafer	Chip wafers; high-efficiency solar cells
Polycrystalline ("poly")	Many small crystals (grains) with boundaries between them	Feedstock for growing crystals; some solar cells; gate and resistor layers inside chips
Amorphous (a-Si)	No long-range order; bonds are strained or "dangling"	Thin-film solar cells, display backplanes

Chips need **single-crystal** silicon because grain boundaries and defects trap and scatter the moving charges. A modern chip wafer is one crystal 300 mm across, purified to better than 99.999 999 9 % ("nine nines"). Chapter 10 explains how it is grown.

GOING DEEPER **Varsity**

The diamond lattice is two interpenetrating face-centred cubic (FCC) lattices offset by $(a/4, a/4, a/4)$. The bonds come from **sp^3 hybrid orbitals**: one s and three p orbitals combine into four equivalent orbitals pointing to the corners of a tetrahedron, 109.5° apart. Crystal direction matters in practice. Wafers are cut along defined planes, written with Miller indices such as (100), and properties like etch rate and electron mobility depend on direction.

TRY IT

Build a tetrahedral silicon "atom" from a ball of Prestik and four toothpicks pointing to the corners of a tetrahedron, then join several to see the 3-D network. Notice that every atom has exactly four neighbours and no spare bonds. Ask yourself what would happen if one atom had five arms (phosphorus) or three (boron). You have just predicted doping.

CHECK YOURSELF

1. **G10** How many protons, electrons and valence electrons does a neutral silicon atom have?
2. **G10** Why does silicon form four covalent bonds?
3. **G10** Name one element you could add to silicon to give an extra electron, and one that leaves an electron missing.
4. **G11** Germanium has $a = 0.5658$ nm and the same structure. Calculate its atom density.
5. **G11** Explain, using the bond model, why breaking one bond creates two charge carriers.
6. **G11** Why do chips use single-crystal rather than amorphous silicon?

WATCH & EXPLORE

- [Siyavula Physical Sciences, Grades 10–11](#) — the chapters on atomic structure, the periodic table and chemical bonding underpin this chapter. [siyavula.com](#) **G10**
- [PhET: Semiconductors](#) — University of Colorado. Add donor and acceptor atoms and watch electrons and holes appear. This is an older Java simulation, so it may need to be downloaded. [phet.colorado.edu/en/simulations/semiconductor](#) **G11**
- [Chenming Hu, *Modern Semiconductor Devices for Integrated Circuits*](#) — chapter 1 covers the silicon crystal and bond model. The chapters can be downloaded as PDFs from the author's page. [chu.berkeley.edu](#) **Varsity**

IN PLAIN WORDS

A silicon atom has four outer electrons. It shares them with four neighbours, so every atom is held by four bonds. Billions of atoms lock together in a neat, repeating pattern called a crystal. A cubic centimetre of silicon has about 50 thousand billion billion atoms. If heat or light breaks one bond, an electron escapes and can move. The gap it leaves, called a hole, can move too. Moving electrons and holes are what carry current in a chip.

SOURCES

1. Lattice constant, density and molar mass of Si: S. M. Sze & K. K. Ng, *Physics of Semiconductor Devices*, 3rd ed., Wiley, 2007, Appendix; Ioffe Institute NSM archive, "Silicon: basic parameters".
2. C. Hu, *Modern Semiconductor Devices for Integrated Circuits*, Pearson, 2010, ch. 1.

Electricity essentials

GRADE 10

GRADE 11

Before we can understand why semiconductors are special, we need the ideas every electrician and chip designer uses daily: charge, current, voltage, resistance and power. This chapter builds them from the Grade 10 basics and ends with a surprise: electrons in a wire crawl slower than a snail.

IN THIS CHAPTER YOU WILL LEARN

- what charge, current and potential difference are, with units
- Ohm's law, resistance and resistivity, with real wire calculations
- how series and parallel circuits share current and voltage
- power, energy and what your electricity actually costs
- emf and internal resistance of real batteries
- how fast electrons really move, and how to use a multimeter safely

Electric charge G10

Charge is a basic property of matter, like mass. It comes in two kinds, positive and negative. Like charges repel and unlike charges attract (Grade 10 Electrostatics). Charge is measured in **coulombs** (C).

Charge comes in fixed lumps. The smallest free lump is the charge of one electron, $-e$, or of one proton, $+e$, where the **elementary charge** is

$$e = 1.602\,176\,634 \times 10^{-19} \text{ C} \quad (3.1)$$

This value is exact: since 2019 it has been used to define the SI units^[1]. One coulomb is therefore the charge of $1/e = 6.24 \times 10^{18}$ electrons. Charge is also **conserved**: it can move from place to place, but the total never changes. In a semiconductor, when a bond breaks (Chapter 2) one electron ($-e$) and one hole ($+e$) appear together, so the total stays zero.

Current: charge on the move G10

Current is the rate at which charge passes a point in a circuit:

$$I = \frac{Q}{\Delta t} \quad (3.2)$$

I is in **amperes** (A), Q in coulombs and Δt in seconds. One ampere is one coulomb per second. Chips use small currents, so you will often see milliamperes ($1 \text{ mA} = 10^{-3} \text{ A}$), microamperes ($1 \mu\text{A} = 10^{-6} \text{ A}$) and nanoamperes ($1 \text{ nA} = 10^{-9} \text{ A}$).

WORKED EXAMPLE Charging a phone

Question. A phone charges at a steady 2.0 A for 1.0 hour. How much charge flows, and how many electrons is that?

Step 1. $\Delta t = 1.0 \text{ h} = 3600 \text{ s}$, so $Q = I\Delta t = 2.0 \times 3600 = 7200 \text{ C}$.

Step 2. Number of electrons $N = Q/e = 7200/(1.602 \times 10^{-19}) = 4.49 \times 10^{22}$.

Answer. 7 200 C, carried by about 4.5×10^{22} electrons, roughly the number of atoms in a cubic centimetre of silicon (Chapter 2).

Conventional current and electron flow

Long before electrons were discovered, scientists agreed that current flows from the positive terminal of a cell, through the circuit, to the negative terminal. This is **conventional current**, and every circuit diagram, every formula and every datasheet still uses it. In a metal wire the moving charges are actually electrons, which are negative, so they move the *opposite* way. Both descriptions give the same answers, as long as you are consistent.

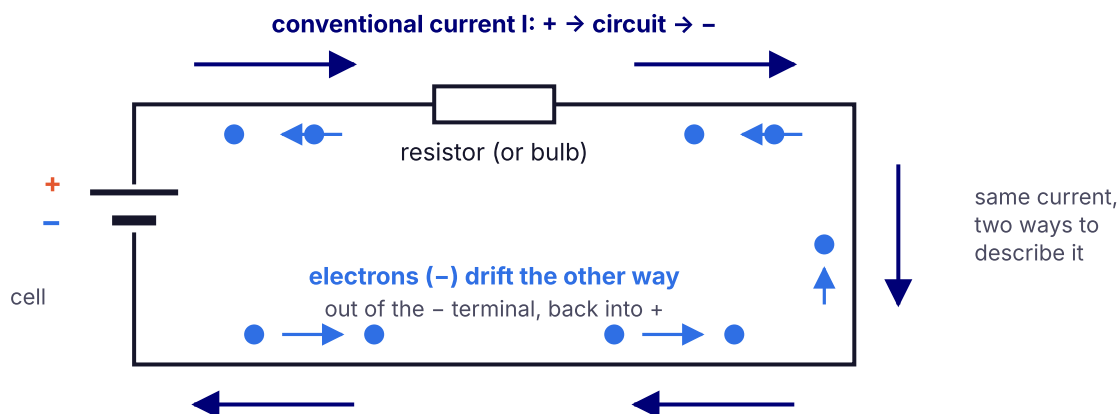


Figure 3.1 In a wire, negative electrons (blue) drift from the negative terminal round to the positive terminal, while conventional current (navy arrows) is drawn from positive to negative. A flow of negative charge one way is equivalent to a flow of positive charge the other way. In semiconductors both happen at once: electrons drift one way and holes, which are positive, drift the other, and their currents add (Chapter 6).

Potential difference and energy G10

Charges flow only when something pushes them. A cell or battery does work on the charges, giving them energy, which they hand over in the circuit (as heat in a resistor, light in a bulb, motion in a motor). The **potential difference** (p.d., or **voltage**) between two points is the energy transferred per unit charge:

$$V = \frac{W}{Q} \quad \Longleftrightarrow \quad W = QV \quad (3.3)$$

V is in **volts** (V), where $1 \text{ V} = 1 \text{ joule per coulomb}$. A 5 V USB charger gives every coulomb that passes through it 5 J of energy. A single electron moved through 1 V gains $1.602 \times 10^{-19} \text{ J}$, an amount so useful in semiconductor physics that it has its own name, the **electron-volt** (eV). You will use it from Chapter 4 onwards.

WORKED EXAMPLE Energy in a phone battery

Question. (a) In the charging example, 7 200 C passed through a 5.0 V charger. How much energy was delivered? (b) A phone battery label says 4 000 mAh, 3.85 V. How much energy does it store?

(a) $W = QV = 7200 \times 5.0 = 36\,000 \text{ J} = 36 \text{ kJ}$.

(b) "mAh" is a unit of charge: $4000 \text{ mAh} = 4.000 \text{ A} \times 3600 \text{ s} = 14\,400 \text{ C}$. $W = QV = 14\,400 \times 3.85 = 55\,440 \text{ J} \approx 55 \text{ kJ}$, also written $4.000 \text{ Ah} \times 3.85 \text{ V} = 15.4 \text{ Wh}$.

Answer. 36 kJ and about 55 kJ (15.4 Wh). For comparison, lifting a 1 kg bag of sugar by 1 m takes about 10 J.

Resistance and Ohm's law G11

When a p.d. is applied to a conductor, a current flows. The **resistance** R tells us how much p.d. is needed per ampere:

$$R = \frac{V}{I} \quad (3.4)$$

R is in **ohms** (Ω). For many materials at constant temperature, R stays the same whatever the current: double the p.d. and the current doubles. This is **Ohm's law** (Grade 11): the current through a conductor is directly proportional to the p.d. across it, at constant temperature. A conductor that obeys it is called **ohmic**; its current–voltage (I–V) graph is a straight line through the origin with gradient $1/R$.

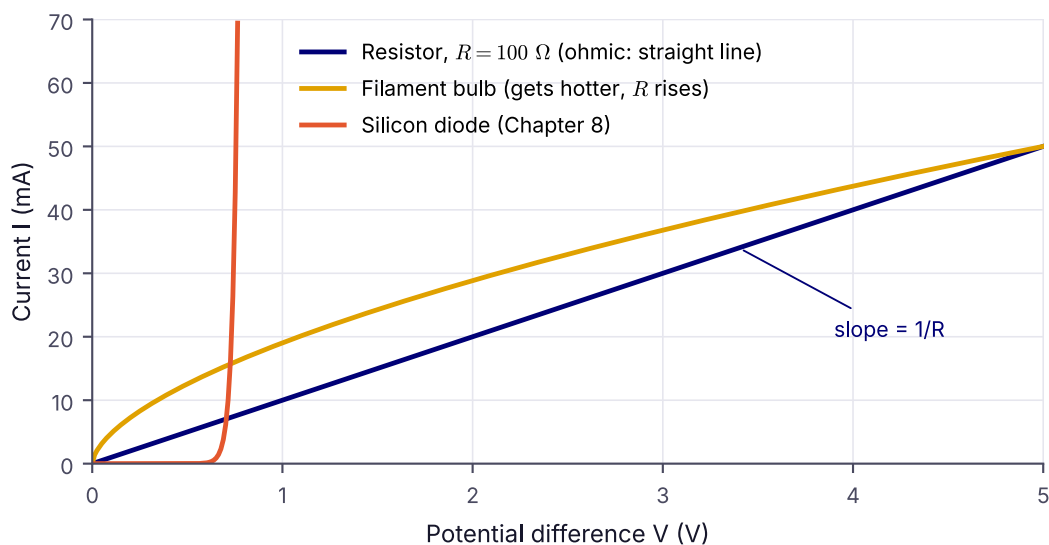


Figure 3.2 Current–voltage graphs. A resistor is ohmic (straight line). A filament bulb is not: as it heats up its resistance rises and the curve bends over. A silicon diode is strongly non-ohmic: almost no current until about 0.6 V, then a steep rise. Illustrative models: resistor $I = V/R$ with $R = 100 \Omega$; bulb $I \propto V^{0.6}$; diode $I = I_s(e^{V/V_T} - 1)$ with $I_s = 10^{-14} \text{ A}$, $V_T = 25.85 \text{ mV}$ at 300 K (Chapter 7). Semiconductor devices are interesting precisely because they are not ohmic.

COMMON MISTAKE

$R = V/I$ is the *definition* of resistance and works for any device at any point on its curve. Ohm's law is the separate, experimental statement that R is constant. Diodes and transistors do not obey Ohm's law, but you can still divide V by I at a given point.

Resistivity: the material, not the shape G11

Chapter 1 introduced resistivity ρ , the property that makes copper a conductor and glass an insulator. Resistance grows with length L and shrinks with cross-section area A :

$$R = \frac{\rho L}{A} \quad (3.5)$$

Think of a crowd leaving a stadium: a longer tunnel slows them more; a wider tunnel lets more through at once. The inverse of resistivity is the **conductivity** $\sigma = 1/\rho$, in siemens per metre (S/m). Semiconductor engineers often quote resistivity in $\Omega \cdot \text{cm}$ ($1 \Omega \cdot \text{m} = 100 \Omega \cdot \text{cm}$).

WORKED EXAMPLE A house wire and a bar of pure silicon

Question. (a) Find the resistance of a 20 m length of copper house wire with a 1.5 mm^2 cross-section ($\rho_{\text{Cu}} = 1.68 \times 10^{-8} \Omega \cdot \text{m}$). (b) Find the resistance of a bar of pure silicon 1.0 cm long with a 1.0 mm^2 cross-section ($\rho = 2.3 \times 10^3 \Omega \cdot \text{m}$).

Units first. $1.5 \text{ mm}^2 = 1.5 \times 10^{-6} \text{ m}^2$ (because $1 \text{ mm} = 10^{-3} \text{ m}$, so $1 \text{ mm}^2 = 10^{-6} \text{ m}^2$). $1.0 \text{ cm} = 0.010 \text{ m}$.

$$\text{(a)} \quad R = \frac{1.68 \times 10^{-8} \times 20}{1.5 \times 10^{-6}} = 0.224 \Omega.$$

$$\text{(b)} \quad R = \frac{2.3 \times 10^3 \times 0.010}{1.0 \times 10^{-6}} = 2.3 \times 10^7 \Omega = 23 \text{ M}\Omega.$$

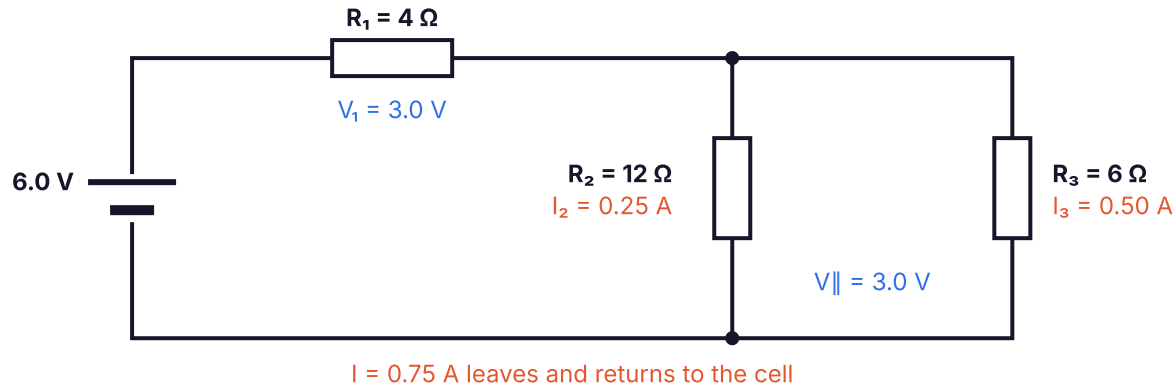
Answer. 0.22Ω for 20 m of copper, but 23 million ohms for a tiny bar of pure silicon. That is why pure silicon is useless as a wire, and why doping (Chapter 5) matters so much.

Series and parallel circuits G10 G11

In a **series** circuit the components are in one loop, so the *same current* passes through each, and the p.d.s add up. In a **parallel** circuit components are connected across the same two points, so they have the *same p.d.*, and the currents add up. The rules for combining resistors follow:

$$R_{\text{series}} = R_1 + R_2 + \dots \qquad \frac{1}{R_{\text{parallel}}} = \frac{1}{R_1} + \frac{1}{R_2} + \dots \quad (3.6)$$

Adding a resistor in parallel always *lowers* the total resistance, because it opens an extra path. This is why plugging in more appliances draws more current from the supply.



Check: $0.25 \text{ A} + 0.50 \text{ A} = 0.75 \text{ A}$ (currents add in parallel); $3.0 \text{ V} + 3.0 \text{ V} = 6.0 \text{ V}$ (p.d.s add in series)

Figure 3.3 A series-parallel circuit, solved in the worked example below. The same 0.75 A flows through R_1 and then splits between R_2 and R_3 ; the bigger share takes the easier (smaller-resistance) path. Every chip is, at heart, a huge network like this, solved by computer.

WORKED EXAMPLE Solving the circuit in Figure 3.3

Step 1. Parallel pair. $1/R_{||} = 1/12 + 1/6 = 3/12$, so $R_{||} = 4 \Omega$.

Step 2. Total. $R_{\text{total}} = R_1 + R_{||} = 4 + 4 = 8 \Omega$. Current from the cell $I = V/R = 6.0/8 = 0.75 \text{ A}$.

Step 3. P.d.s. $V_1 = IR_1 = 0.75 \times 4 = 3.0 \text{ V}$; $V_{||} = 0.75 \times 4 = 3.0 \text{ V}$.

Step 4. Branch currents. $I_2 = 3.0/12 = 0.25 \text{ A}$; $I_3 = 3.0/6 = 0.50 \text{ A}$. Check: $0.25 + 0.50 = 0.75 \text{ A}$. ✓

Power and energy: what electricity costs G11

Power is the rate of transferring energy, in watts ($1 \text{ W} = 1 \text{ J/s}$). Combining $W = QV$ with $Q = I\Delta t$ gives

$$P = \frac{W}{\Delta t} = VI = I^2R = \frac{V^2}{R} \quad (3.7)$$

The last two forms use $V = IR$. The form $P = I^2R$ explains why wires and chips heat up: power lost in a resistance grows with the *square* of the current. Chip designers fight this every day (Chapter 9).

Electricity suppliers sell energy, not power. Their unit is the **kilowatt-hour** (kWh), the energy used by a 1 kW appliance in 1 hour: $1 \text{ kWh} = 1000 \text{ W} \times 3600 \text{ s} = 3.6 \text{ MJ}$. Your prepaid slip shows how many kWh ("units") you bought.

WORKED EXAMPLE The kettle and the geyser

Question. A 2 000 W kettle runs on South Africa's 230 V mains. (a) What current does it draw and what is its resistance? (b) It is used for 15 minutes a day. How many kWh is that in a 30-day month, and what does it cost at an illustrative R 3.50 per kWh? (c) Compare a 3 kW geyser element that heats for 2 hours a day.

(a) $I = P/V = 2000/230 = 8.7 \text{ A}$. $R = V^2/P = 230^2/2000 = 26 \Omega$ (the hot nichrome element).

(b) Energy = $2.0 \text{ kW} \times 0.25 \text{ h/day} \times 30 \text{ days} = 15 \text{ kWh}$. Cost = $15 \times \text{R } 3.50 = \text{R } 52.50$.

(c) $3.0 \times 2 \times 30 = 180 \text{ kWh}$, costing $180 \times \text{R } 3.50 = \text{R } 630$, twelve times the kettle.

Answer. 8.7 A and 26Ω ; 15 kWh (R 52.50); the geyser uses 180 kWh (R 630). Heating water is where most household electricity goes, which is why solar geysers and geyser timers save the most money.

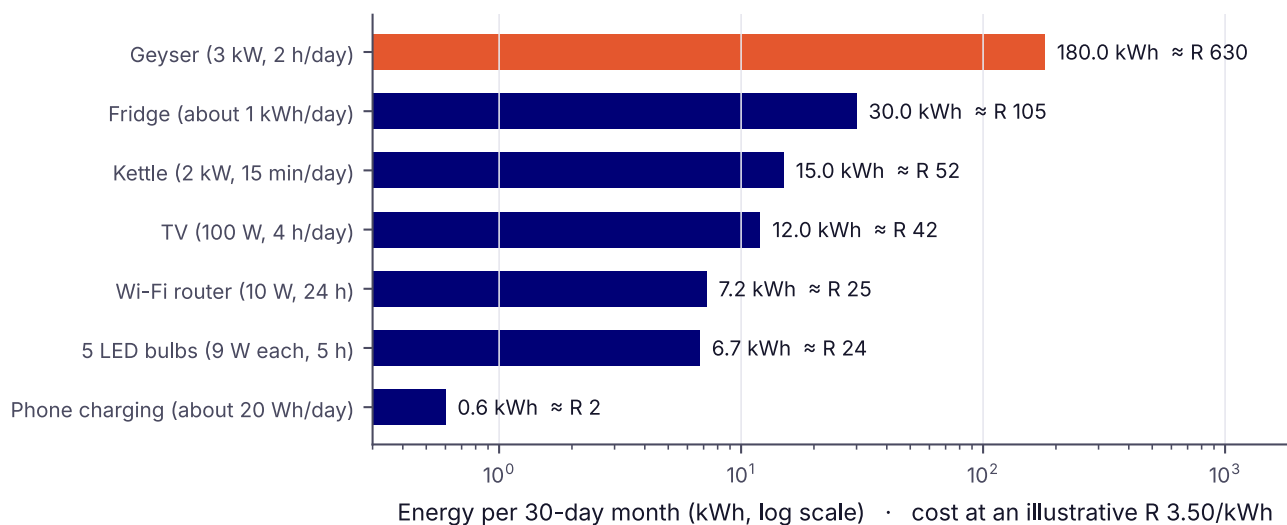


Figure 3.4 Monthly energy use of common appliances, for the stated power ratings and hours of use (assumptions, not measurements; check your own appliances' labels). The log scale is needed because the geyser uses about 300 times more energy than charging a phone. Cost uses an illustrative tariff of R 3.50/kWh **ESTIMATE**. Together these appliances use about 252 kWh per month, of which the geyser is about 72 %.

ABOUT THESE NUMBERS

Electricity tariffs in South Africa differ between Eskom-supplied and municipal customers, between prepaid and credit meters, and between tariff blocks (the price per kWh often rises after a monthly threshold); they change every year. We did not use a specific published tariff here, so R 3.50/kWh is an **illustrative round number** **ESTIMATE**. Replace it with the rand-per-kWh figure on your own prepaid slip or bill: divide the rands paid by the kWh received.

TRY IT

Find the rating plate on three appliances at home (power in W, often next to "230 V ~ 50 Hz"). Estimate how many hours each is on per day and calculate its kWh per month. Then check your prepaid meter reading at the same time on two days, 24 hours apart, to measure your household's daily energy use. How close is your estimate?

Real batteries: emf and internal resistance G11 G12

A real battery is not a perfect source of p.d. Some of its energy is lost inside it, because the chemicals and connections have resistance. We model a real battery as an ideal source of **emf** ε (the energy it gives each coulomb, in volts) in series with a small **internal resistance** r .

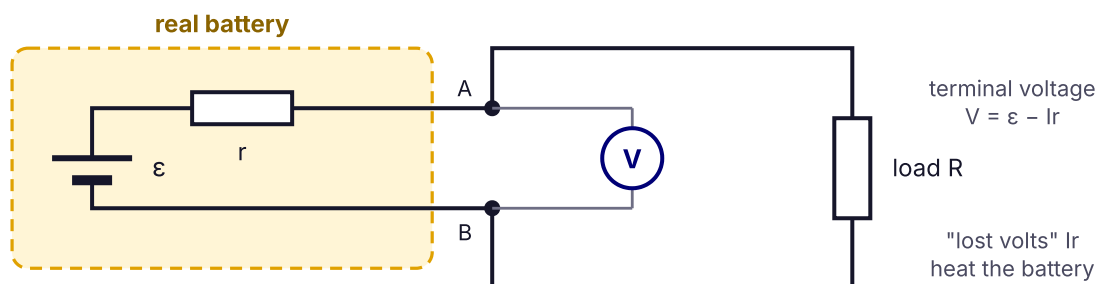


Figure 3.5 Model of a real battery: an ideal emf ε in series with an internal resistance r , inside the dashed box. A voltmeter across the terminals A and B reads the terminal voltage, which drops as more current is drawn. The same model describes solar panels, inverter batteries and the power supply of a chip.

$$\varepsilon = I(R + r) \quad \Rightarrow \quad V_{\text{terminal}} = IR = \varepsilon - Ir \quad (3.8)$$

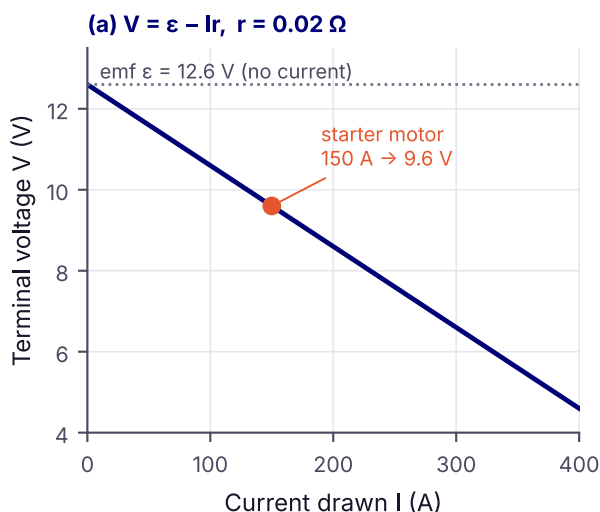


Figure 3.6 Terminal voltage of a car battery model ($\varepsilon = 12.6 \, \text{V}$, $r = 0.02 \, \Omega$, illustrative values). Starting the engine draws about 150 A and the voltage dips to 9.6 V, which is why headlights dim when you start the car.

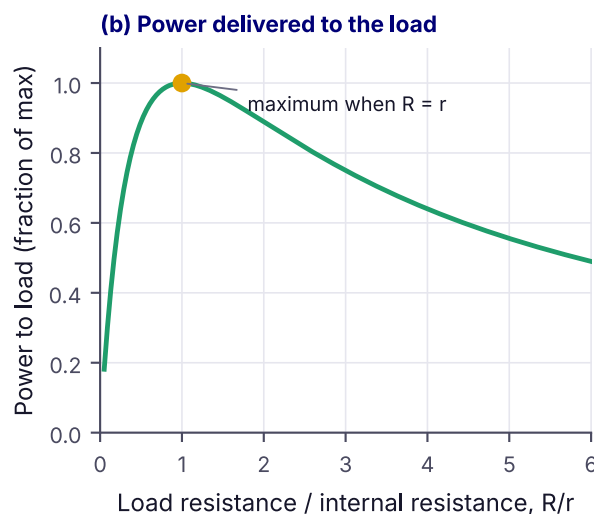


Figure 3.7 Power delivered to the load, $P = \varepsilon^2 R / (R + r)^2$, peaks when the load resistance equals the internal resistance (Going deeper below). Curve is the exact model, scaled to its maximum.

WORKED EXAMPLE Starting a car

Question. A car battery has $\varepsilon = 12.6 \, \text{V}$ and $r = 0.020 \, \Omega$. The starter motor draws 150 A. Find the terminal voltage and the power wasted inside the battery.

Step 1. Lost volts $= Ir = 150 \times 0.020 = 3.0 \, \text{V}$. Terminal voltage $= 12.6 - 3.0 = 9.6 \, \text{V}$.

Step 2. Power wasted $= I^2 r = 150^2 \times 0.020 = 450 \, \text{W}$.

Answer. 9.6 V at the terminals, and 450 W heating the battery while the starter turns. Cold batteries have higher r , which is why cars struggle to start on winter mornings.

GOING DEEPER G12

Maximum power transfer. The power in the load is $P = I^2 R = \varepsilon^2 R / (R + r)^2$. Setting $dP/dR = 0$ gives $\varepsilon^2 [(R + r)^2 - 2R(R + r)] / (R + r)^4 = 0$, so $R = r$, and then $P_{\max} = \varepsilon^2 / 4r$. At that point half the energy is wasted inside the source, so power systems never run there; but radio and chip designers often match impedances deliberately to get the most signal power. Solar inverters do something related: their "maximum power point tracking" constantly adjusts the load the panel sees (Chapter 7).

How fast do electrons move? Drift velocity G11 G12

When you flip a switch the light comes on instantly. So electrons must move fast? Surprisingly, no. In a metal the free electrons already rush about randomly at enormous speeds, around 10^6 m/s^[2], colliding with vibrating atoms and changing direction all the time. With no p.d. the random motion averages to zero: no current. Applying a p.d. adds a tiny, steady push, so the electrons slowly *drift* along the wire while still zig-zagging. This slow average speed is the **drift velocity** v_d .

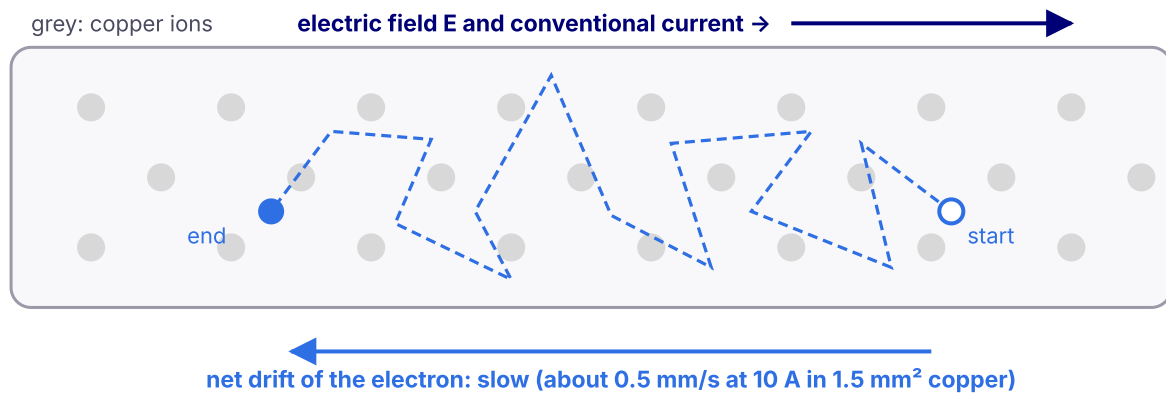


Figure 3.8 A free electron in copper zig-zags at about 10^6 m/s between collisions with the vibrating ions, but the electric field nudges it a tiny bit to the left (opposite to E , because it is negative) each time. The net drift is millions of times slower than the random motion. Signals travel fast because the field, not any one electron, moves along the wire at a large fraction of the speed of light.

To find v_d , picture a length of wire with cross-section A containing n free electrons per cubic metre, each with charge q . In time Δt the electrons move a distance $v_d \Delta t$, so the charge passing a point is $Q = n A v_d \Delta t q$. Dividing by Δt :

$$I = n A q v_d \quad (3.9)$$

WORKED EXAMPLE The snail-speed electron

Question. A 1.5 mm^2 copper wire carries 10 A. Copper has density 8.96 g/cm^3 and molar mass 63.55 g/mol , and each atom gives one free electron. Find the drift velocity.

Step 1. Free-electron density. Moles per $\text{m}^3 = 8.96 \times 10^6 \text{ g/m}^3 / 63.55 = 1.410 \times 10^5$. So $n = 1.410 \times 10^5 \times 6.022 \times 10^{23} = 8.49 \times 10^{28} \text{ m}^{-3}$.

Step 2. $v_d = \frac{I}{n A q} = \frac{10}{8.49 \times 10^{28} \times 1.5 \times 10^{-6} \times 1.602 \times 10^{-19}} = 4.9 \times 10^{-4} \text{ m/s}$.

Answer. About 0.5 mm/s. An electron would take $1 / (4.9 \times 10^{-4}) \approx 2000 \text{ s}$, about 34 minutes, to drift one metre. A garden snail is faster.

KEY IDEA

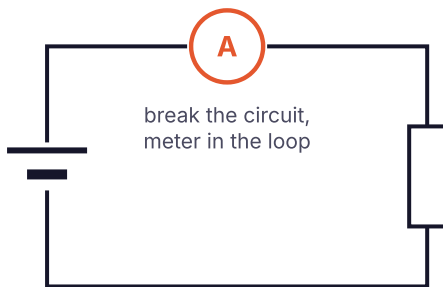
Current = (number of carriers per volume) $\times$ (charge) $\times$ (drift speed) $\times$ (area). Metals have a huge n , about 10^{29} m^{-3} , so a slow drift gives a big current. Pure silicon at room temperature has only about 10^{16} free electrons per m^3 (Chapter 5), ten million million times fewer, so the same current would need impossibly fast drift. That single comparison is why we must *add* carriers to silicon by doping.

Measuring safely with a multimeter G10

A **multimeter** measures p.d. (voltmeter), current (ammeter) and resistance (ohmmeter). It is the first tool of every electronics engineer, and cheap ones are good enough for learning. The two key rules come straight from series and parallel circuits:

- A **voltmeter** is connected **in parallel**, across the component, because p.d. is measured between two points. A good voltmeter has a very high resistance, so it draws almost no current.
- An **ammeter** is connected **in series**, by breaking the circuit and letting the current flow through the meter. A good ammeter has a very low resistance.

(a) Current: ammeter in series



(b) P.d.: voltmeter in parallel

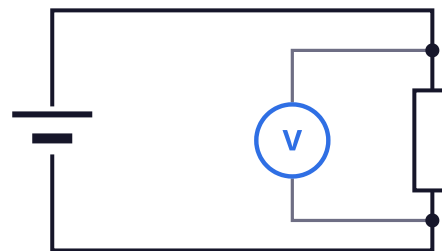


Figure 3.9 (a) An ammeter must carry the current, so it goes in series. (b) A voltmeter compares two points, so it goes in parallel. Connecting an ammeter in parallel across a battery creates a near short circuit and usually blows the meter's fuse.

SAFETY FIRST

Practise only on batteries and low-voltage circuits (up to about 12 V). **Never** probe the 230 V mains, a plug socket, an inverter's output or a solar panel string unless you are trained and supervised: these can kill. Before each measurement: choose the correct function and a range higher than you expect; put the black lead in COM and the red lead in the V/ Ω socket (move it to the A or mA socket only for current); measure resistance only with the power off; and check that the meter's category rating (CAT II, CAT III ...) suits the circuit.

CHECK YOURSELF

1. **G10** How many electrons pass a point each second in a wire carrying 1.0 mA?
2. **G10** A torch bulb draws 0.30 A from a 3.0 V battery for 10 minutes. How much charge flows and how much energy is transferred?
3. **G10** Explain the difference between conventional current and electron flow. Which is used on circuit diagrams?
4. **G11** Three resistors of 10 Ω , 20 Ω and 20 Ω are connected (a) in series, (b) in parallel. Find the total resistance each time.
5. **G11** A 100 m extension cord has two copper conductors, each 1.0 mm² (so the current travels 200 m in total). Find the resistance of the copper and the power lost when it carries 8.0 A.
6. **G11** A 60 W bulb and a 9 W LED give similar light. If each runs 5 hours a day, how many kWh does the LED save in a 30-day month, and what is that worth at R 3.50/kWh?
7. **G12** A battery has $\epsilon = 9.0$ V. When a 10 Ω resistor is connected the terminal voltage is 8.0 V. Find the current and the internal resistance.
8. **G12** Aluminium has about 1.8×10^{29} free electrons per m³ (three per atom). An aluminium wire of area 2.0 mm² carries 5.0 A. Find the drift velocity.
9. **G11** Why is a voltmeter designed to have a very high resistance and an ammeter a very low one?

WATCH & EXPLORE

- ▶ **Electric Current: Crash Course Physics #28** — CrashCourse. Charge, current, voltage and resistance with clear animations. **G10** youtube.com/watch?v=HX0ok3mfMLM
- ▶ **How Electricity Actually Works** — Veritasium. Why the energy travels in the fields around wires, not inside the slow electrons. **G12** youtube.com/watch?v=oI_X2cMHNe0
- **Circuit Construction Kit: DC** — PhET. Build series and parallel circuits, add meters and watch the electrons. **G10** phet.colorado.edu/en/simulations/circuit-construction-kit-dc
- **Resistance in a Wire and Ohm's Law** — PhET. Change ρ , L and A, or V and R, and watch the equation respond. **G11** phet.colorado.edu/en/simulations/resistance-in-a-wire
- **Siyavula Grade 11: Electric circuits and Grade 12: Electric circuits** — free CAPS chapters on Ohm's law, power and internal resistance. **G11** siyavula.com/read/za/physical-sciences/grade-11/electric-circuits
- **Ohm's Law, Microscopic View** — HyperPhysics. Drift velocity and the free-electron picture of conduction. **G12** hyperphysics.phy-astr.gsu.edu/hbase/electric/ohmmic.html

IN PLAIN WORDS

Charge is a basic property of tiny particles. Electrons carry negative charge. Current is how much charge passes each second.

A battery gives each bit of charge some energy. The voltage tells you how much.

Resistance says how hard it is to push current through something. Long, thin wires resist more. Some materials resist much more than others.

Power is how fast energy is used. You pay for energy in kilowatt-hours. Heating water uses the most.

Real batteries lose some voltage inside themselves when you draw a big current.

Electrons in a wire drift very slowly, slower than a snail. A wire works because it has a huge number of free electrons. Pure silicon has very few. That is why we must add electrons or holes to silicon.

Use a multimeter only on low voltages. Never touch mains wiring.

SOURCES

1. BIPM, *The International System of Units (SI)*, 9th ed., 2019 (exact value of the elementary charge); NIST CODATA, "elementary charge". <https://physics.nist.gov/cgi-bin/cuu/Value?e> (accessed 15 Oct 2026).
2. C. Kittel, *Introduction to Solid State Physics*, 8th ed., Wiley, 2005, ch. 6 (free-electron Fermi gas; Fermi velocity of copper about 1.6×10^6 m/s).
3. Siyavula, *Physical Sciences Grade 10, 11 and 12*, chapters "Electric circuits", free CAPS textbooks. <https://www.siyavula.com/read/za/physical-sciences/grade-11/electric-circuits> (accessed 15 Oct 2026).
4. Wikipedia, "Electrical resistivity and conductivity" (resistivities of copper and silicon at 20 °C). https://en.wikipedia.org/wiki/Electrical_resistivity_and_conductivity (accessed 15 Oct 2026).
5. D. Halliday, R. Resnick and J. Walker, *Fundamentals of Physics*, 10th ed., Wiley, 2014, ch. 26–27 (current, drift speed, resistivity, emf and internal resistance).

Energy Bands and the Band Gap

GRADE 11

GRADE 12

VARSITY

Why does silicon look shiny and grey while glass is clear? Why do LEDs come in different colours? One idea answers both, and it is the most important picture in semiconductor physics: the energy band diagram.

IN THIS CHAPTER YOU WILL LEARN

- how energy levels in atoms become bands in solids
- valence band, conduction band and band gap
- the electron-volt, the natural energy unit
- why metals, semiconductors and insulators differ
- how photon energy links band gap to colour: $\lambda \approx 1240/E$
- direct and indirect band gaps

From energy levels to energy bands G11 G12

In Grade 12 you meet atomic spectra: an electron in an atom can only have certain energies, called **energy levels**. It jumps between them by absorbing or emitting a photon of exactly the right energy, which is why each element has its own line spectrum.

Now bring two silicon atoms close together. Their outer electrons start to feel both nuclei, and each energy level splits into two slightly different levels. Bring together N atoms and each level splits into N levels. In a real crystal N is around 10^{22} per cubic centimetre, so the levels lie so close together that they form an almost continuous **band** of allowed energies. Between bands there can be ranges of energy that no electron in the crystal may have: **band gaps**.

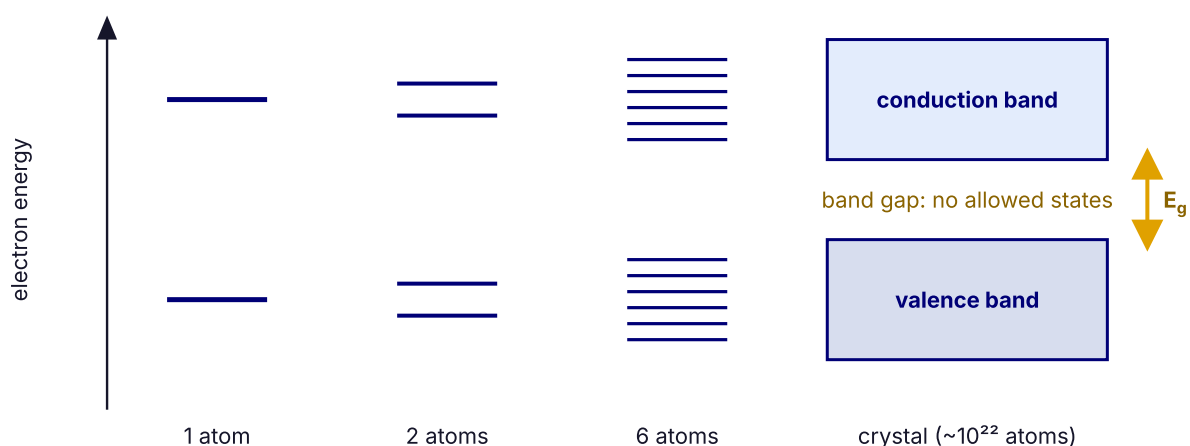


Figure 4.1 As atoms join into a crystal, each sharp atomic level spreads into a band. In silicon, the band formed from the bonding (valence) electrons is the **valence band**, and the next band up is the **conduction band**. The gap between them, E_g , is 1.12 eV at 300 K.

The electron-volt

Energies of single electrons are tiny in joules, so we use the **electron-volt (eV)**: the energy an electron gains moving through a potential difference of 1 V.

$$1 \text{ eV} = 1.602 \times 10^{-19} \text{ J} \quad (4.1)$$

Why some solids conduct and others don't G11

To carry current, an electron must be able to gain a little energy from the electric field, which means moving into a slightly higher **empty** state. A completely full band cannot carry current, because there is nowhere for its electrons to go. A completely empty band carries none either, because it has no electrons. What matters is **partly filled** bands.

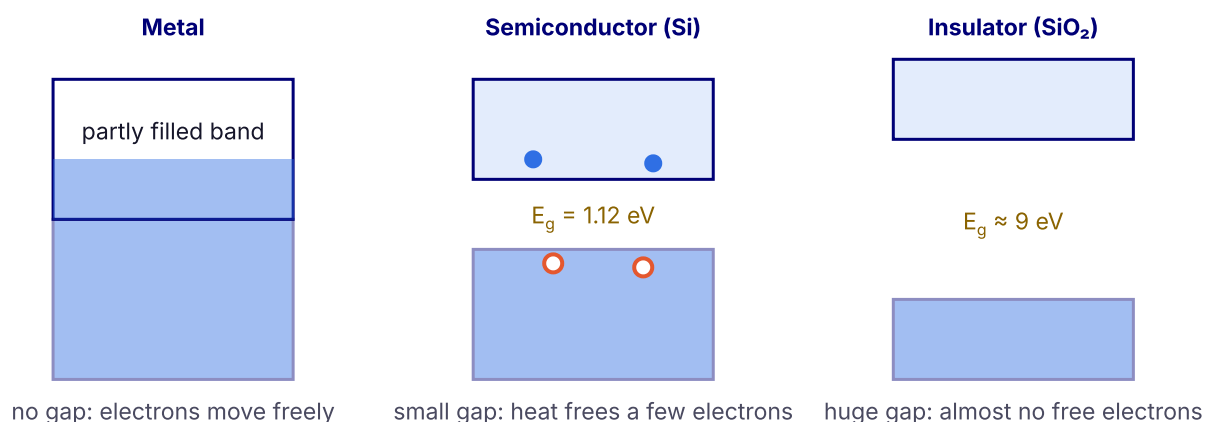


Figure 4.2 Band pictures. Shading shows filled states. In a metal the top band is only partly filled, so conduction is easy. In a semiconductor the valence band is full and the conduction band empty at absolute zero, but the gap is small enough that heat or light lifts a few electrons across, leaving holes (red rings) behind. In an insulator the gap is too wide for that to happen at ordinary temperatures.

KEY IDEA

The **band gap** E_g is the minimum energy needed to free an electron from its bond so it can carry current. It is the energy-band version of "breaking a bond" from Chapter 2. When an electron is lifted into the conduction band, it leaves a **hole** in the valence band. Both can carry current.

Band gap, light and colour G12

A photon of light carries energy $E = hf = hc/\lambda$, where $h = 6.626 \times 10^{-34} \text{ J}\cdot\text{s}$ is Planck's constant, $c = 3.00 \times 10^8 \text{ m/s}$ and λ is the wavelength. Putting E in eV and λ in nm gives a handy shortcut:

$$E (\text{eV}) \approx \frac{1240}{\lambda (\text{nm})} \quad (4.2)$$

A semiconductor **absorbs** photons with $E > E_g$ (they have enough energy to lift an electron across the gap) and is **transparent** to photons with $E < E_g$. Run the same process backwards and an electron falling across the gap can **emit** a photon of energy about E_g . That is how an LED makes light (Chapter 7).

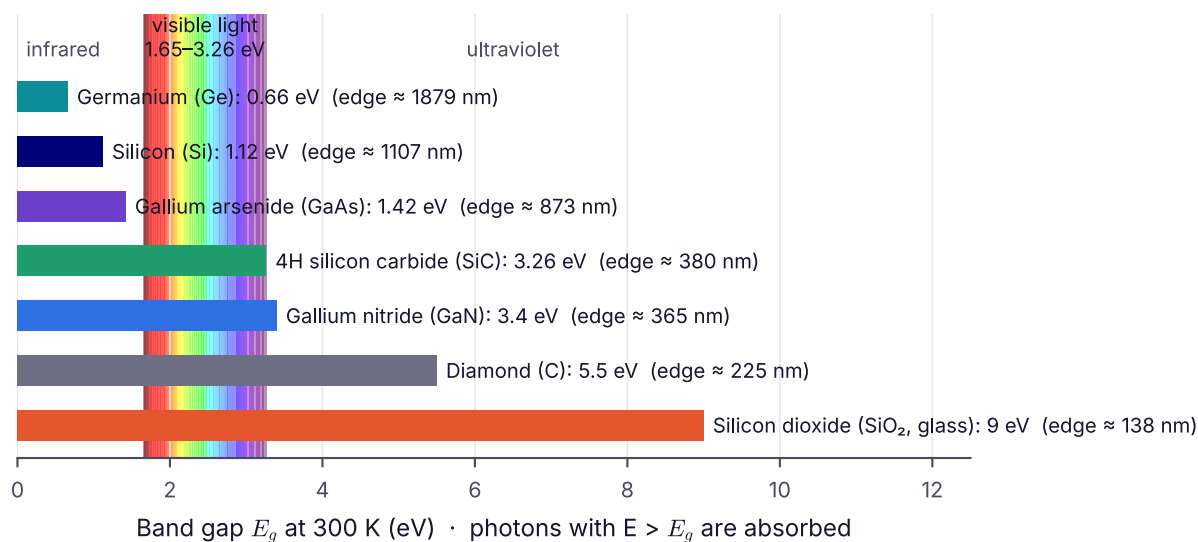


Figure 4.3 Band gaps at 300 K, with the corresponding absorption-edge wavelength $\lambda = 1240/E_g$. Materials whose gap lies above the visible band (SiC, GaN, diamond, glass) are transparent to visible light. Silicon and germanium absorb all visible light, so they look grey and metallic.

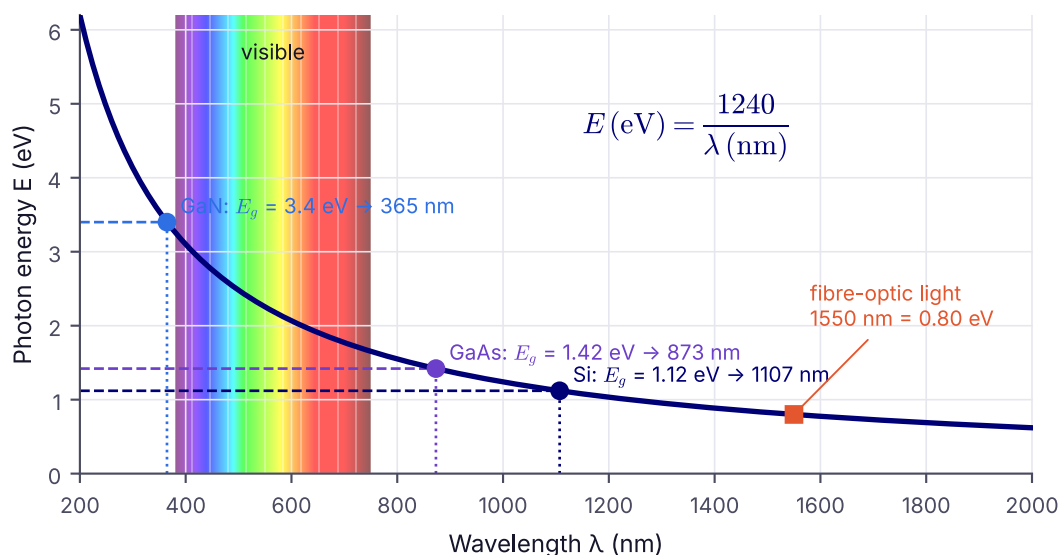


Figure 4.4 Photon energy against wavelength. Silicon absorbs everything shorter than about 1100 nm, which is why it makes good solar cells and camera sensors. It is transparent at 1550 nm, the wavelength used in fibre-optic internet cables, which is why silicon can also guide that light ("silicon photonics").

WORKED EXAMPLE Can silicon see a TV remote?

Question. A TV remote's infrared LED emits at 940 nm. Will a silicon photodiode detect it? Will a GaN detector?

Step 1. Photon energy: $E = 1240/940 = 1.32 \text{ eV}$.

Step 2. Silicon: $1.32 > 1.12 \text{ eV}$, so the photon is absorbed and a silicon photodiode responds. (Your phone camera, which is silicon, can often "see" the remote flashing.)

Step 3. GaN: $1.32 < 3.4 \text{ eV}$, so the photon passes straight through. GaN cannot detect it.

Answer. Silicon yes, GaN no. This is why GaN makes "solar-blind" ultraviolet detectors.

TRY IT

Point a TV remote at your phone's camera, using the front (selfie) camera if the back one has an infrared filter, and press a button. You will see a flashing light your eyes cannot see. You are watching silicon absorb photons with energy above its 1.12 eV band gap.

Direct and indirect gaps Varsity

An electron in a crystal has both an energy and a crystal momentum, written $\hbar k$. Plotting allowed energy against k gives the **E- k diagram** (band structure). In a **direct-gap** material (GaAs, GaN, InP) the lowest conduction-band energy sits at the same k as the highest valence-band energy, so an electron can drop across the gap and emit a photon directly. In an **indirect-gap** material (Si, Ge) the two are at different k . A photon carries almost no momentum, so the transition also needs a lattice vibration (a **phonon**). That makes light emission rare. Silicon is excellent for chips and solar cells but poor for LEDs and lasers.

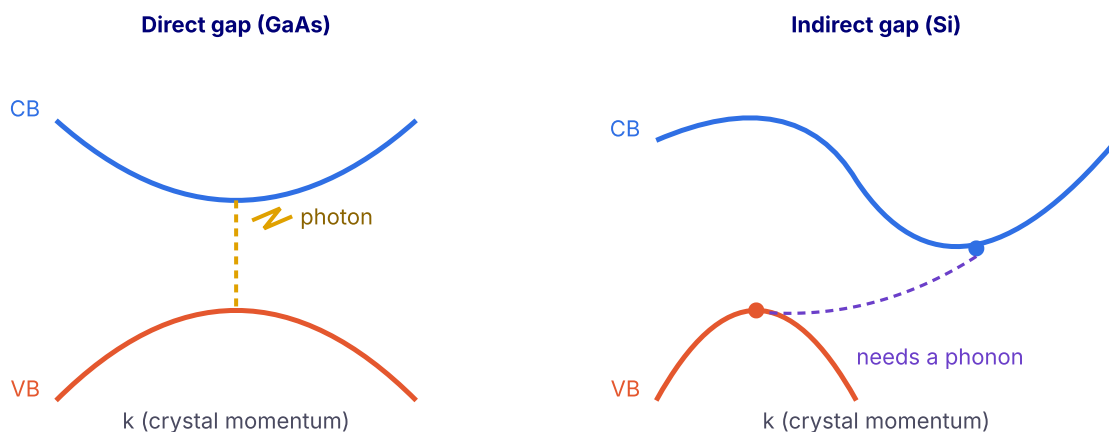


Figure 4.5 Simplified E- k diagrams. CB is the conduction band, VB the valence band. In a direct-gap material the band edges line up vertically, so light emission is efficient. In silicon the conduction-band minimum is shifted in k , so emitting a photon needs help from a lattice vibration (phonon) and is very slow.

GOING DEEPER Varsity

Bands come from solving the Schrödinger equation for an electron in a periodic potential. Bloch's theorem says the solutions are plane waves modulated by a function with the lattice period. Near a band edge, $E(k) \approx E_C + \hbar^2 k^2 / 2m^*$, where m^* is the **effective mass**. It packs the effect of the crystal into one number, so the electron behaves almost like a free particle with a different mass. The band gap also shrinks slightly as temperature rises. For silicon it is about 1.17 eV near 0 K and 1.12 eV at 300 K (the Varshni relation).

CHECK YOURSELF

1. G11 Convert 1.12 eV to joules.
2. G11 Why can't a completely full band carry current?
3. G12 Calculate the photon energy of green light at 530 nm. Is it absorbed by silicon? By GaN?
4. G12 Which band gap would an LED need to emit red light at 630 nm?
5. G12 Explain why window glass is transparent but a silicon wafer is not.
6. Varsity Why is silicon a poor light emitter, even though it absorbs light well?

WATCH & EXPLORE

- [PhET: Semiconductors](#) — see electrons fill bands and jump the gap. phet.colorado.edu/en/simulations/semiconductor **G11**
- [Siyavula Grade 12: Optical phenomena](#) — the photoelectric effect and atomic spectra (CAPS). siyavula.com **G12**
- [MIT OpenCourseWare 6.012: Microelectronic Devices and Circuits](#) — free lecture notes; the early lectures cover energy bands and carriers. ocw.mit.edu/courses/6-012-microelectronic-devices-and-circuits-spring-2009 **VARSITY**

IN PLAIN WORDS

Electrons in a solid can only have energies inside certain ranges, called bands. In silicon the lower band is full and the upper band is empty. Between them is a gap with no allowed energies. To carry current, an electron must jump this gap. Silicon's gap is small, so a bit of heat or light lets some electrons jump. Glass has a very big gap, so it hardly conducts. The gap also sets which light a material absorbs or gives off. That is why silicon turns sunlight into electricity, and why LEDs made from different materials glow in different colours.

SOURCES

1. Band gaps at 300 K and Varshni parameters: Ioffe Institute, *New Semiconductor Materials: Characteristics and Properties* (NSM archive), ioffe.ru/SVA/NSM; S. M. Sze & K. K. Ng, *Physics of Semiconductor Devices*, 3rd ed., Wiley, 2007, Appendix G.
2. Physical constants: CODATA 2018 recommended values, NIST.
3. C. Kittel, *Introduction to Solid State Physics*, 8th ed., Wiley, 2005, ch. 7–8 (bands, Bloch theorem, effective mass).

Electrons, Holes and Doping

GRADE 11

GRADE 12

VARSITY

Pure silicon is a poor conductor. Add one phosphorus atom for every few million silicon atoms and it becomes a useful one. This chapter shows how we count the charge carriers, how temperature changes them, and how doping puts us in control.

IN THIS CHAPTER YOU WILL LEARN

- what a hole is and why it acts as a positive charge
- intrinsic carrier concentration n_i and its huge temperature dependence
- why semiconductors conduct better when hot (metals do the opposite)
- n-type and p-type doping with donors and acceptors
- the mass-action law $np = n_i^2$
- the Fermi level, a first look

The hole: a missing electron that moves G11

When a bond breaks, the free electron moves off and leaves a gap in the bond. A neighbouring bonded electron can hop into that gap, which moves the gap one step the other way. Repeat this, and the gap wanders through the crystal. Because the gap is a *missing* negative charge, it behaves exactly like a **positive** charge $+e$ moving in the opposite direction. We call it a **hole**.

Analogy: one empty seat in a full row of a minibus taxi

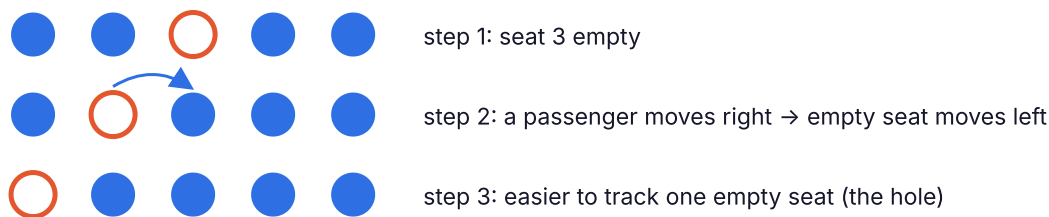


Figure 5.1 Electrons (blue) shuffling one way make the hole (red ring) travel the other way. Keeping track of one positive hole is far easier than tracking many bound electrons. In this book electrons are always filled blue circles and holes open red circles.

Intrinsic silicon and temperature G11 G12

Pure (**intrinsic**) silicon has free electrons and holes only from thermal bond-breaking, so their numbers are equal: $n = p = n_i$, where n is the electron concentration, p the hole concentration (both per cm^3) and n_i the **intrinsic carrier concentration**. At 300 K, $n_i \approx 1 \times 10^{10} \text{ cm}^{-3}$ for silicon. (Modern measurements give 9.65×10^9 ; many textbooks round to 10^{10} .) That is one free electron per 5×10^{12} atoms.

Thermal energy per particle is about kT , where $k = 8.617 \times 10^{-5}$ eV/K is Boltzmann's constant. At 300 K, $kT = 0.0259$ eV, much smaller than the 1.12 eV gap. The probability of a bond breaking follows a **Boltzmann factor**, which leads to

$$n_i = \sqrt{N_C N_V} e^{-E_g/2kT} \quad (5.1)$$

where N_C and N_V (about 10^{19} cm $^{-3}$ for Si, growing as $T^{3/2}$) count the available states near each band edge (see the Going deeper box). The exponential makes n_i extremely sensitive to temperature and band gap.

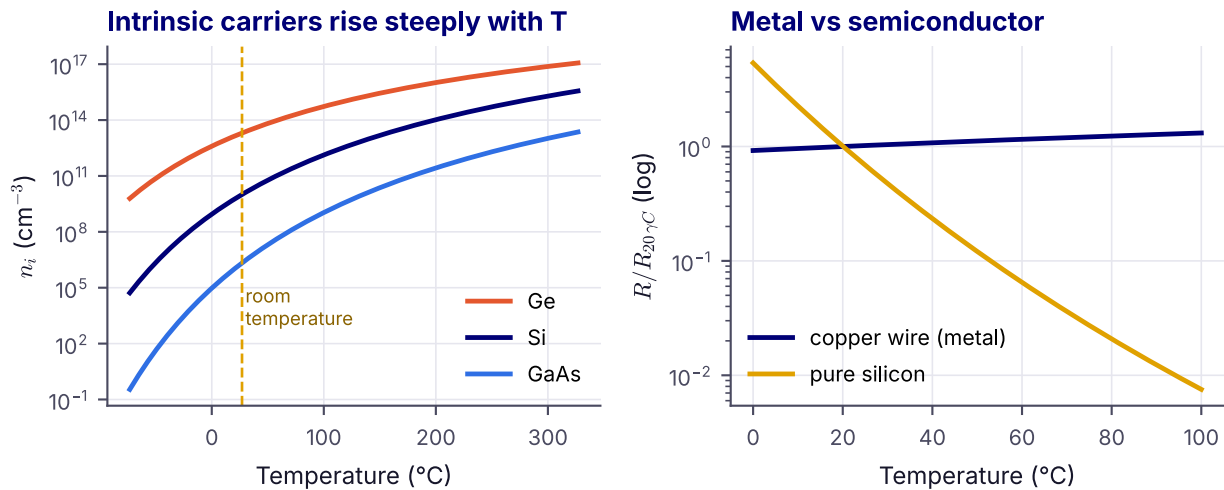


Figure 5.2 Left: n_i for Ge, Si and GaAs (model of Eq. 5.1 with temperature-dependent band gap, scaled to the 300 K values 2×10^{13} , 1×10^{10} and 2.1×10^6 cm $^{-3}$). Smaller gap means more carriers. Right: the resistance of a copper wire rises slowly with temperature, but the resistance of pure silicon falls steeply because n_i rises so fast.

WORKED EXAMPLE How much does n_i change from 27 °C to 77 °C?

Using Eq. 5.1 with $N_C N_V \propto T^3$ and E_g fixed at 1.12 eV:

$$\frac{n_i(350 \text{ K})}{n_i(300 \text{ K})} = \left(\frac{350}{300}\right)^{3/2} \exp\left[-\frac{1.12}{2 \times 8.617 \times 10^{-5}} \left(\frac{1}{350} - \frac{1}{300}\right)\right] = 1.26 \times 22.1 \approx 28$$

Answer. A 50 °C rise multiplies n_i by about 28. This is why chips leak more current when hot, and why semiconductor **thermistors** make sensitive thermometers.

Doping: taking control G11 G12

Doping means replacing a tiny fraction of silicon atoms with atoms from group 15 or group 13 (Figure 2.2).

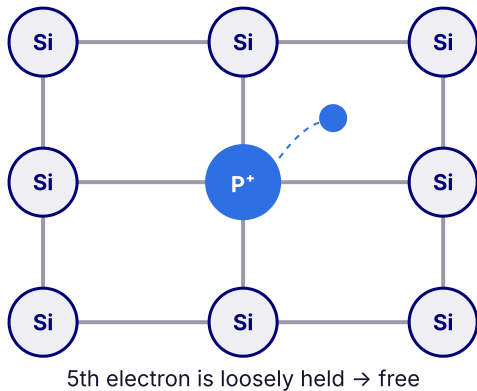


Figure 5.3 n-type: phosphorus has five valence electrons. Four form bonds; the fifth is bound by only about 0.045 eV, so at room temperature it is set free. The phosphorus ion left behind is fixed and positive. Phosphorus is a **donor**.

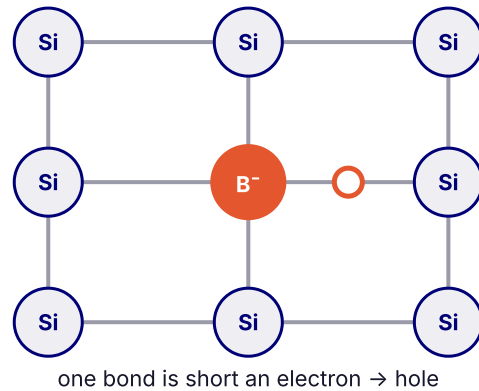


Figure 5.4 p-type: boron has three valence electrons, so one bond is missing an electron. A neighbouring electron easily fills it (about 0.045 eV), making the boron a fixed negative ion and creating a mobile hole. Boron is an **acceptor**.

Write N_D for the donor concentration and N_A for the acceptor concentration (atoms per cm^3). At room temperature nearly every dopant is ionised, so:

$$\text{n-type: } n \approx N_D \quad \text{p-type: } p \approx N_A \quad (5.2)$$

The carriers the dopant provides are the **majority carriers** (electrons in n-type, holes in p-type). The other kind are **minority carriers**, and there are very few of them, because doping does not only add carriers: it also speeds up recombination of the other kind. In thermal equilibrium the product of the two concentrations is fixed:

$$np = n_i^2 \quad (5.3)$$

This is the **mass-action law**. It holds for doped and undoped material alike, as long as the material is in equilibrium (no light, no applied current) and not too heavily doped.

WORKED EXAMPLE Carriers in doped silicon

Question. Silicon is doped with $N_D = 10^{16} \text{ cm}^{-3}$ phosphorus at 300 K. Find n and p .

Step 1. Majority electrons: $n \approx N_D = 1 \times 10^{16} \text{ cm}^{-3}$, a million times n_i .

Step 2. Minority holes from Eq. 5.3: $p = n_i^2/n = (10^{10})^2/10^{16} = 1 \times 10^4 \text{ cm}^{-3}$.

Answer. Electrons outnumber holes by 10^{12} to 1. Doping one atom in five million (Chapter 2) gave a million times more electrons and a million times fewer holes than pure silicon.

Table 5.1 Typical doping levels in silicon devices (orders of magnitude)

Doping (cm ⁻³)	Dopant atoms per Si atom	Name	Typical use
10 ¹⁴ –10 ¹⁵	1 in 50 million – 500 million	lightly doped	Wafer substrate, high-voltage drift regions, detectors
10 ¹⁶ –10 ¹⁷	1 in 0.5 – 5 million	moderate	Transistor wells and channels
10 ¹⁸ –10 ¹⁹	1 in 5 000 – 50 000	heavy (n ⁺ , p ⁺)	Source/drain regions
10 ²⁰ +	1 in 500 or more	degenerate	Contacts, tunnel junctions; the material behaves almost like a metal

COMMON MISTAKE

"n-type silicon is negatively charged." No: doped silicon is electrically **neutral**. Every free electron in n-type material is balanced by a fixed positive donor ion. The "n" and "p" describe which carrier is in the majority, not the overall charge.

A first look at the Fermi level G12 VARSITY

Band diagrams usually show one more line: the **Fermi level** E_F . Think of it as the "water level" of electrons: states well below E_F are almost certainly full, states well above are almost certainly empty. In intrinsic silicon E_F sits near the middle of the gap (written E_i). Donors push it up towards the conduction band; acceptors pull it down towards the valence band:

$$E_F - E_i = kT \ln \frac{n}{n_i} \quad (\text{n-type}), \quad E_i - E_F = kT \ln \frac{p}{n_i} \quad (\text{p-type}) \quad (5.4)$$

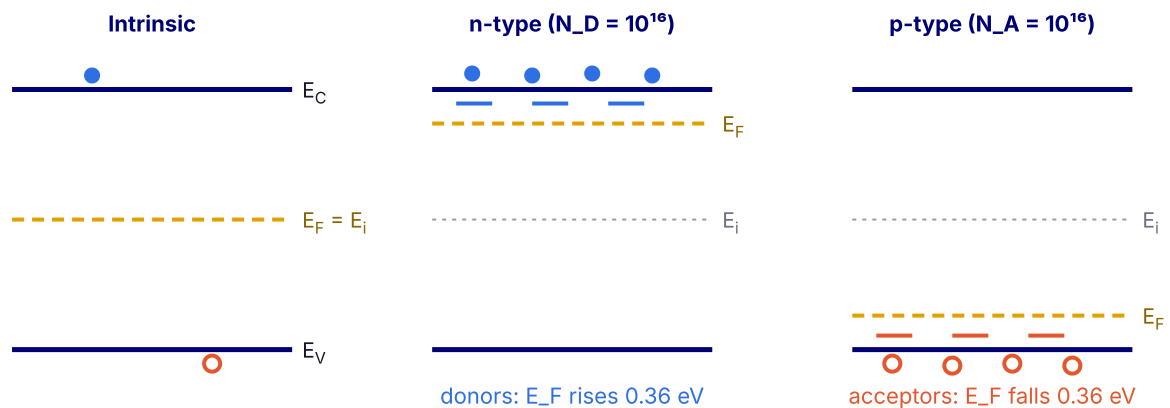


Figure 5.5 Band diagrams at 300 K. Short blue dashes are donor levels and short red dashes acceptor levels, each about 0.045 eV from the nearest band. The dashed amber line is the Fermi level. For 10¹⁶ cm⁻³ doping, Eq. 5.4 gives $|E_F - E_i| = 0.0259 \ln(10^6) = 0.36$ eV. When we join p- and n-type material, the difference between their Fermi levels creates the built-in voltage of a diode (Chapter 7).

GOING DEEPER VARSITY

The probability that a state at energy E is occupied is the **Fermi–Dirac function** $f(E) = 1 / (1 + e^{(E-E_F)/kT})$. Multiplying by the density of states, $g_C(E) \propto \sqrt{E - E_C}$, and integrating over the conduction band gives $n = N_C e^{-(E_C-E_F)/kT}$, with $N_C = 2 (2\pi m_n^* kT / h^2)^{3/2} \approx 2.8 \times 10^{19} \text{ cm}^{-3}$ for Si at 300 K. Similarly $p = N_V e^{-(E_F-E_V)/kT}$. Multiply the two and E_F cancels, which gives $np = N_C N_V e^{-E_g/kT} = n_i^2$. This is the derivation of Eqs. 5.1 and 5.3. The Boltzmann approximation is valid while E_F stays more than about $3kT$ inside the gap. Above roughly 10^{19} cm^{-3} doping it fails, and the material is called **degenerate**.

CHECK YOURSELF

1. **G11** Explain in your own words why a hole behaves like a positive charge.
2. **G11** Is n-type silicon negatively charged? Explain.
3. **G12** Silicon is doped with $N_A = 5 \times 10^{17} \text{ cm}^{-3}$ boron. Find p and n at 300 K.
4. **G12** Why does the resistance of a semiconductor fall as it warms up, while a metal's rises?
5. **G12** Germanium has a smaller band gap than silicon. Which has the larger n_i ? Why does that make Ge worse for chips that run hot?
6. **Varsity** Using Eq. 5.4, find how far E_F sits above E_i for $N_D = 10^{18} \text{ cm}^{-3}$.

WATCH & EXPLORE

- **PhET: Semiconductors** — dope a crystal with donors or acceptors and watch the carriers. phet.colorado.edu/en/simulations/semiconductor **G11**
- **Chenming Hu, *Modern Semiconductor Devices for Integrated Circuits*, ch. 1** — electrons, holes, doping and the Fermi function, with worked problems. [chu.berkeley.edu](https://www.eecs.berkeley.edu) **Varsity**
- **MIT OCW 6.012** — lecture notes on carrier statistics. ocw.mit.edu **Varsity**

IN PLAIN WORDS

When a bond breaks, you get a free electron and a hole. The hole is a missing electron, and it moves like a positive charge. Pure silicon has very few of these, but heat makes many more, which is why hot silicon conducts better. To take control, we add a tiny amount of another element. Phosphorus gives spare electrons (n-type). Boron gives holes (p-type). The more we add, the better it conducts. A simple rule links the two kinds: electrons times holes always equals the same number, n_i^2 .

SOURCES

1. n_i of Si = $9.65 \times 10^9 \text{ cm}^{-3}$ at 300 K: P. P. Altermatt et al., *J. Appl. Phys.* 93, 1598 (2003); the traditional rounded 10^{10} is from Sze & Ng (2007).
2. Dopant ionisation energies, N_C , N_V : S. M. Sze & K. K. Ng, *Physics of Semiconductor Devices*, 3rd ed., Wiley, 2007, ch. 1.
3. R. F. Pierret, *Semiconductor Device Fundamentals*, Addison-Wesley, 1996, ch. 2.

How Charges Move: Drift and Diffusion

GRADE 12

UNIVERSITY

Electrons and holes move for only two reasons: they are pushed by an electric field, or they spread out from where they are crowded. Every semiconductor device is built from these two motions, drift and diffusion.

IN THIS CHAPTER YOU WILL LEARN

- drift velocity and mobility
- why mobility falls with doping and temperature
- conductivity $\sigma = q(n\mu_n + p\mu_p)$
- velocity saturation in strong fields
- diffusion and the Einstein relation
- recombination, lifetime and diffusion length; the Hall effect

Drift: pushed by a field G12

Inside a crystal, a free electron rushes about randomly at roughly 10^5 m/s, bumping into vibrating atoms and impurities. With no field, these random motions cancel out and there is no current. Apply an electric field E (in V/cm) and each carrier gains a small average **drift velocity** along the field (holes) or against it (electrons):

$$v_d = \mu E \quad (6.1)$$

The constant μ is the **mobility**, measured in $\text{cm}^2/\text{V}\cdot\text{s}$. It says how easily a carrier moves. In lightly doped silicon at 300 K, $\mu_n \approx 1400 \text{ cm}^2/\text{V}\cdot\text{s}$ for electrons and $\mu_p \approx 470 \text{ cm}^2/\text{V}\cdot\text{s}$ for holes. Electrons are about three times more mobile, which is why n-channel transistors are faster than p-channel ones of the same size.



Figure 6.1 A field adds a small, steady sideways bias to the random zig-zag. Collisions with vibrating atoms (phonons) and with dopant ions limit how fast carriers drift, and that limit is what mobility measures.

Conductivity from carriers

Current density (current per unit area) from drift is $J = q(n\mu_n + p\mu_p)E$. Comparing with Ohm's law, $J = \sigma E$, gives the conductivity σ and resistivity $\rho = 1/\sigma$:

$$\sigma = q(n\mu_n + p\mu_p), \quad \rho = \frac{1}{\sigma} \quad (6.2)$$

Mobility is not constant. More dopant ions mean more scattering, so mobility falls as doping rises (Figure 6.2). Higher temperature means more lattice vibration, so mobility also falls as the device heats up. The combined effect of several scattering processes is estimated with **Matthiessen's rule**, $1/\mu = 1/\mu_{\text{lattice}} + 1/\mu_{\text{impurity}} + \dots$

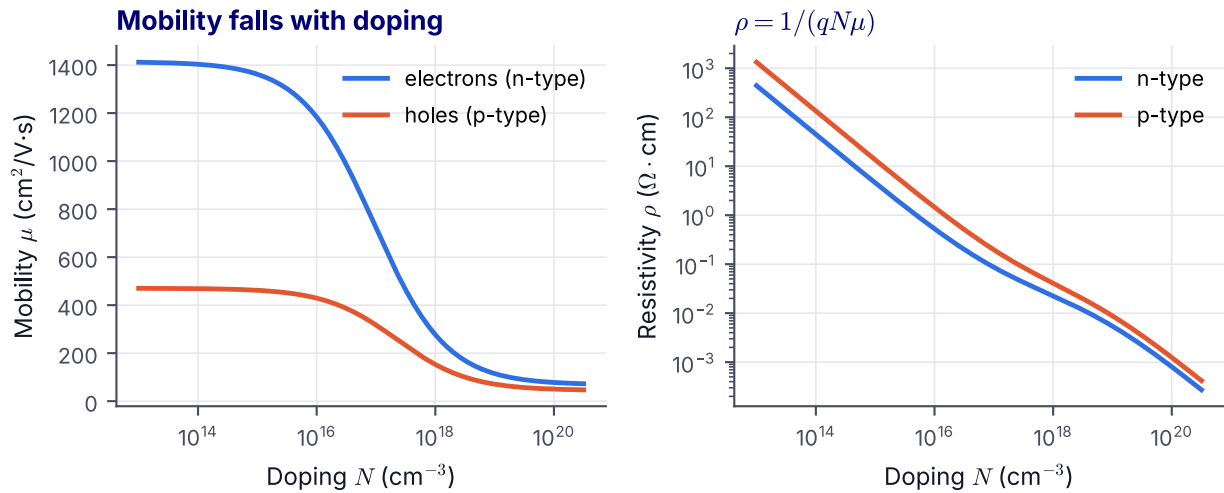


Figure 6.2 Silicon at 300 K. Left: mobility against doping (empirical Caughey–Thomas model with Masetti-type parameters). Right: resistivity from Eq. 6.2. Engineers read this "Irvin curve" backwards: measure a wafer's resistivity and get its doping.

WORKED EXAMPLE Resistivity of a doped wafer

Question. Silicon has $N_D = 10^{16} \text{ cm}^{-3}$. At this doping the model gives $\mu_n = 1184 \text{ cm}^2/\text{V}\cdot\text{s}$. Find ρ .

Step 1. $n \approx N_D = 10^{16} \text{ cm}^{-3}$; holes are negligible (10^4 cm^{-3} , Chapter 5).

Step 2. $\sigma = 1.602 \times 10^{-19} \times 10^{16} \times 1184 = 1.90 (\Omega \cdot \text{cm})^{-1}$.

Answer. $\rho = 1/1.90 = 0.53 \Omega \cdot \text{cm}$. Compare pure silicon at about $2 \times 10^5 \Omega \cdot \text{cm}$: doping one atom in five million cut the resistivity by a factor of about 400 000.

Velocity saturation Varsity

Eq. 6.1 fails in strong fields. Above roughly 10^4 V/cm , carriers lose energy to the lattice as fast as they gain it, and the drift velocity levels off at the **saturation velocity**, about 10^7 cm/s in silicon. Inside a modern transistor the field along a channel only tens of nanometres long easily exceeds 10^5 V/cm , so velocity saturation, not mobility, often limits the current (Chapter 9).

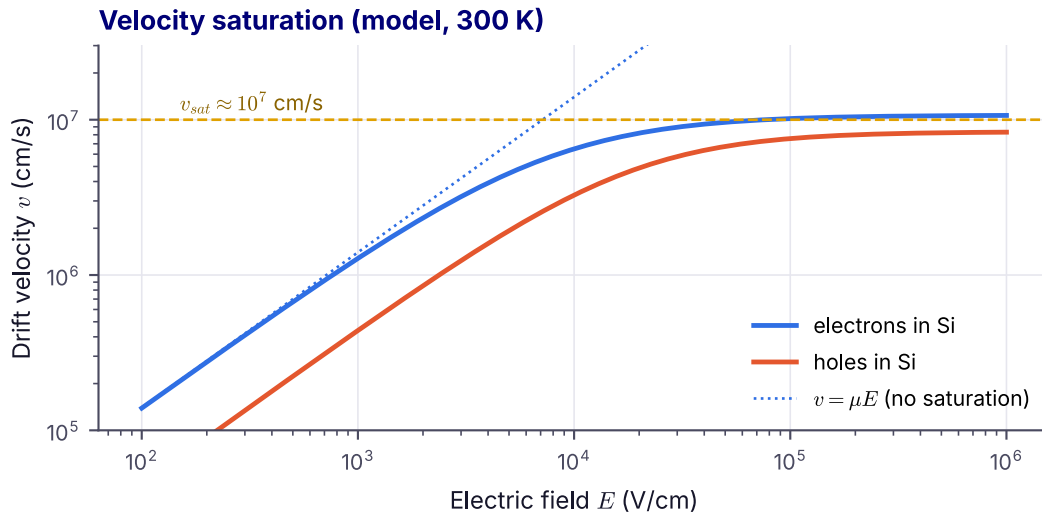


Figure 6.3 Drift velocity against field for electrons and holes in silicon at 300 K (Caughey–Thomas field model). At low field $v = \mu E$ (dotted); at high field the velocity saturates near 10^7 cm/s.

Diffusion: spreading out G12

Drop food colouring into still water and it spreads, even with no stirring. Random motion carries particles from where they are crowded to where they are sparse. Charge carriers do the same. If the electron concentration changes with position, a **diffusion current** flows down the concentration gradient:

$$J_{n,\text{diff}} = q D_n \frac{dn}{dx}, \quad J_{p,\text{diff}} = -q D_p \frac{dp}{dx} \quad (6.3)$$

D is the **diffusion coefficient** (cm^2/s). The signs differ because electrons carry negative charge. Diffusion needs no field at all: it is why a p–n junction works (Chapter 7).

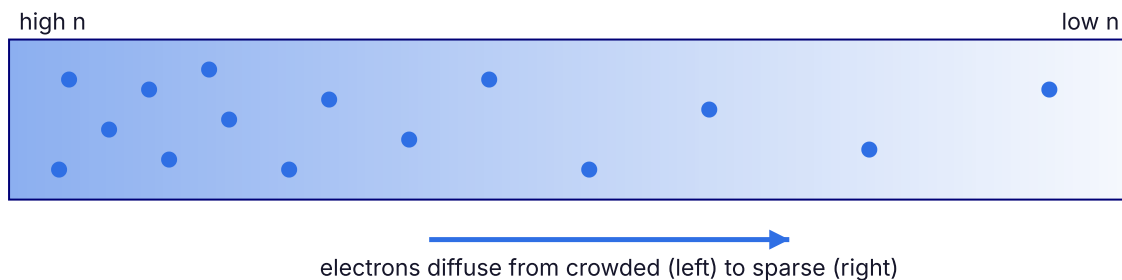


Figure 6.4 A concentration gradient alone drives a current. Because electrons are negative, electrons diffusing to the right make a conventional current to the left.

The Einstein relation VARSITY

Drift and diffusion are two faces of the same random motion, so D and μ are linked:

$$\frac{D}{\mu} = \frac{kT}{q} \approx 25.9 \text{ mV at } 300 \text{ K} \quad (6.4)$$

The total electron current is the sum of drift and diffusion, $J_n = qn\mu_n E + qD_n \frac{dn}{dx}$, and similarly for holes. These **drift–diffusion equations**, together with Poisson's equation for the field, are what device-simulation software (TCAD) solves.

Recombination, lifetime and diffusion length UNIVERSITY

Extra carriers do not last forever. An electron can fall back into a hole and both disappear:

recombination. If we inject extra minority carriers, for example with light or a forward-biased junction, they survive for an average **lifetime** τ , typically microseconds in good silicon and much shorter in defective material. In that time they diffuse an average distance called the **diffusion length**:

$$L = \sqrt{D\tau} \quad (6.5)$$

WORKED EXAMPLE How far do injected electrons travel?

For electrons in silicon with $\mu_n = 1184 \text{ cm}^2/\text{V}\cdot\text{s}$: $D_n = 0.02585 \times 1184 = 30.6 \text{ cm}^2/\text{s}$. With $\tau = 1 \mu\text{s}$, $L = \sqrt{30.6 \times 10^{-6}} = 5.5 \times 10^{-3} \text{ cm} = 55 \mu\text{m}$. A solar cell must collect carriers within about this distance of its junction, which is why solar-grade silicon must be very clean.

In silicon, most recombination happens through defects (traps) in the band gap, called **Shockley-Read-Hall** recombination. Direct band-to-band recombination that gives off light is rare in silicon (Chapter 4) but dominant in GaAs and GaN, which is why those make LEDs.

Measuring carriers: the Hall effect UNIVERSITY

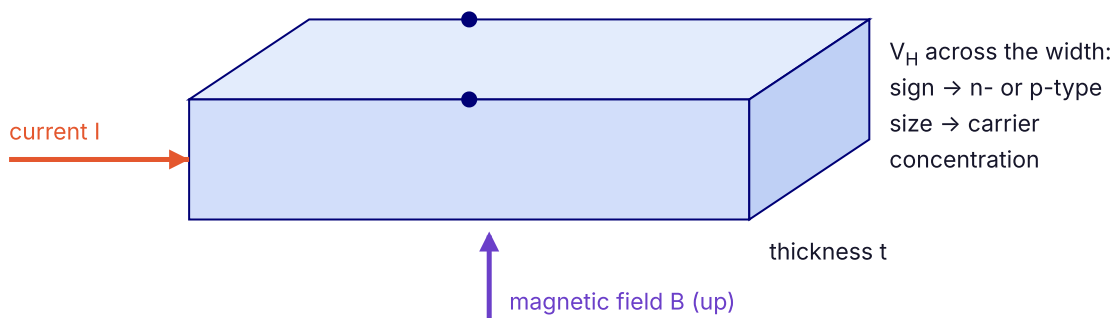


Figure 6.5 Hall measurement. The magnetic force pushes moving carriers to one side of the bar, building up a transverse voltage $V_H = IB/(qnt)$. Electrons and holes pile up on the same side, so the sign of V_H shows which carrier dominates.

WORKED EXAMPLE Hall voltage

A bar with $n = 10^{16} \text{ cm}^{-3} = 10^{22} \text{ m}^{-3}$ and thickness $t = 0.5 \text{ mm}$ carries $I = 1 \text{ mA}$ in $B = 0.5 \text{ T}$. Then $V_H = (10^{-3})(0.5)/(10^{22} \times 1.602 \times 10^{-19} \times 5 \times 10^{-4}) = 0.62 \text{ mV}$, easily measured. In copper, with about 10^{29} electrons per m^3 , the same bar would give 10^7 times less, which shows how few carriers semiconductors have.

CHECK YOURSELF

1. **G12** Find the drift velocity of electrons ($\mu_n = 1400 \text{ cm}^2/\text{V}\cdot\text{s}$) in a field of 100 V/cm .
2. **G12** Why are electrons in silicon faster than holes?
3. **G12** Explain the difference between drift and diffusion with an everyday example of each.
4. **VARSIITY** Find ρ for p-type Si with $N_A = 10^{16} \text{ cm}^{-3}$ and $\mu_p = 429 \text{ cm}^2/\text{V}\cdot\text{s}$.
5. **VARSIITY** Use the Einstein relation to find D_p for $\mu_p = 429 \text{ cm}^2/\text{V}\cdot\text{s}$ at 300 K , then L_p for $\tau = 2 \mu\text{s}$.
6. **VARSIITY** A Hall voltage is negative in a setup where holes would give a positive one. What type is the sample?

WATCH & EXPLORE

- [MIT OCW 6.012](#) — lecture notes on drift, diffusion and the continuity equation. [ocw.mit.edu](#) **VARSIITY**
- [Chenming Hu, ch. 2: Motion and Recombination of Electrons and Holes](#) — clear derivations and problems. [chu.berkeley.edu](#) **VARSIITY**
- [PhET: Semiconductors](#) — apply a voltage and watch carriers drift. [phet.colorado.edu/en/simulations/semiconductor](#) **G12**

IN PLAIN WORDS

Charges move for two reasons. They are pushed by a voltage (drift), or they spread from crowded places to empty places (diffusion). How easily they move is called mobility. Electrons in silicon move about three times more easily than holes. Too many dopant atoms or too much heat slows them down. If you push very hard, they hit a speed limit. Extra electrons and holes also meet and cancel out after a short time, so they only travel a small distance first.

SOURCES

1. Mobility model parameters: G. Masetti, M. Severi & S. Solmi, *IEEE Trans. Electron Devices* 30, 764 (1983); D. M. Caughey & R. E. Thomas, *Proc. IEEE* 55, 2192 (1967).
2. Saturation velocity and field-dependent mobility parameters: C. Canali et al., *IEEE Trans. Electron Devices* 22, 1045 (1975).
3. S. M. Sze & K. K. Ng, *Physics of Semiconductor Devices*, 3rd ed., Wiley, 2007, ch. 1; R. F. Pierret, *Semiconductor Device Fundamentals*, 1996, ch. 3.

The p–n Junction: Diodes, LEDs and Solar Cells

GRADE 12 VARSITY

Put p-type and n-type silicon side by side and something remarkable happens at the boundary: an invisible one-way valve for current appears. That valve is the diode, and run with light instead of a battery it becomes an LED or a solar cell.

IN THIS CHAPTER YOU WILL LEARN

- how the depletion region and built-in voltage form
- forward and reverse bias, with band diagrams
- the diode equation and the 0.6–0.7 V "knee"
- rectifiers: how a charger turns AC into DC
- how LEDs choose their colour
- how a solar cell makes power, and its limits

Making a junction G12

Imagine joining a block of p-type silicon (many holes) to a block of n-type silicon (many electrons). Electrons near the boundary diffuse into the p side and holes diffuse into the n side (Chapter 6), where they recombine. That leaves behind **fixed ions**: positive donor ions on the n side and negative acceptor ions on the p side. This thin zone, emptied of mobile carriers, is the **depletion region**. Its ions create an electric field pointing from n to p, which pushes back against further diffusion. Very quickly a balance is reached.

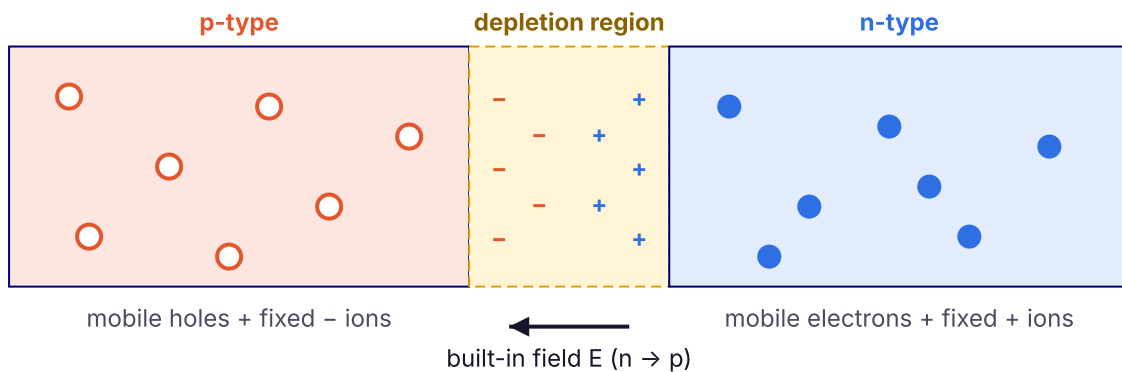


Figure 7.1 A p–n junction in equilibrium. Diffusion empties a thin region of mobile carriers, leaving fixed negative acceptor ions (–) and positive donor ions (+). Their field opposes further diffusion. For moderate doping the region is well under a micrometre wide.

The field corresponds to a voltage across the junction, the **built-in potential**:

$$V_{bi} = \frac{kT}{q} \ln \left(\frac{N_A N_D}{n_i^2} \right) \quad (7.1)$$

WORKED EXAMPLE Built-in voltage and depletion width

For $N_A = N_D = 10^{16} \text{ cm}^{-3}$ at 300 K: $V_{bi} = 0.02585 \ln(10^{32}/10^{20}) = 0.714 \text{ V}$.

The depletion width is $W = \sqrt{\frac{2\epsilon_{Si} V_{bi}}{q} \left(\frac{1}{N_A} + \frac{1}{N_D} \right)}$ with $\epsilon_{Si} = 11.7\epsilon_0 = 1.04 \times 10^{-12} \text{ F/cm}$, giving

$W = 0.43 \mu\text{m}$. The peak field is about $3 \times 10^4 \text{ V/cm}$. With 5 V of reverse bias (add 5 to V_{bi}), W grows to $1.2 \mu\text{m}$.

You cannot measure V_{bi} with a voltmeter. Contact potentials at the metal leads cancel it exactly; otherwise a diode on a shelf would be a battery.

Forward and reverse bias G12

Connect a battery with + to the p side (**forward bias**) and it opposes the built-in field. The barrier drops, and carriers pour across: holes into the n side and electrons into the p side. Reverse the battery (**reverse bias**) and the barrier grows. Only a tiny leakage current flows, carried by the few minority carriers.

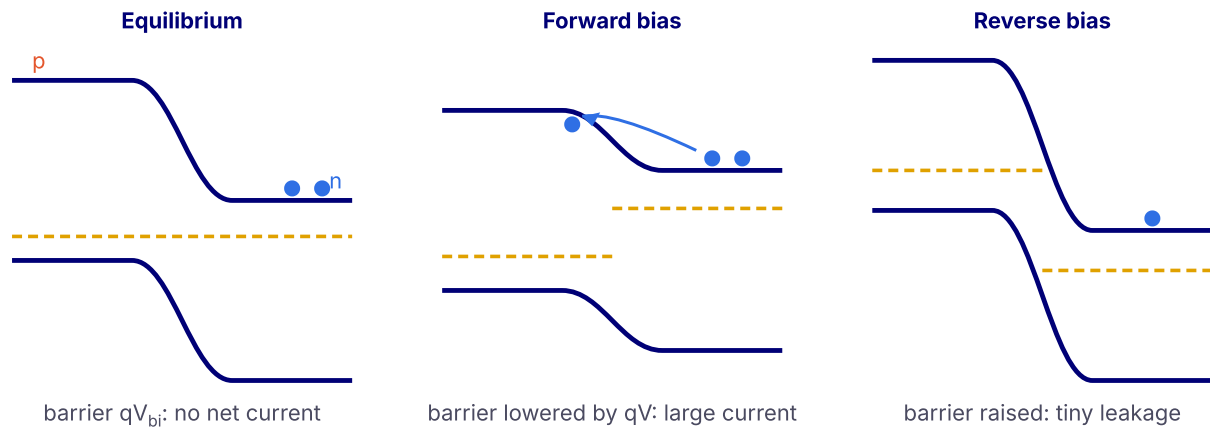


Figure 7.2 Energy bands across the junction (conduction band on top, valence band below; the dashed amber line is the Fermi level, which splits into two under bias). Electrons roll "downhill" in energy. Forward bias lowers the hill so many electrons can climb it; reverse bias makes the hill taller.

The diode equation

The number of carriers able to climb the barrier grows exponentially as the barrier drops (a Boltzmann factor again), giving the **Shockley diode equation**:

$$I = I_s \left(e^{qV/nkT} - 1 \right) \quad (7.2)$$

Here I_s is the tiny **saturation current** (around 10^{-14} A for a small silicon diode) and n is the **ideality factor**, between 1 and 2. Because of the exponential, current rises about tenfold for every 60 mV of extra forward voltage. In practice a silicon diode seems to switch on near 0.6–0.7 V.

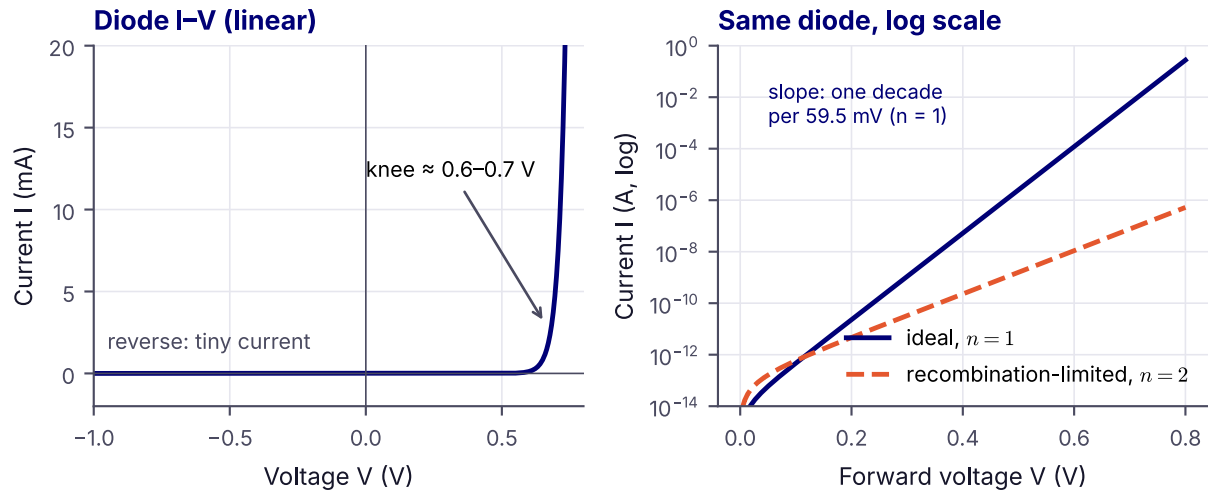


Figure 7.3 Model diode with $I_s = 10^{-14}$ A at 300 K. Left: on a linear scale the diode seems to "turn on" at a knee. Right: the same curve on a log scale is a straight line. At $n = 1$ it rises one decade per 59.5 mV, which shows that the knee is just the exponential becoming visible.

Diodes at work: chargers and power supplies G12

South African mains is 230 V AC at 50 Hz, but phones and electronics need steady DC. A **bridge rectifier** of four diodes steers both halves of the AC wave the same way through the load. A **smoothing capacitor** then fills in the dips.

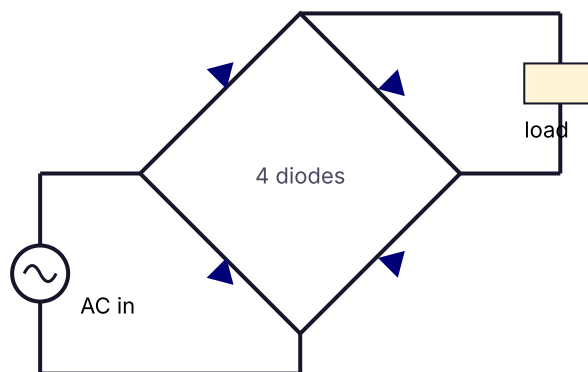


Figure 7.4 A bridge rectifier. On each half-cycle, two of the four diodes conduct, so current through the load always flows the same way.

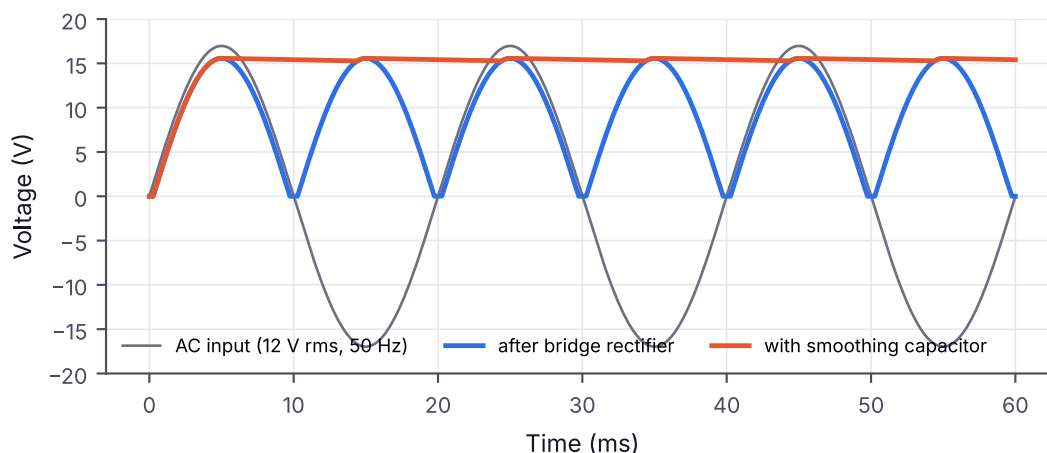


Figure 7.5 Simulated waveforms: 12 V rms, 50 Hz input (grey), full-wave rectified output, which loses about 1.4 V to two diode drops (blue), and output with a 4 700 μF smoothing capacitor and a 100 Ω load (red). Ripple is at 100 Hz.

WORKED EXAMPLE Choosing an LED resistor

A red LED needs about 2.0 V across it and 20 mA. From a 5 V USB supply, the resistor must drop $5.0 - 2.0 = 3.0$ V at 0.020 A, so $R = 3.0/0.020 = 150\ \Omega$. Without the resistor, the exponential diode law would let the current run away and destroy the LED.

Light out: LEDs G12

In a forward-biased junction made of a **direct-gap** material (Chapter 4), injected electrons and holes recombine and each recombination can emit a photon with energy close to the band gap. That is a **light-emitting diode**. The colour is set by the band gap, $\lambda \approx 1240/E_g$. By mixing elements, for example indium into GaN or aluminium into GaAs, engineers tune the gap and so the colour. Efficient **blue** LEDs from GaN were the missing piece for white lighting: a blue LED coated with a yellow phosphor looks white. Isamu Akasaki, Hiroshi Amano and Shuji Nakamura received the 2014 Nobel Prize in Physics for them.

Light in: solar cells and photodiodes G12 VARSITY

Run the LED backwards. A photon with $E > E_g$ is absorbed near the junction and creates an electron–hole pair. The built-in field sweeps the electron to the n side and the hole to the p side, which drives a current through an external load with no battery. This is the **photovoltaic effect**. A **photodiode** (in a camera sensor, a TV remote receiver or a fibre-optic link) uses the same physics, usually under reverse bias, to measure light.

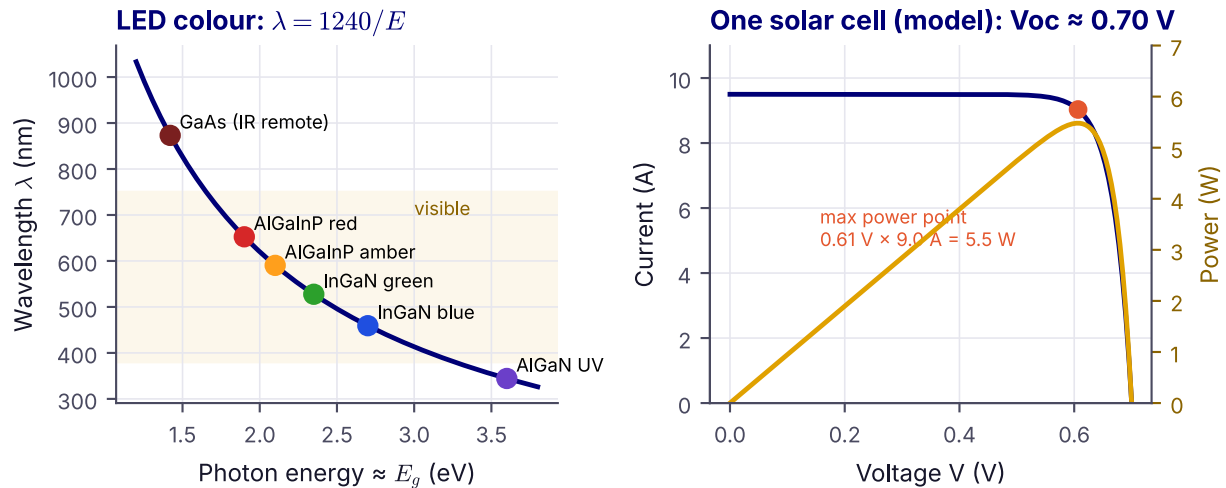


Figure 7.6 Left: LED wavelength set by band gap, with typical material systems. Right: a single silicon solar cell (model: $I_{sc} = 9.5$ A, $n = 1.2$, 300 K). Current is nearly constant until the voltage nears $V_{oc} \approx 0.70$ V. Power $P = VI$ peaks at the **maximum power point**, here about 0.61 V $\times$ 9.0 A = 5.5 W.

Three numbers describe a cell: short-circuit current I_{sc} , open-circuit voltage V_{oc} and **fill factor** $FF = P_{max}/(V_{oc}I_{sc})$. In the model above, $FF = 5.48/(0.70 \times 9.5) = 0.82$. If that cell is 156 mm square in standard sunlight ($1\,000$ W/m²), the light power in is $0.156^2 \times 1000 = 24.3$ W, so the efficiency is $5.48/24.3 = 22.5\%$. That is typical of good commercial silicon cells.

GOING DEEPER VARSITY

Why not 100 %? Photons below E_g pass through unused. Photons far above E_g waste their excess energy as heat. Some carriers recombine before they are collected. Balancing these for a single junction under sunlight gives the **Shockley–Queisser limit** of about 33 % at a band gap near 1.34 eV. Silicon (1.12 eV) has a practical ceiling near 29 %. Stacking materials with different gaps ("tandem" cells, such as perovskite on silicon) gets past the single-junction limit. Hot panels also lose efficiency, because V_{oc} falls by roughly 2 mV per °C per cell as n_i rises (Chapter 5).

SOUTH AFRICAN CONTEXT

South Africa has one of the strongest solar resources in the world. Large parts of the Northern Cape receive more than 6 kWh of solar energy per square metre per day on average. Rooftop solar grew very quickly during the load-shedding years. Every panel is a string of silicon p–n junctions like Figure 7.1, and every inverter's "maximum power point tracker" continually adjusts the load so the panel sits at the peak of the red power curve in Figure 7.6.

CHECK YOURSELF

- G12** What is in the depletion region, and why is it "depleted"?
- G12** Sketch the I–V curve of a silicon diode and label the knee voltage.
- G12** A blue LED needs 3.0 V and 15 mA. What resistor do you need from a 5 V supply?
- G12** Explain why a bridge rectifier's output voltage is about 1.4 V below the input peak.
- Varsity** Find V_{bi} for $N_A = 10^{18}$, $N_D = 10^{16}$ cm^{−3} at 300 K.
- Varsity** A diode carries 1 mA at 0.60 V ($n = 1$). Estimate the current at 0.72 V.
- Varsity** A 60-cell panel has cells with $V_{oc} = 0.70$ V. Estimate the panel's V_{oc} on a cold morning and at a cell temperature 40 °C higher.

WATCH & EXPLORE

- [PhET: Semiconductors](#) — build a p–n junction, apply forward and reverse bias and watch current flow one way only. phet.colorado.edu/en/simulations/semiconductor **G12**
- [PhET: Circuit Construction Kit: DC](#) — practise series resistors and measure voltage drops before building an LED circuit. phet.colorado.edu **G12**
- [Chenming Hu, ch. 4: PN and Metal–Semiconductor Junctions](#) — including LEDs and solar cells. chu.berkeley.edu **VARSAITY**

IN PLAIN WORDS

When p-type and n-type silicon touch, a thin layer forms between them with no free charges. It acts like a one-way gate. Push current one way and, after about 0.6 V, it flows easily. Push the other way and almost nothing flows. That gate is a diode. Chargers use diodes to turn AC into DC. Made from the right materials, a diode gives off light: an LED. Its colour depends on the material. Shine light on a diode and it makes electricity: a solar cell.

SOURCES

1. S. M. Sze & K. K. Ng, *Physics of Semiconductor Devices*, 3rd ed., Wiley, 2007, ch. 2 (p–n junctions) and ch. 13 (solar cells).
2. The Nobel Prize in Physics 2014, press release, nobelprize.org (Akasaki, Amano, Nakamura: efficient blue LEDs).
3. Shockley–Queisser limit: W. Shockley & H. J. Queisser, *J. Appl. Phys.* 32, 510 (1961); S. Rühle, *Solar Energy* 130, 139 (2016) (33.7 % at 1.34 eV).
4. South African solar resource: Solargis / World Bank Global Solar Atlas, globalsolaratlas.info (long-term average global horizontal irradiation maps).

The Transistor: From Switch to Logic and Memory

GRADE 12

VARSITY

A transistor is a valve for current, opened by a voltage. Join two of them the right way and you get a NOT gate. Join a few more and you can add numbers, remember bits and, with billions, run a phone.

IN THIS CHAPTER YOU WILL LEARN

- how a MOSFET switches, step by step
- threshold voltage and the transistor's I-V curves
- nMOS, pMOS and why CMOS uses both
- logic gates, truth tables and a binary adder
- how chips store bits: flip-flops, SRAM, DRAM, flash
- where a chip's power goes: $P = \alpha CV^2 f$

Why switches, and why transistors G12

Early computers used **relays** (magnet-driven switches) and then **vacuum tubes**, which were bulky, hot and failed often. The **transistor** does the same job in solid silicon with no moving parts. There are two main families. The **bipolar junction transistor** (BJT) uses a small base current to control a larger collector current (gain β , often around 100). It is still used in analogue and radio circuits. The **MOSFET** (metal-oxide-semiconductor field-effect transistor) is controlled by a *voltage* and draws almost no steady current at its control terminal. That property lets billions of them sit on one chip, so this book concentrates on it.

How a MOSFET works G12

An n-channel MOSFET (**nMOS**) has two heavily doped n^+ regions, the **source** and the **drain**, in a p-type silicon body. Between them, separated from the silicon by a very thin insulating layer (the **gate oxide**), sits the **gate** electrode.

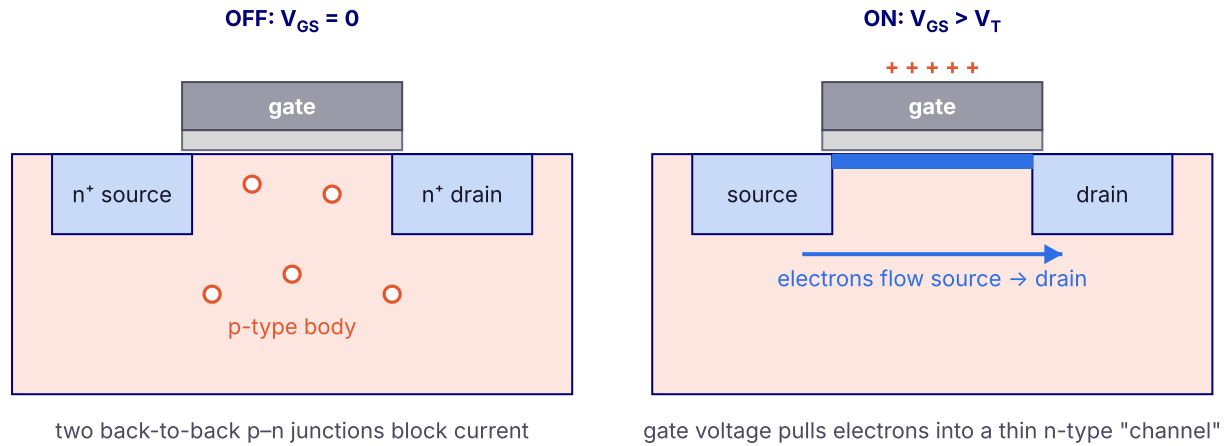


Figure 8.1 An nMOS transistor in cross-section. Off: source and drain are separated by p-type silicon, so no current flows. On: a positive gate voltage repels holes and attracts electrons to the surface just under the oxide, forming an **inversion layer** (blue) that connects source to drain. The gate's oxide insulates it, so the gate draws essentially no steady current.

The gate voltage at which the channel forms is the **threshold voltage** V_T , a few tenths of a volt in modern chips. Think of a tap: the gate is the handle, V_T is the point where water starts to flow, and the drain–source voltage is the water pressure.

The transistor's curves

For a simple long-channel MOSFET, once $V_{GS} > V_T$ and the drain voltage is large enough (the **saturation** region), the current follows the **square law**:

$$I_D = \frac{1}{2} \mu_n C_{ox} \frac{W}{L} (V_{GS} - V_T)^2 \quad (8.1)$$

W and L are the channel width and length, μ_n the electron mobility and C_{ox} the gate capacitance per unit area. Wider, shorter channels and thinner oxides give more current. Below threshold the current does not drop to zero. It falls exponentially, by a factor of ten for every 60 mV or more of gate voltage, which matters a lot for leakage (Chapter 9).

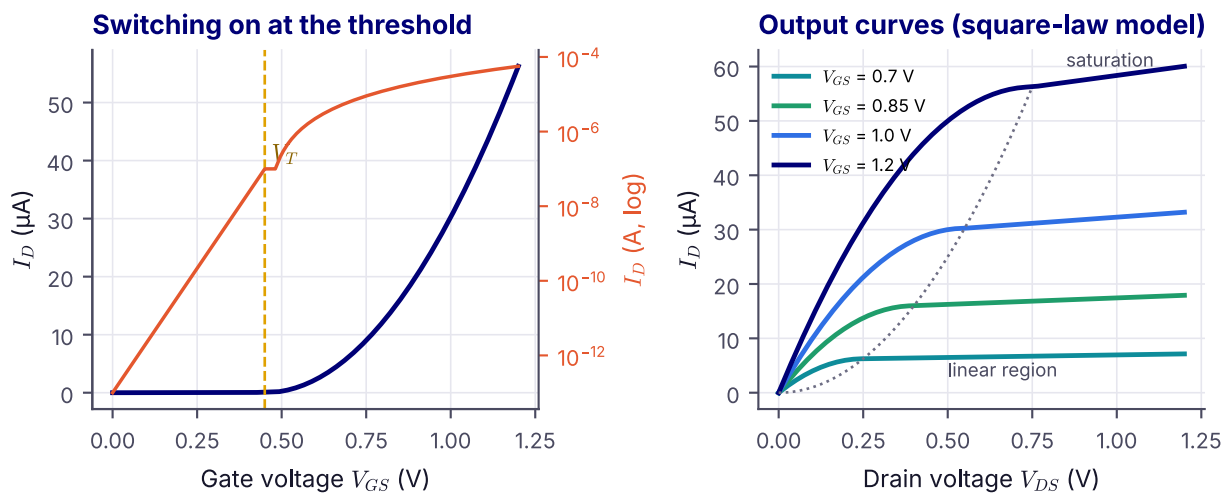


Figure 8.2 Illustrative long-channel nMOS ($V_T = 0.45$ V). Left: drain current against gate voltage, on linear (navy) and log (red) scales. The log scale shows the exponential subthreshold region below V_T . Right: output curves. At small V_{DS} the device acts like a resistor whose value the gate sets (linear region). At larger V_{DS} the current levels off (saturation).

nMOS, pMOS and CMOS

A **pMOS** transistor is the mirror image: p^+ source and drain in an n-type body. It turns on when its gate is pulled *below* its source, and conducts with holes. **CMOS** (complementary MOS) logic uses nMOS and pMOS in pairs, arranged so that in either steady state one of the pair is off. Ideally, no current flows from supply to ground except while switching. This low standby power is the main reason CMOS won.

Logic: the CMOS inverter and gates G12

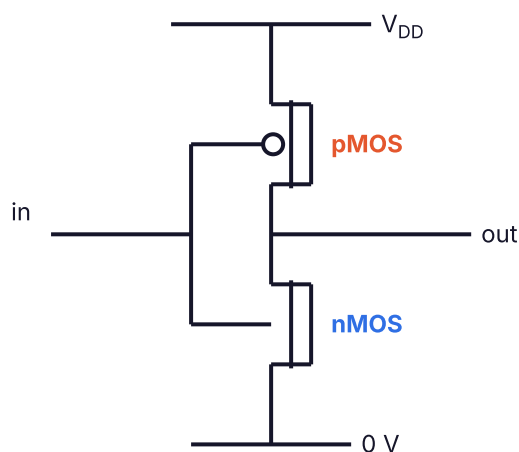


Figure 8.3 CMOS inverter. Input high: nMOS on, pMOS off, so the output is pulled to 0 V. Input low: pMOS on, nMOS off, so the output is pulled to V_{DD} . The bubble marks the pMOS gate.

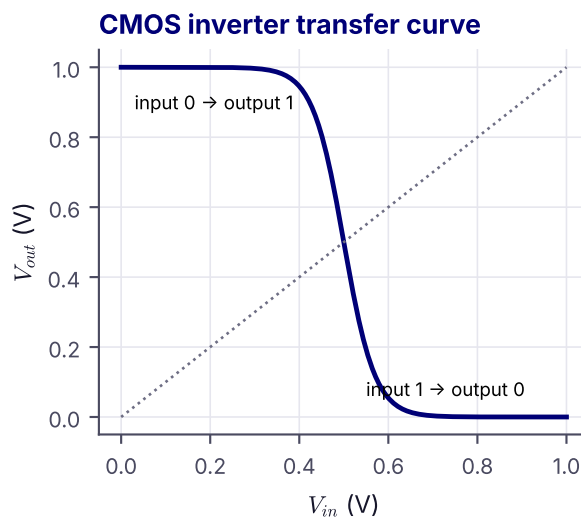


Figure 8.4 Inverter transfer curve (illustrative, $V_{DD} = 1$ V). The sharp drop in the middle restores logic levels: a slightly noisy input still produces a clean output, so signals can pass through billions of gates without fading.

Combining transistors in series and parallel gives other **logic gates**. A CMOS **NAND** gate puts two nMOS in series and two pMOS in parallel. Every digital function can be built from NAND gates alone.

Table 8.1 Truth tables (A, B inputs; 1 = high voltage, 0 = low)

A	B	NOT A	A AND B	A OR B	A NAND B	A XOR B
0	0	1	0	0	1	0
0	1	1	0	1	1	1
1	0	0	0	1	1	1
1	1	0	1	1	0	0

Doing arithmetic: the half adder

Adding two one-bit numbers gives a **sum** bit and a **carry** bit: $1 + 1 = 10$ in binary (sum 0, carry 1). From Table 8.1, the sum is A XOR B and the carry is A AND B. Chain these with an extra input for the incoming carry (a **full adder**) and you can add numbers of any length. Every calculator and processor adds this way.

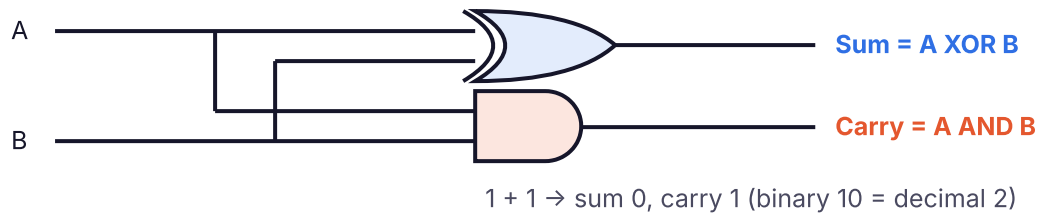


Figure 8.5 A half adder: one XOR gate (blue) and one AND gate (red). Built from CMOS it takes 14–20 transistors, depending on the design.

Remembering bits G12 VARSITY

Two inverters connected in a loop hold their state indefinitely: each keeps the other's input fixed. That loop is the heart of a **latch**. Add gates to control when it can change and you get a **flip-flop**, which updates only on the edge of a **clock** signal. Groups of flip-flops form **registers**. A clock ticking billions of times a second (gigahertz) keeps every part of a processor in step.

Table 8.2 How chips store bits

Memory	Cell	How it stores a bit	Keeps data without power?	Used for
SRAM	6 transistors	Two cross-coupled inverters (a latch)	No	Fast caches inside processors
DRAM	1 transistor + 1 capacitor	Charge on a tiny capacitor; leaks, so it is refreshed every few tens of milliseconds	No	Main memory ("RAM"), HBM for AI chips
Flash (NAND)	1 transistor with a charge-trap layer	Electrons trapped in an insulated layer shift V_T	Yes (years)	Phone storage, SSDs, memory cards

Where the power goes VARSITY

Each time a CMOS gate switches, it charges or discharges the capacitance C of the wires and gates it drives. The energy drawn from the supply per full charge–discharge cycle is CV^2 . With a fraction α of the chip's capacitance switching on each clock cycle at frequency f :

$$P_{\text{dynamic}} = \alpha C V_{DD}^2 f \quad (8.2)$$

WORKED EXAMPLE Why lower voltage matters so much

A processor has 100 nF of total switchable capacitance, $\alpha = 0.1$, $V_{DD} = 0.8$ V and $f = 3$ GHz. Then $P = 0.1 \times 100 \times 10^{-9} \times 0.8^2 \times 3 \times 10^9 = 19.2$ W. Dropping V_{DD} to 0.6 V gives $19.2 \times (0.6/0.8)^2 = 10.8$ W, a 44 % saving from a 25 % voltage cut. This V^2 term is why chip designers fight for every millivolt, and why phones lower voltage and clock speed when idle.

There is also **static** (leakage) power, from transistors that are "off" but not completely off (the subthreshold current of Figure 8.2), which grows quickly with temperature (Chapter 5). Chapter 9 explains why leakage stopped voltages from falling much further.

KEY IDEA

A processor is built from layers of abstraction: transistors → gates → adders and flip-flops → registers and arithmetic units → a CPU that fetches, decodes and executes instructions. Each layer hides the one below, which is how engineers manage designs with billions of transistors (Chapter 10).

TRY IT

Play [NandGame](#) in your browser. Starting from a single NAND gate, you build NOT, AND, OR, XOR, a half adder, a full adder, memory and finally a working computer. It is free and needs no installation. It is the best way to feel how gates become a processor.

CHECK YOURSELF

1. **G12** Name the three terminals of a MOSFET and say what the gate oxide does.
2. **G12** What is the threshold voltage? Use the tap analogy.
3. **G12** Explain why a CMOS inverter uses almost no power when its input is steady.
4. **G12** Write the truth table of a half adder and check it for all four input pairs.
5. **VARSIITY** Using Eq. 8.1, by what factor does I_D change if $V_{GS} - V_T$ doubles? If L halves?
6. **VARSIITY** A chip uses 12 W at 1.0 V and 2 GHz. Estimate its dynamic power at 0.8 V and 1.5 GHz.
7. **VARSIITY** Why must DRAM be refreshed but SRAM need not be?

WATCH & EXPLORE

- [NandGame](#) — build a computer from NAND gates, step by step. [nandgame.com](#) **G12**
- [Ben Eater: Build an 8-bit computer](#) — a famous video series building a working computer from simple logic chips on breadboards; the page links all videos and parts. [eater.net/8bit](#) **G12**
- [MIT OCW 6.012](#) — MOSFET models and CMOS inverter analysis. [ocw.mit.edu](#) **VARSIITY**

IN PLAIN WORDS

A MOSFET is a switch with three parts: source, drain and gate. Put a voltage on the gate and a thin path for electrons opens between source and drain. Remove it and the path closes. Pairing two kinds of these switches (CMOS) makes logic gates that hardly use power when they sit still. Gates can add numbers and, when looped, remember bits. A chip's power use depends on its voltage squared, so lowering the voltage saves a lot of energy.

SOURCES

1. D. Kahng, "Silicon-silicon dioxide field induced surface devices", IRE Solid-State Device Research Conference, 1960; S. M. Sze & K. Ng, *Physics of Semiconductor Devices*, 3rd ed., Wiley, 2007, ch. 6.
2. N. Weste & D. Harris, *CMOS VLSI Design*, 4th ed., Addison-Wesley, 2010, ch. 1 (CMOS logic, power).
3. B. Jacob, S. Ng & D. Wang, *Memory Systems: Cache, DRAM, Disk*, Morgan Kaufmann, 2008 (SRAM, DRAM refresh).

Inside the MOSFET: Physics, Limits and Scaling

VARSIY **PROFESSIONAL**

For fifty years the industry made transistors smaller, faster and cheaper at the same time. This chapter derives how a MOSFET really works, finds the physical limit that ended easy scaling, and explains the new 3-D transistor shapes the industry turned to.

IN THIS CHAPTER YOU WILL LEARN

- MOS band bending: accumulation, depletion, inversion
- where the threshold voltage comes from
- the drain-current equation and its assumptions
- the 60 mV/decade subthreshold limit, derived
- short-channel effects, leakage and high-k gates
- Moore's law, Dennard scaling, FinFET, GAA, and why node names are not sizes

The MOS capacitor **VARSIY**

Strip away source and drain and a MOSFET is a capacitor: metal gate, oxide insulator, p-type silicon. The gate voltage bends the silicon's bands near the surface. Writing ψ_s for the surface potential (how far the bands bend) and $\phi_F = (kT/q) \ln(N_A/n_i)$ for the distance from E_i down to E_F in the body, there are three regimes:

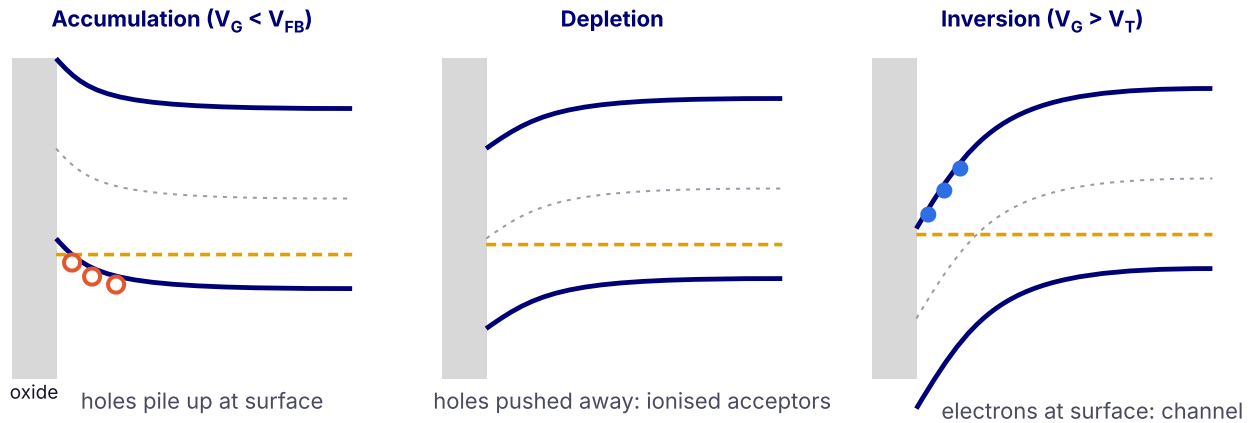


Figure 9.1 Bands near the oxide–silicon surface (left edge of each panel) in p-type silicon. The dashed amber line is E_F ; the dotted grey line is E_i . In **inversion** the bands bend so far that E_i drops below E_F at the surface. The surface then behaves as n-type: this electron layer is the transistor channel. The usual criterion for strong inversion is $\psi_s = 2\phi_F$.

At threshold, the gate must support the depletion charge $Q_{dep} = \sqrt{2\epsilon_{Si} q N_A (2\phi_F)}$ across the oxide capacitance $C_{ox} = \epsilon_{ox}/t_{ox}$, on top of bending the bands by $2\phi_F$:

$$V_T = V_{FB} + 2\phi_F + \frac{\sqrt{2\epsilon_{Si} q N_A (2\phi_F)}}{C_{ox}} \quad (9.1)$$

WORKED EXAMPLE Threshold voltage terms

For $N_A = 10^{17} \text{ cm}^{-3}$ and $t_{ox} = 2 \text{ nm}$ of SiO_2 at 300 K: $\phi_F = 0.0259 \ln(10^7) = 0.417 \text{ V}$; $C_{ox} = 3.9 \times 8.854 \times 10^{-14} / 2 \times 10^{-7} = 1.73 \text{ } \mu\text{F}/\text{cm}^2$; $Q_{dep} = 1.66 \times 10^{-7} \text{ C}/\text{cm}^2$, so $Q_{dep}/C_{ox} = 0.096 \text{ V}$. Hence $V_T - V_{FB} = 0.834 + 0.096 = 0.93 \text{ V}$. The flat-band voltage V_{FB} depends on the work-function difference between gate metal and silicon. Engineers choose gate metals (and adjust doping) to land V_T at the few tenths of a volt they want.

The drain current VARSITY

Assume (i) the field along the channel is much weaker than the field across it (the **gradual-channel approximation**), (ii) constant mobility, and (iii) inversion charge $Q_n(y) = -C_{ox}(V_{GS} - V_T - V(y))$ at position y along a channel of length L and width W . The drift current $I_D = W |Q_n(y)| \mu_n dV/dy$ is the same at every y , so integrating from source ($V = 0$) to drain ($V = V_{DS}$) gives

$$I_D = \mu_n C_{ox} \frac{W}{L} \left[(V_{GS} - V_T) V_{DS} - \frac{1}{2} V_{DS}^2 \right], \quad V_{DS} \leq V_{GS} - V_T \quad (9.2)$$

At $V_{DS} = V_{GS} - V_T$ the channel "pinches off" at the drain end and the current saturates at Eq. 8.1. In real short-channel devices, carriers hit velocity saturation (Chapter 6) before pinch-off, and the current becomes closer to linear in $V_{GS} - V_T$: $I_D \approx W C_{ox} v_{sat} (V_{GS} - V_T)$.

The 60 mV/decade wall VARSITY PRO

Below threshold there is no channel, only a small population of electrons that leak over a barrier. Their number follows the Boltzmann factor $e^{q\psi_s/kT}$. The gate controls ψ_s through a capacitive divider of C_{ox} and the depletion capacitance C_{dep} : $d\psi_s/dV_{GS} = 1/m$, with $m = 1 + C_{dep}/C_{ox}$. The gate voltage needed to change the current tenfold, the **subthreshold swing**, is therefore

$$S = \frac{dV_{GS}}{d(\log_{10} I_D)} = m \frac{kT}{q} \ln 10 \geq 59.5 \text{ mV/decade at 300 K} \quad (9.3)$$

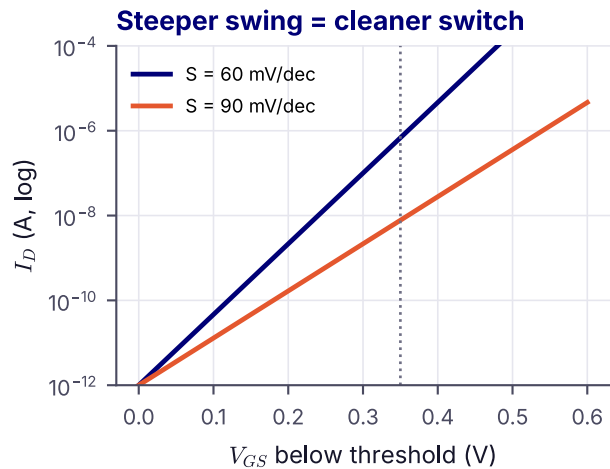
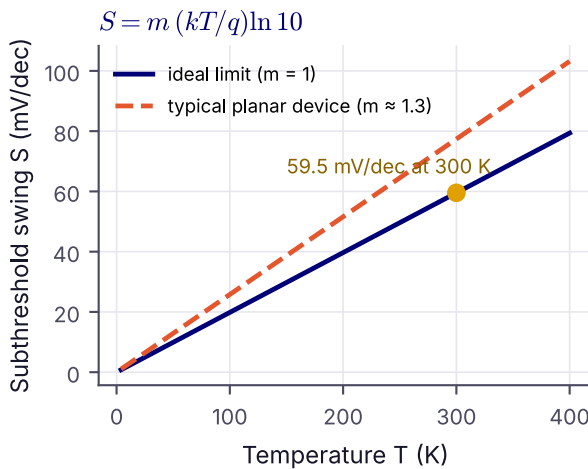


Figure 9.2 Left: the thermionic limit of Eq. 9.3 rises in proportion to temperature. Right: with a fixed 0.35 V between "off" and threshold, a 60 mV/dec device leaks far less than a 90 mV/dec one. You cannot beat $m = 1$ with an ordinary MOSFET at room temperature, because no gate can bend the bands by more than it is driven.

This matters because chips must switch between "on" and "off" currents differing by a factor of about 10^5 – 10^6 . At 60 mV/dec that needs at least 0.3–0.36 V of gate swing below threshold. So V_T cannot drop much below about 0.2–0.3 V, and the supply V_{DD} has stayed in the 0.7–1 V range for over a decade, instead of falling with every generation. Using the numbers from the worked example above, the depletion width at threshold is 104 nm, $C_{dep} = 0.10 \mu\text{F}/\text{cm}^2$, $m = 1.06$ and $S \approx 63 \text{ mV/dec}$, close to the limit. Research devices that aim to go below 60 mV/dec (tunnel FETs, negative-capacitance FETs) remain **research-stage**.

Scaling: Moore and Dennard VARSITY

In 1965 Gordon Moore observed that the number of components on a chip at minimum cost per component was doubling every year. In 1975 he revised it to every two years. In 1974 Robert Dennard and colleagues showed how to scale a MOSFET: shrink every length and voltage by a factor κ and raise doping by κ . Then each transistor gets faster (delay $\propto 1/\kappa$), uses $1/\kappa^2$ the power, and **power density stays constant**.

Table 9.1 Ideal constant-field (Dennard) scaling by a factor $\kappa > 1$

Quantity	Scales as	Quantity	Scales as
Length, width, t_{ox}	$1/\kappa$	Gate delay	$1/\kappa$
Voltage V_{DD} , V_T	$1/\kappa$	Power per transistor	$1/\kappa^2$
Doping	κ	Transistors per area	κ^2
Current per transistor	$1/\kappa$	Power density	constant

Around 2005 this broke down. V_T could not keep falling because of the 60 mV/dec wall, so V_{DD} stopped falling. Then power density rose with each shrink, and clock frequencies stopped climbing (the "power wall"). Since then, performance gains have come from more cores, specialised accelerators, better architecture and packaging, not simply faster clocks.

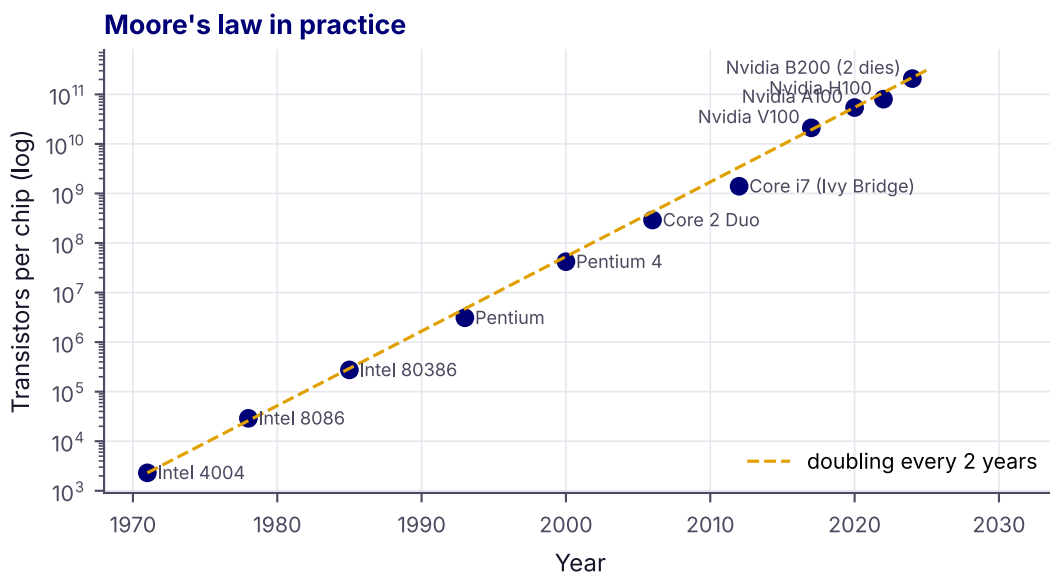


Figure 9.3 Transistors per chip for selected processors (company-stated counts; the 2024 point is a two-die product). The dashed line doubles every two years from the Intel 4004. Counts kept growing after Dennard scaling ended, through density gains, larger dies and multi-die packages.

Short channels, leakage and new materials PRO

When L shrinks to the same scale as the depletion widths of source and drain, the drain starts to control the channel barrier as well as the gate does. Two symptoms follow. In **drain-induced barrier lowering** (DIBL), a higher V_{DS} lowers the barrier and raises the off-current. In V_T **roll-off**, threshold falls for shorter L , so manufacturing variation in L becomes variation in leakage. The cure is to make the gate's grip on the channel stronger.

One route is thinner oxide. But SiO_2 below about 1.2 nm leaks unacceptably by quantum tunnelling. From 2007 the industry moved to **high- k** dielectrics such as hafnium oxide (relative permittivity around 20–25, versus 3.9 for SiO_2) with **metal gates**. A 3 nm HfO_2 layer with $k \approx 22$ gives the same capacitance as $3 \times 3.9/22 = 0.53$ nm of SiO_2 . This is its **equivalent oxide thickness** (EOT), but it is physically thick enough to stop tunnelling.

The other route is geometry: wrap the gate around the channel.

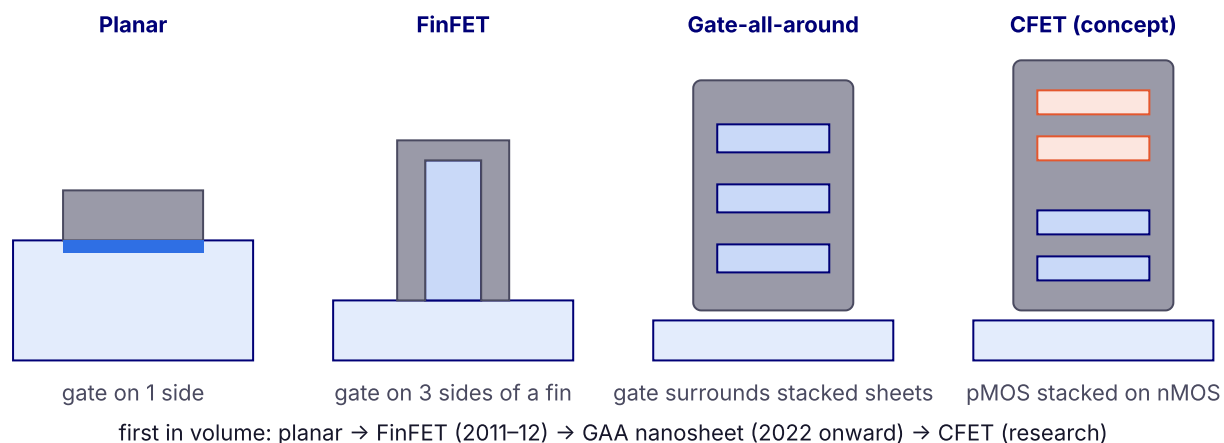


Figure 9.4 Transistor architectures viewed across the channel (current flows into the page). Grey is the gate, blue the silicon channel (n), red p-channels. Surrounding the channel on more sides gives the gate stronger control, which suppresses short-channel effects and lets L keep shrinking.

Table 9.2 Architecture milestones (company announcements; dates are first volume production as reported)

Milestone	Company	When	Status label
High- k metal gate in volume	Intel ("45 nm")	2007	PUBLISHED
FinFET ("tri-gate") in volume	Intel ("22 nm")	2011–12	PUBLISHED
Gate-all-around in mass production	Samsung ("3 nm")	2022	COMPANY CLAIM
Nanosheet GAA ("N2") first revenue	TSMC	Q2 2026 (TrendForce)	ANALYST REPORT
Backside power delivery	Intel (PowerVia, "18A"); TSMC planned on later nodes	2025 onward	COMPANY CLAIM

COMMON MISTAKE: NODE NAMES ARE NOT SIZES

"3 nm", "2 nm" and "18A" are **marketing labels** for process generations. No feature on the chip measures 3 nm or 2 nm. The physically meaningful numbers are pitches: the **contacted gate pitch** (centre-to-centre distance between gates), still several tens of nanometres on today's most advanced processes; the **minimum metal pitch**, in the low tens of nanometres; and transistor or SRAM-cell density per mm^2 . When comparing processes, ask for those, not the name.

GOING DEEPER PRO

Backside power delivery moves the thick power-supply wires to the back of the wafer, under the transistors. This frees the front-side wiring for signals and cuts voltage drop (IR drop). **Variability** is the other big modern problem. With only tens to hundreds of dopant atoms in a channel, random dopant fluctuation, line-edge roughness and metal-gate grain variation make each transistor slightly different. Mismatch between two nominally identical devices scales as $\sigma(\Delta V_T) = A_{VT}/\sqrt{WL}$ (Pelgrom's law), which is why SRAM and analogue circuits are designed statistically, to six sigma and beyond.

CHECK YOURSELF

1. **Varsity** Sketch the bands of a p-type MOS capacitor in inversion and mark where the electrons are.
2. **Varsity** Recompute $V_T - V_{FB}$ for $N_A = 10^{18} \text{ cm}^{-3}$ and $t_{ox} = 2 \text{ nm}$. Which term grew most?
3. **Varsity** Derive Eq. 9.2 from $I_D = W\mu_n C_{ox}(V_{GS} - V_T - V) dV/dy$.
4. **Varsity** What is the minimum subthreshold swing at 85 °C? Why do hot chips leak more?
5. **Pro** Explain in two sentences why the end of V_{DD} scaling ended Dennard scaling.
6. **Pro** Find the EOT of 2.5 nm of a dielectric with $k = 25$.
7. **Pro** Why does a gate-all-around transistor suppress DIBL better than a planar one?

WATCH & EXPLORE

- [Chenming Hu, ch. 5–7: MOS capacitor, MOSFET, MOSFETs in ICs](#) — written by the co-inventor of the FinFET; covers scaling and leakage. chu.berkeley.edu **Varsity**
- [MIT OCW 6.012](#) — MOSFET physics and models with problem sets. ocw.mit.edu **Varsity**
- [Asianometry](#) — YouTube channel with many well-researched videos on transistor history, FinFETs, gate-all-around and the industry; search the channel by topic. youtube.com/@Asianometry **Pro**

IN PLAIN WORDS

A transistor's gate pulls electrons to the surface to make a path. Physics sets a limit on how sharply it can turn off: at room temperature, you need about 60 millivolts to cut the current tenfold. Because of this, chip voltages stopped falling around 2005, and chips stopped getting much faster clocks. Engineers kept shrinking transistors by wrapping the gate around the channel, first on three sides (FinFET), now on all sides (gate-all-around). Names like "2 nm" are labels, not real sizes.

SOURCES

1. R. H. Dennard et al., "Design of ion-implanted MOSFETs with very small physical dimensions", *IEEE J. Solid-State Circuits* SC-9, 256 (1974).
2. G. E. Moore, "Cramming more components onto integrated circuits", *Electronics* 38(8), 1965; "Progress in digital integrated electronics", *IEDM* 1975.
3. Y. Taur & T. H. Ning, *Fundamentals of Modern VLSI Devices*, 2nd ed., Cambridge, 2009 (MOS electrostatics, subthreshold swing, short-channel effects).
4. M. J. M. Pelgrom et al., "Matching properties of MOS transistors", *IEEE JSSC* 24, 1433 (1989).
5. TrendForce press release, 9 Sep 2026: TSMC 2 nm contributed revenue for the first time in 2Q26 (trendforce.com/presscenter/news/20260909-13225.html).
6. Transistor counts (Fig. 9.3): manufacturers' published figures (Intel, Nvidia), as widely reported; treat as company claims.

How Chips Are Made and Designed

GRADE 12

VARSITY

PROFESSIONAL

A chip is printed, not assembled: hundreds of steps of light, chemistry and heat build billions of transistors at once on a wafer of pure silicon. Before that, engineers describe the whole chip in code and check it millions of times. This chapter follows both journeys.

IN THIS CHAPTER YOU WILL LEARN

- how sand becomes a 300 mm single-crystal wafer
- the core process steps and why cleanrooms matter
- photolithography and its physical limit
- the chip design flow, from code to tape-out
- a first piece of Verilog
- how students and start-ups can design real chips today

From sand to wafer G12



In the Czochralski process a seed crystal is dipped into molten silicon (about 1 414 °C) and slowly pulled up while rotating.

Figure 10.1 The silicon supply chain. Purity improves at each step until impurities are below about one part per billion. The final wafer is a single crystal, ready for chipmaking.

Chips are made in **cleanrooms**. A dust particle of 1 µm is huge compared with features tens of nanometres wide, and one particle can kill a chip. The air in the most critical areas is filtered to a few particles per cubic metre (ISO class 1–3), and workers wear full "bunny suits". Modern fabs move wafers between machines in sealed pods so that human hands never touch them.

The process steps VARSITY

A chip is built in layers. The bottom layers are the transistors (**front end of line**, FEOL). Above them, ten or more levels of copper wiring connect them (**back end of line**, BEOL). Every layer is made by repeating a few basic steps.

Table 10.1 The core toolbox of chipmaking

Step	What it does	Physics or chemistry
Oxidation	Grows SiO ₂ on silicon at 800–1 200 °C	Diffusion of O ₂ /H ₂ O through the oxide (Deal–Grove model)
Deposition	Adds thin films: CVD (from gases), PVD (sputtering), ALD (one atomic layer at a time), epitaxy (crystal on crystal)	Surface chemistry; ALD gives sub-nanometre control
Lithography	Prints the pattern onto light-sensitive resist	Optics; limited by diffraction (below)
Etching	Removes material where the resist is open; plasma (dry) etch gives vertical walls	Ion bombardment + reactive chemistry
Ion implantation	Shoots dopant ions into silicon at controlled depth	Ion stopping; then an anneal repairs damage and activates dopants
CMP	Chemical-mechanical polishing makes each layer flat	Slurry chemistry + abrasion
Metallisation	Copper wiring filled into etched trenches ("damascene")	Electroplating; barrier layers stop copper poisoning silicon
Metrology & inspection	Measures thickness, line width, overlay and defects after many steps	Optical and electron-beam measurement

Photolithography, step by step

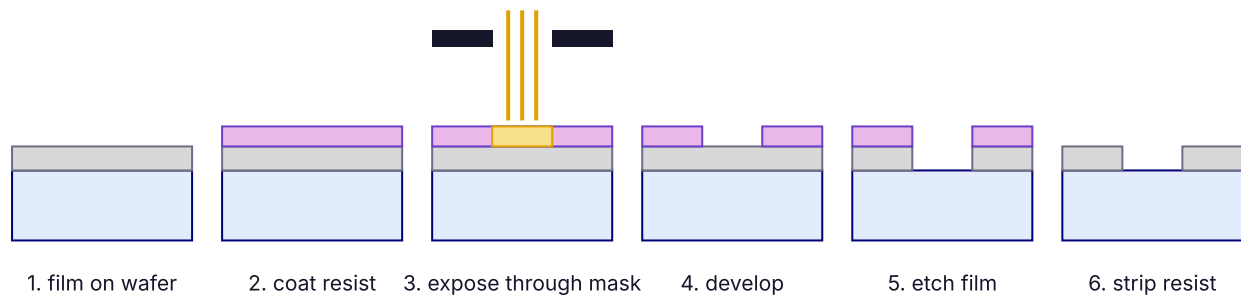


Figure 10.2 One patterning cycle (positive resist). Light passing through the clear part of the mask makes the resist soluble; developing removes it; etching transfers the opening into the film below. A leading-edge chip repeats cycles like this for many dozens of mask layers, each aligned to the ones beneath within a few nanometres ("overlay").

The limit: diffraction

Light bends (diffracts) around small openings, so a projection system cannot print arbitrarily small features. The smallest printable half-pitch is given by the **Rayleigh criterion**:

$$CD = k_1 \frac{\lambda}{NA} \quad (10.1)$$

Here λ is the wavelength, NA the numerical aperture of the lens (how wide a cone of light it collects), and k_1 a process factor with a physical floor of 0.25 for a single exposure. To print smaller, you can shorten the wavelength, raise NA , or lower k_1 with clever tricks. The industry did all three: from mercury lamps (365 nm) to deep-ultraviolet lasers (248 and 193 nm), then water between lens and wafer (**immersion**, NA up to 1.35), then **multiple patterning** (splitting one layer into two or four exposures), and finally **extreme ultraviolet** (EUV) at 13.5 nm.

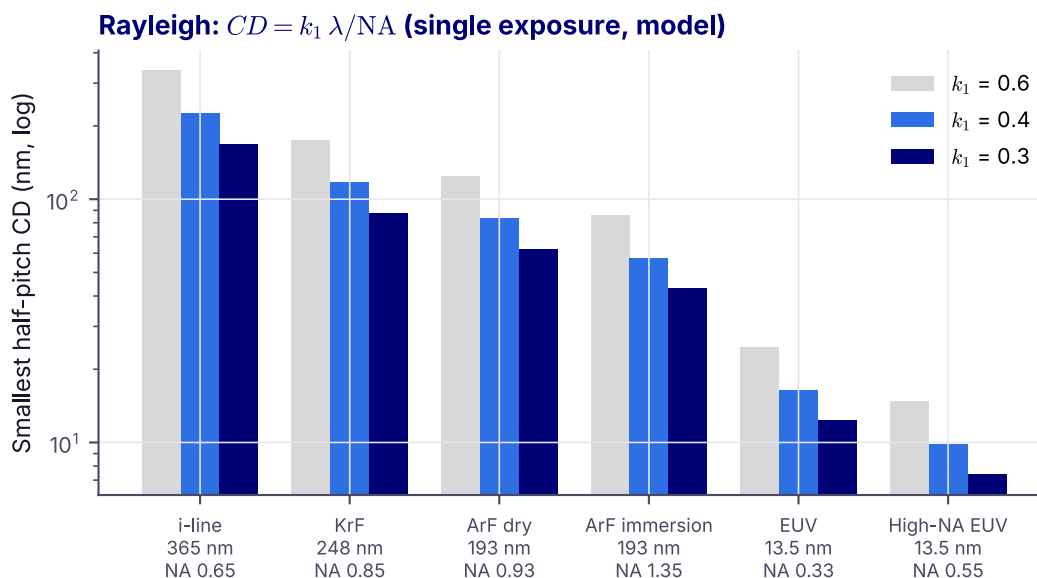


Figure 10.3 Half-pitch from Eq. 10.1 for each generation of lithography and three values of k_1 (model). Immersion ArF bottoms out near 40 nm half-pitch in one exposure; EUV reaches the low teens; High-NA EUV goes below 10 nm.

ABOUT THESE NUMBERS

EUV light is so strongly absorbed by everything, including air and glass, that the whole optical path is in vacuum and uses multilayer mirrors instead of lenses. The light comes from tin droplets hit by a powerful laser. A single supplier, ASML (Netherlands), makes all EUV scanners. Press reports in May 2026 put a standard EUV scanner (NA 0.33) at roughly US\$ 180 million and a High-NA scanner (NA 0.55) at roughly US\$ 380 million

PRESS/ANALYST ESTIMATE. Prices are not officially published.

Putting it together, and testing

A simplified CMOS process: grow isolation trenches between transistors; implant n-wells (for pMOS) and p-wells (for nMOS); form the gate stack; implant the source and drain regions and anneal; form contacts; build the copper wiring levels; seal with a passivation layer. Then each chip on the wafer is probed electrically (**wafer sort**); the wafer is sawn into **dies**; good dies are packaged; and packaged parts are tested again. From bare wafer to finished chip typically takes a few months at the leading edge. Yield, the fraction of good dies, decides the cost (Chapter 13).

Designing a chip Varsity Pro

Nobody draws billions of transistors by hand. Digital chips are **written**, in a **hardware description language** (HDL) such as Verilog, SystemVerilog or VHDL, then turned into transistors and wires by software tools (**EDA**, electronic design automation).

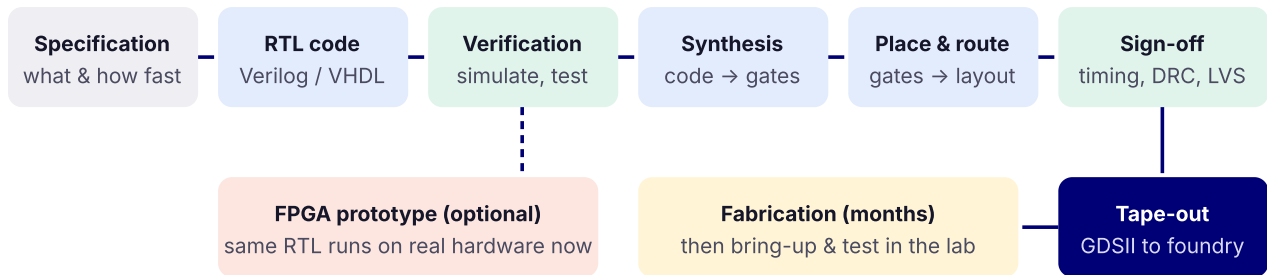


Figure 10.4 The digital design flow. Green boxes are checking steps; in a real project verification takes a large share of the engineering effort, because a mistake found after tape-out means paying for new masks and months of delay. The same RTL can be tested on an FPGA long before any silicon exists.

Your first Verilog

Here is a complete 8-bit counter. It describes hardware, not a sequence of instructions: on every rising clock edge the register `count` updates. Synthesis turns it into eight flip-flops and an adder (Chapter 8).

```

module counter8 (
    input wire    clk,      // clock
    input wire    rst,      // synchronous reset, active high
    input wire    enable,   // count only when 1
    output reg [7:0] count  // 8-bit output, 0..255
);
    always @(posedge clk) begin
        if (rst)
            count ≤ 8'd0;
        else if (enable)
            count ≤ count + 8'd1; // wraps from 255 back to 0
    end
endmodule

```

A **testbench**, more Verilog that drives `clk`, `rst` and `enable` and checks `count`, runs in a free simulator such as Icarus Verilog or Verilator before anything is built.

FPGA or custom chip (ASIC)?

An **FPGA** (field-programmable gate array) is a chip full of small look-up tables, flip-flops and programmable wiring that can be configured to become your circuit in seconds. An **ASIC** (application-specific integrated circuit) is a custom chip made for one design. FPGAs cost nothing to "re-spin" but are larger, slower and more power-hungry per function. ASICs need large up-front costs (masks, tools, engineering) but are cheap per unit at high volume. Chapter 14 works out the break-even point.

Open tools and shared wafers: chips for everyone

Since about 2020, open-source tools and open **process design kits** (PDKs, the rule books and device models of a manufacturing process) have made it possible for students to design real chips. Examples are SkyWater's SKY130 (130 nm), GlobalFoundries' GF180MCU (180 nm) and IHP's SG13G2

(130 nm SiGe BiCMOS), with free tools such as Yosys (synthesis), OpenROAD/OpenLane (place and route), Magic and KLayout (layout viewing and checks). **Multi-project wafers** (shuttles) split one wafer run among many designs, so each pays a small share of the mask cost.

ABOUT THESE PROGRAMMES (STATUS CHANGES, CHECK BEFORE RELYING)

Efabless, which ran the main open-source shuttle programme, closed in March 2025. That delayed the community projects that depended on it. Its founders then launched ChipFoundry, offering SKY130 shuttles at a reported US\$ 15 000 for 100 chips (eeNews Europe, May 2025). **Tiny Tapeout** continues: as of October 2026 its website lists shuttles on SkyWater ("SKY"), IHP and GlobalFoundries ("GF") processes, with per-tile pricing from a calculator on the site. **wafer.space** runs crowd-funded GF180MCU runs through Crowd Supply. Prices and schedules change often.

TRY IT

Start with [HDLBits](#): free, browser-based Verilog exercises that simulate your answer instantly. Then read [tinytapeout.com](#). A small digital design such as the counter above fits in one Tiny Tapeout tile, and a growing number of school and university students have had their own designs manufactured this way.

SOUTH AFRICAN CONTEXT

South Africa has no leading-edge wafer fab, and building one is not realistic: a single EUV scanner costs more than many national research budgets. But every step in Figure 10.4 except fabrication can be done from a laptop in Durban, Johannesburg or Cape Town. That is the route open to small South African design companies, Naple Semiconductor included: design, verify, prototype on FPGA, and manufacture through shuttles or commercial foundries abroad. Chapter 11 shows a worked example.

CHECK YOURSELF

1. **G12** List the steps from quartz sand to a polished wafer.
2. **G12** Why must chips be made in cleanrooms?
3. **Varsity** Using Eq. 10.1, find the half-pitch for EUV ($\lambda = 13.5$ nm, $NA = 0.33$) with $k_1 = 0.35$. How much does High-NA (0.55) improve it?
4. **Varsity** Put these in order: synthesis, RTL, tape-out, place and route, specification, sign-off.
5. **Varsity** Modify the Verilog counter so it counts down and stops at zero.
6. **Pro** Give two reasons a start-up might ship a product on an FPGA before making an ASIC.

WATCH & EXPLORE

- ▶ [The Ridiculous Engineering of the World's Most Important Machine](#) — Veritasium, on ASML's EUV lithography machines. youtube.com/watch?v=MiUHjLxm3V0 **G12**
- ▶ [Branch Education](#) — animated explainers; search the channel for "How are Microchips Made?". youtube.com/@BranchEducation **G12**
- [HDLBits](#) — practise Verilog in the browser with instant checking. hdlbits.01xz.net **Varsity**
- [Tiny Tapeout](#) — get your own small design manufactured on a shared chip; guides for beginners. tinytapeout.com **Varsity**
- [wafer.space GF180MCU runs](#) — crowd-funded open-PDK manufacturing runs (check current run and terms). crowdsupply.com/wafer-space **Pro**

IN PLAIN WORDS

Chips start as sand. The sand is cleaned into super-pure silicon, grown into one giant crystal and sliced into thin wafers. Then, in rooms cleaner than a hospital, machines add and remove very thin layers again and again. They use light to print patterns, a bit like a photo. Light can only print things so small, so the newest machines use extreme ultraviolet light and cost hundreds of millions of dollars each. Before any of that, engineers write the chip as code, test it carefully, and can try it on a reprogrammable chip (an FPGA). Today, students can design a small chip and have it made on a shared wafer.

SOURCES

1. S. Wolf & R. N. Tauber, *Silicon Processing for the VLSI Era*, vol. 1, Lattice Press; J. D. Plummer, M. D. Deal & P. B. Griffin, *Silicon VLSI Technology*, Prentice Hall, 2000.
2. B. E. Deal & A. S. Grove, "General relationship for the thermal oxidation of silicon", *J. Appl. Phys.* 36, 3770 (1965).
3. EUV tool prices: A. Shilov, "ASML's roadmap for chipmaking lithography tools examined", *Tom's Hardware*, 1 May 2026.
4. N. Flaherty, "Tiny Tapeout sees industrial boost as it recovers from eFabless closure", *eeNews Europe*, 16 May 2025.
5. Tiny Tapeout website, tinytapeout.com (accessed 15 Oct 2026); wafer.space GF180MCU runs, crowdsupply.com (accessed 15 Oct 2026).

Case Study: Raven-01, a Chip to Warn of Sewer Blockages

GRADE 12

VARSITY

PROFESSIONAL

This chapter follows a real South African chip-design project from a community problem to a verified digital design, using only the physics and methods in this book. It shows how a small team decides what a chip should do, proves it works before building anything, and stays honest about what is still unknown.

IN THIS CHAPTER YOU WILL LEARN

- how a real-world problem becomes a chip specification
- how a mass-balance equation decided what to measure
- why temperature matters to an ultrasonic distance sensor
- how designs are verified with a "digital twin"
- what simulation can and cannot prove
- the path from code to FPGA to silicon

WHAT THIS CASE STUDY IS, AND IS NOT

Raven-01 is a research design by Naple Semiconductor. It is a digital chip design that has been **simulated and implemented in FPGA design tools only**. As of October 2026 no hardware has been built or field-tested. All site values are estimates, not measurements. Detailed design information is kept confidential; this chapter shares the physics, the method and the headline simulation results for teaching.

The problem G12

In parts of eThekwin (Durban), sewer manholes overflow without warning, spilling raw sewage into streets and streams. Community interviews by Mableroot, a venture working on this problem with Naple, found that residents often wait days for repairs. In one documented incident, the overflow took 4 days to be reported and 14 days to be fixed.

The idea: put a small battery-powered sensor in each manhole that measures the water level and raises an alarm *before* it overflows. Mableroot owns the sensor device; Naple's job is the **chip** inside it. The first engineering question is not "which chip?" but "**what must the chip detect, and how early?**"

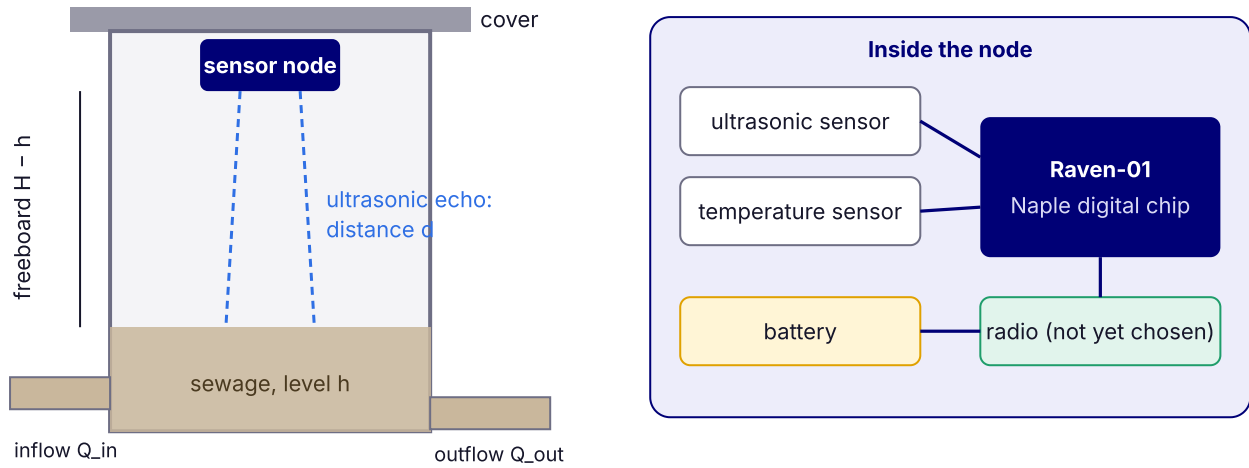


Figure 11.1 The system and its boundary. Off-the-shelf parts (sensors, radio, battery) surround Naple's part, the digital chip. The chip schedules measurements, powers the ultrasonic sensor only while measuring, converts echo time to distance and water level, learns the manhole's normal daily pattern and sends short alarm and "I'm alive" messages.

Physics decides what to measure G12 VARSITY

How fast does a blocked manhole fill?

Treat the manhole as a vertical tank of plan area A_m . Conservation of mass for an incompressible liquid gives

$$A_m \frac{dh}{dt} = Q_{in} - Q_{out} \quad (11.1)$$

If the outlet is completely blocked ($Q_{out} = 0$) and inflow is steady, the time to rise from an alarm level h_a to the rim H is

$$t_w = \frac{A_m (H - h_a)}{Q_{in}} \quad (11.2)$$

Check the units: $\text{m}^2 \times \text{m} \div (\text{m}^3/\text{s}) = \text{s}$. Check the limits: no inflow gives infinite warning; a tiny chamber gives none. For chambers 1.0–2.0 m across, each metre of free space lasts only 13–52 minutes at an inflow of 1 litre per second.

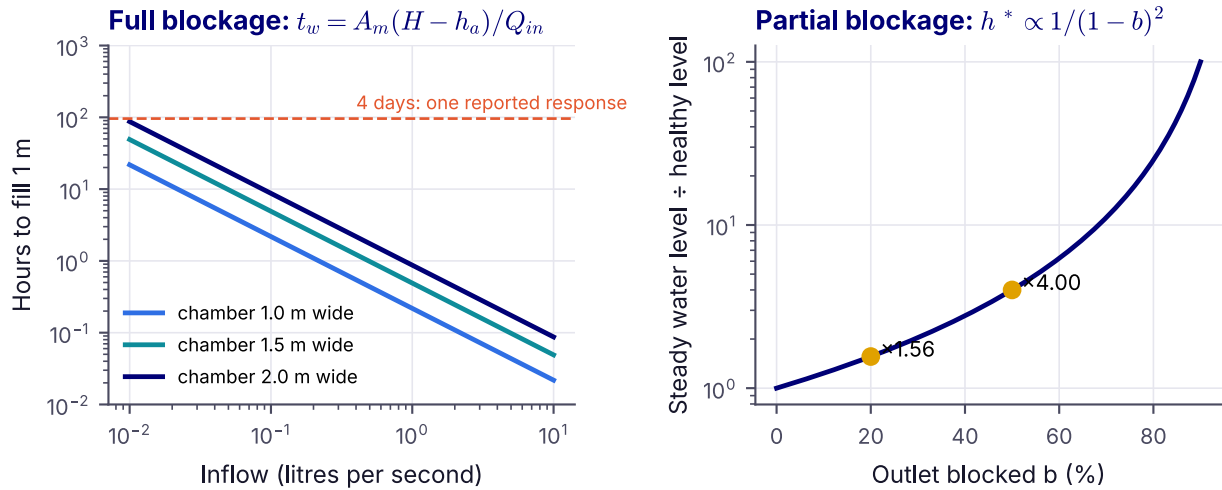


Figure 11.2 Left: Eq. 11.2 per metre of freeboard. With a *full* blockage, overflow comes within minutes to a few days, far shorter than the reported 4-day response time. Right: with a *partial* blockage b of the outlet (idealised orifice-like flow law), the steady water level rises by a factor $1/(1-b)^2$: 1.56 times at 20 % blocked.

KEY IDEA: THE DECISION THE EQUATION MADE

A warning hours before overflow is useless to a service that takes days to respond. So the chip's real job is to notice a **partial** blockage while it is still small, days before overflow. It does this by learning what "normal" looks like at each time of day and flagging a sustained rise above it. One equation changed the whole specification.

Measuring distance with sound

The sensor times an ultrasonic pulse to the water and back: $d = ct/2$. But the speed of sound in air depends on temperature:

$$c(T) = 331.3 \sqrt{1 + \frac{T}{273.15}} \text{ m/s} \quad (T \text{ in } ^\circ\text{C}) \quad (11.3)$$

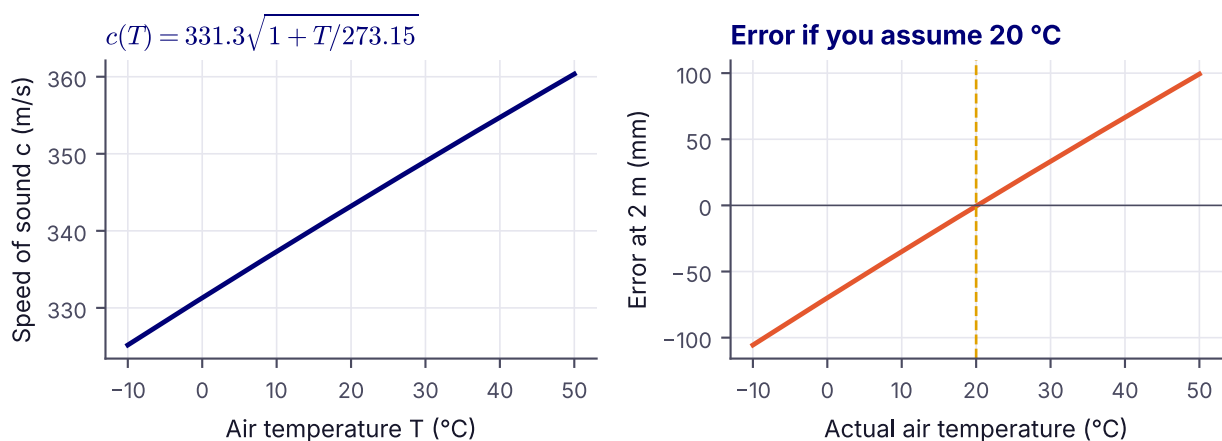


Figure 11.3 Speed of sound changes by about 0.17 % per $^\circ\text{C}$. Assuming 20°C when the air is really 30°C causes a 34 mm error at 2 m, larger than the healthy water depth in a small sewer. So the node also measures temperature, and the chip corrects every reading.

The hardest physics problem

Estimates from Manning's open-channel flow equation suggest that healthy flow in a small (160 mm) sewer pipe is very shallow: roughly 5 to 56 mm deep, depending on how many households feed it. That is comparable to the ± 10 mm accuracy of a low-cost ultrasonic sensor. In simulation, the smallest blockage the chip could detect grew from about 18 % at a healthy depth of 20 mm to about 46 % at 5 mm (with a placeholder sensor-noise model). **Whether this works in a real Illovo manhole cannot be settled by analysis; it needs site measurements.**

Engineering method: prove it before you build it Varsity Pro

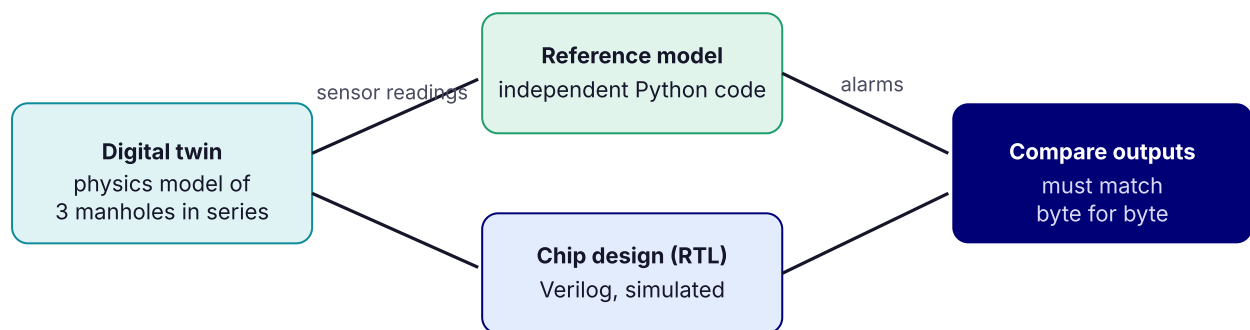


Figure 11.4 The verification strategy. A physics-based digital twin generates realistic sensor readings for healthy days, slowly growing blockages, pump-station failures and rain. Two independent implementations, a reference model and the actual chip design, process the same readings, and their outputs must be identical. Deliberately planted bugs ("mutation testing") check that the tests would catch real mistakes.

Table 11.1 Headline simulation results with estimated Illovo values (Naple research report RA-001, October 2026). These are simulation results, not field measurements.

Scenario	Result in simulation
Healthy flow, 9 days after a learning period	No false alarms on any of the three manholes
Blockage growing over 7 days	Alarm when about 18 % blocked, about 119 hours before the first overflow
Pump-station failure	Warning about 1.4 hours before overflow at the lowest manhole
Rain	All manholes rise together; a single node cannot tell rain from blockage, so the municipal platform must compare nodes
Chip design vs reference model	Byte-identical outputs over all scenarios and long random sequences
FPGA implementation (design tools only)	Uses under 40 % of a small low-cost FPGA and meets its timing target with margin
Sensor power	Switching the ultrasonic sensor on only while measuring cuts its average current from about 8 mA to about 15 μ A (once a minute)

From code to silicon, and the economics Pro

The same Verilog can run on an FPGA board for bench and field trials. But an FPGA draws far too much power for years on a battery, so a deployed node would need a low-power custom chip. Following Chapters 10 and 14, the realistic route for a small South African company is: FPGA prototype → a multi-project-wafer shuttle on a mature open process → a volume foundry run only if orders justify the masks. Each step is chosen by break-even arithmetic, not by ambition.

What is still unknown Varsity

- Real flow depth, chamber size and rim depth at the site (this decides whether early detection works).
- Whether failures are mostly blockages or pump-station stoppages.
- The real sensor's error: random noise averages out, but a **bias** does not.
- Echoes from chamber walls, ladder rungs and foam; corrosion from hydrogen sulphide.
- Radio choice, total energy budget, enclosure, and municipal permission.

THE LESSON

Good engineering says clearly what has been *proved* (the logic is correct and matches the reference), what has been *estimated* (site values), and what is *unknown* (real sensor behaviour). A confident wrong answer about a sewer is worse than an honest "we need to measure this first".

CHECK YOURSELF

1. G12 A manhole 1.2 m in diameter has 1.5 m of free space and an inflow of 0.5 L/s. How long until it overflows if fully blocked?
2. G12 Why did the team decide that detecting full blockage was not good enough?
3. G12 Find the speed of sound at 0 °C and at 35 °C. What distance error at 1.5 m would you make by ignoring the difference?
4. Varsity Show that the steady level rises by $1/(1-b)^2$ when $Q = K(1-b)\sqrt{h}$ and inflow is constant.
5. Varsity Why compare two independent implementations instead of just testing one?
6. Pro Why can't one sensor tell rain from a blockage? What extra information solves it?

WATCH & EXPLORE

- [HDLBits](https://hdlbits.01xz.net) — practise the Verilog skills used to write designs like Raven-01. Varsity
- [Tiny Tapeout](https://tinytapeout.com) — the low-cost route to turning a small design into silicon. Varsity
- [Siyavula Grade 10-12](https://siyavula.com) — waves and sound (speed of sound, echoes) in the CAPS curriculum. G12

IN PLAIN WORDS

Sewers in parts of Durban overflow and take days to fix. Naple designed a chip for a sensor that sits in a manhole and measures the water level with sound. A simple equation showed that once a pipe is fully blocked, it overflows within hours, which is too fast to help. So the chip learns what a normal day looks like and warns when the water stays a little too high. That catches a partly blocked pipe days early. Sound travels faster in warm air, so the chip also corrects for temperature. In computer tests it worked well. It has not yet been tested in a real manhole, and we say so openly.

SOURCES

1. Naple Semiconductor, RA-001: Raven-01 sewer blockage early-warning chip, internal research report, Draft 1, 2 Oct 2026 (public summary in this chapter).
2. Manhole chamber sizes 1 000–2 000 mm: SANS 1294 precast manhole components (as cited by Global Precast).
3. Speed of sound in dry air: ideal-gas relation $c = \sqrt{\gamma RT/M}$; any first-year physics text, e.g. Halliday, Resnick & Walker, *Fundamentals of Physics*.

4. Manning's equation for open-channel flow: V. T. Chow, *Open-Channel Hydraulics*, McGraw-Hill, 1959.

PART 2

The Economics of Semiconductors

Physics decides what is possible. Economics decides what gets built. Cost per chip, yield, factories that cost more than dams, design costs, business models and the chip cycle.

- 12 Why Chips Are an Economics Story
- 13 The Cost of a Chip: Wafers, Dies and Yield
- 14 Fabs, Design Costs and Business Models



Why Chips Are an Economics Story

GRADE 10

GRADE 11

GRADE 12

Physics decides what is possible. Economics decides what actually gets built. A chip that costs billions to develop can sell for less than a loaf of bread, but only if enough of them are made. This chapter explains why.

IN THIS CHAPTER YOU WILL LEARN

- fixed costs and variable costs
- why average cost falls with volume (economies of scale)
- learning curves and Wright's law
- supply, demand and why chip prices swing
- how big the chip business is (2025–2026)
- the questions every chip idea must answer

Fixed and variable costs G10

Think of a bakery. Before it sells a single loaf, it must pay for an oven, the shop and the licences: these are **fixed costs**, the same whether it bakes 10 loaves or 10 000. Each loaf also needs flour, yeast and electricity: these are **variable costs**, paid per unit.

Chips are an extreme version of the bakery. Designing a modern chip and making its masks can cost tens or hundreds of millions of US dollars (fixed). Once that is paid, each extra chip may cost only a few dollars (variable). Write F for fixed cost, c for variable cost per unit and Q for the number of units:

$$\text{Average cost per unit} = \frac{F + cQ}{Q} = \frac{F}{Q} + c \quad (12.1)$$

WORKED EXAMPLE A US\$ 50 million chip

A chip costs US\$ 50 million to develop (F) and US\$ 2 to make each (c).

Sell 1 million: average cost = $50\,000\,000/1\,000\,000 + 2 = \text{US\$ } 52$ per chip.

Sell 100 million: average cost = $0.50 + 2 = \text{US\$ } 2.50$ per chip.

Lesson. The same chip is 20 times cheaper per unit at 100 times the volume. Chip businesses live or die on volume. A brilliant chip that sells only a few thousand units can never pay back a large fixed cost.

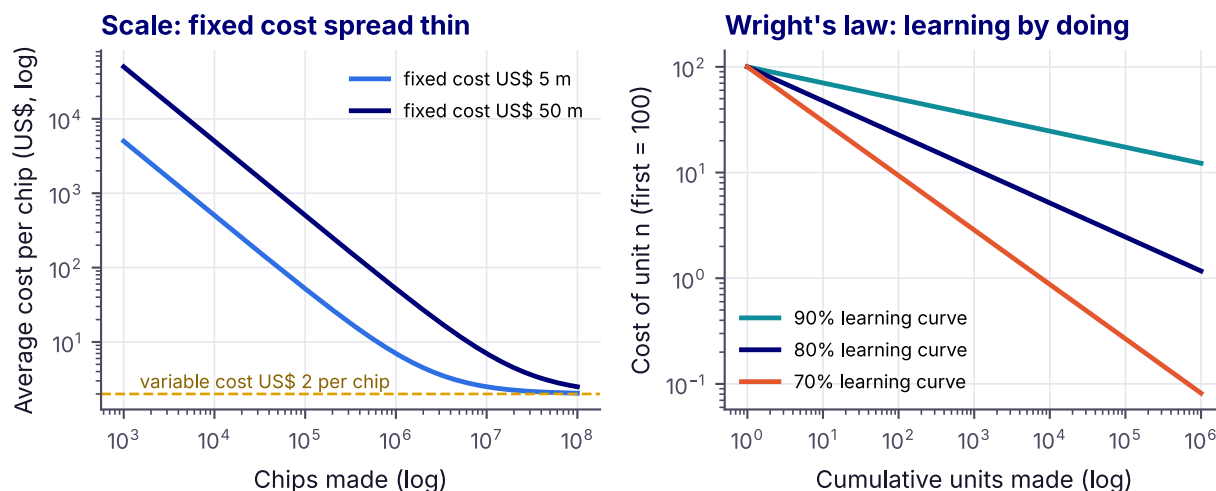


Figure 12.1 Left: Eq. 12.1 for two fixed costs. Average cost approaches the variable cost (dashed) only at very high volume. Right: learning curves. With an 80 % learning rate, each doubling of total output cuts unit cost to 80 % of what it was (model).

Learning by doing: Wright's law G11

In 1936 Theodore Wright noticed that each time aircraft factories doubled the total number of planes they had built, the labour per plane fell by a steady percentage. Chips behave the same way: fabs tune their processes, defects fall and yields rise. If the cost of the first unit is C_1 and the learning rate is r (for example 0.8):

$$C_n = C_1 n^{\log_2 r} \quad (12.2)$$

With $r = 0.8$, the 1 024th unit (10 doublings) costs $0.8^{10} = 0.107$ of the first. Memory chips and solar panels are famous examples of steep learning curves. Learning is also why **being first** to high volume is so valuable: the leader moves down the curve first and can cut prices that followers cannot match.

Supply, demand and price swings G10 G11

In Grade 10 Economics you learn that price settles where supply meets demand. Chips add a twist: **supply takes years to change** (a new fab takes years to build), while **demand can change in months**. When demand jumps, prices soar and everyone orders more. When demand falls, factories built for the boom sit half-empty. This makes the industry famously cyclical:

- **2020–2022:** a global shortage. Car makers in many countries, South Africa included, cut production because they could not get chips.
- **2023:** a downturn. Global sales fell 8.2 % to US\$ 526.8 billion as customers worked through stockpiles (SIA).
- **2025–2026:** an AI-driven boom. Sales rose 25.6 % to US\$ 791.7 billion in 2025 (SIA). In 2026 memory prices surged as AI data centres absorbed supply. WSTS's August 2026 update expects the 2026 market to roughly double, to about US\$ 1.66 trillion.

ABOUT THESE NUMBERS

2023 and 2025 sales are SIA annual totals based on WSTS data **PUBLISHED**. The 2026 figure is a forecast, not a result **FORECAST**. Most of the 2026 jump comes from memory prices, so it reflects price as well as volume. Forecasts in boom years are often revised, sometimes sharply. Chapter 15 has the full market picture.

Five questions every chip idea must answer **G12**

Table 12.1 The economic screen for any semiconductor idea (developed in Chapters 13–14)

#	Question	Why it matters
1	How many units will really sell, and at what price?	Volume spreads the fixed cost (Eq. 12.1)
2	What is the cost per <i>good</i> chip?	Defects mean not every chip on a wafer works (Chapter 13)
3	What are the one-off design and mask costs (NRE)?	These dominate at low volume (Chapter 14)
4	What is the break-even volume?	If it exceeds the whole market, the idea is dead, however good the physics
5	Is there a cheaper way: an existing chip, an FPGA, a module?	Usually yes. A custom chip must beat the alternatives on cost, power or size.

SOUTH AFRICAN CONTEXT

A small South African company cannot build a fab. A modern leading-edge fab costs tens of billions of US dollars: TSMC's planned US investment of US\$ 165 billion covers six fabs, two packaging plants and an R&D centre (announced 2025). But Eq. 12.1 also shows the opportunity. Design work has a much smaller fixed cost when done on mature processes through shared wafer runs. Selling design skill, rather than factory output, is where small players start.

CHECK YOURSELF

- G10** Classify as fixed or variable: mask set, silicon wafers, test time per chip, design engineers' salaries, packaging per chip.
- G10** A chip has $F = \text{US\$ } 10 \text{ million}$ and $c = \text{US\$ } 1$. Find the average cost at 100 000 and at 10 million units.
- G11** With a 90 % learning rate, what fraction of the first unit's cost does unit 256 cost?
- G11** Why do chip prices swing more than bread prices?
- G12** Using SIA figures, by what percentage did sales grow from 2023 to 2025?
- G12** Explain question 4 in Table 12.1 with your own example.

WATCH & EXPLORE

- SIA: Global annual semiconductor sales 2025** — the official yearly sales press release. semiconductors.org **G11**
- WSTS: market update, Q2 2026** — the industry's own statistics body. wsts.org **G12**
- Asianometry** — many videos on chip-industry economics, the memory cycle and fabs. youtube.com/@Asianometry **G12**

IN PLAIN WORDS

Making chips costs a huge amount up front: design, machines and masks. After that, each extra chip is cheap. So the more chips you sell, the cheaper each one becomes. Factories also get better with practice, which cuts costs further. But factories take years to build, while demand changes quickly, so the industry swings between shortages and gluts. Before building any chip, you must ask: will enough sell to pay back the start-up cost?

SOURCES

1. SIA, "Global Annual Semiconductor Sales Increase 25.6% to \$791.7 Billion in 2025", 6 Feb 2026; SIA, "Global Semiconductor Sales Decrease 8.2% in 2023", Feb 2024 (semiconductors.org).
2. WSTS, "Global Semiconductor Market records exceptional growth in Q2 2026", August 2026 (wsts.org).
3. T. P. Wright, "Factors affecting the cost of airplanes", *J. Aeronautical Sciences* 3, 122 (1936).
4. TSMC US investment: Arizona Technology Council, "TSMC to scale up to a gigafab cluster in Arizona", 28 Apr 2025.

The Cost of a Chip: Wafers, Dies and Yield

GRADE 12

VARSITY

A wafer is bought as a whole, but sold as many small chips, and not all of them work. Three numbers decide what each good chip costs: the wafer's cost, how many dies fit on it, and how many of those survive defects.

IN THIS CHAPTER YOU WILL LEARN

- the cost identity for a good die
- how to count dies on a round wafer
- defect density and yield models
- why big chips are disproportionately expensive
- why chiplets can be cheaper
- test, packaging and binning

The cost identity G12

Every cost estimate for a chip starts here:

$$\text{cost per good die} = \frac{\text{wafer cost}}{\text{good dies per wafer}}, \quad \text{good dies} = \text{gross dies} \times Y \quad (13.1)$$

Y is the **yield**, the fraction of dies that work. Then add the cost of testing and packaging each good die.

How many dies fit on a wafer?

A 300 mm wafer has area $\pi(150 \text{ mm})^2 = 70\,686 \text{ mm}^2$. Square dies cannot tile a circle, so some are lost around the edge. A good approximation for the number of whole dies of area A on a wafer of diameter d is

$$N_{\text{gross}} \approx \frac{\pi(d/2)^2}{A} - \frac{\pi d}{\sqrt{2A}} \quad (13.2)$$

The second term estimates the dies cut by the edge. Real counts also lose a few millimetres of edge exclusion and scribe lines between dies.

Defects and yield G12 VARSITY

Despite cleanrooms, a few random defects land on every wafer. Measure them as a **defect density** D_0 (killer defects per cm^2). If defects fall randomly and independently, the chance that a die of area A has none is given by the Poisson distribution:

$$Y_{\text{Poisson}} = e^{-AD_0}, \quad Y_{\text{Murphy}} = \left(\frac{1 - e^{-AD_0}}{AD_0} \right)^2 \quad (13.3)$$

Real defects cluster, so engineers also use Murphy's model and the negative-binomial model $Y = (1 + AD_0/\alpha)^{-\alpha}$, where α is a clustering factor. All of them agree on the key point: **yield falls steeply with die area**.

300 mm wafer, 10 × 10 mm dies: 616 whole dies,
60 hit by a defect (red), 556 good

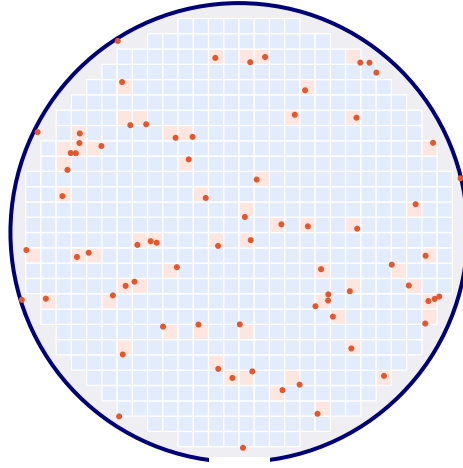


Figure 13.1 A simulated 300 mm wafer with 10 × 10 mm dies, a 3 mm edge exclusion and random defects at $D_0 = 0.1 \text{ cm}^{-2}$ (red dots). Dies hit by a defect (red) are scrapped. Count them yourself: about 90 % survive, as Eq. 13.3 predicts for $AD_0 = 0.1$.

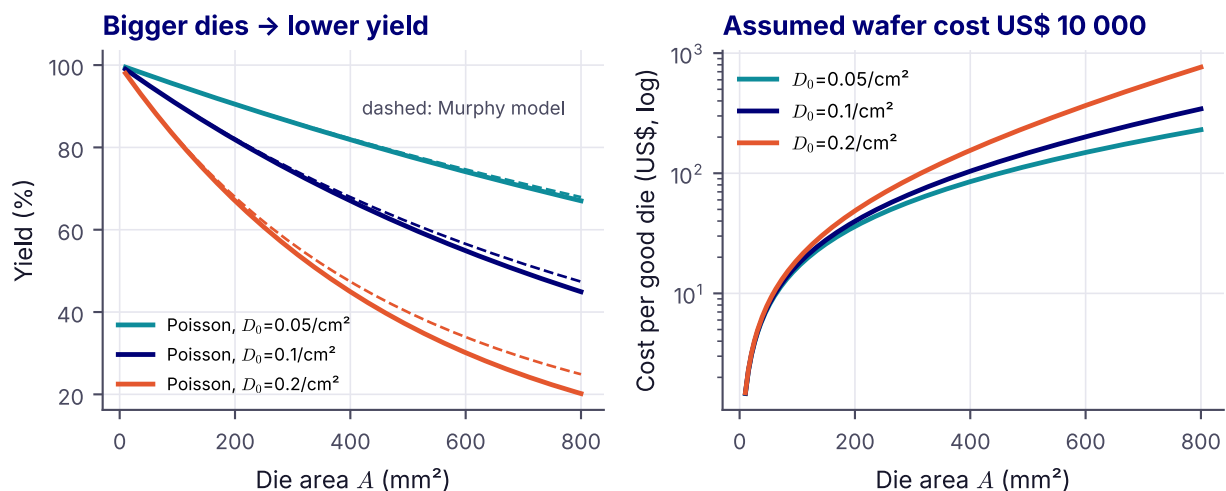


Figure 13.2 Left: yield against die area for three defect densities (solid: Poisson; dashed: Murphy). Right: cost per good die for an **assumed** wafer cost of US\$ 10 000, on a log scale. Doubling die area roughly doubles the cost from area alone, and lower yield makes it worse, so cost rises faster than area.

WORKED EXAMPLE Small die versus big die

Assume a wafer cost of US\$ 10 000 and $D_0 = 0.1 \text{ cm}^{-2}$.

Small die, 100 mm² (1 cm²): Eq. 13.2 gives $706.9 - 66.6 = 640$ dies. $Y = e^{-0.1} = 90.5\%$, so 579 are good. Cost per good die = $10\,000/579 = \text{US\$ } 17.27$.

Big die, 800 mm² (8 cm²): $88.4 - 23.6 = 64$ dies. $Y = e^{-0.8} = 44.9\%$, so 28.8 are good. Cost = US\$ 348.

Lesson. Eight times the area costs about 20 times as much per good die. That is why the largest AI processors are so expensive, and why designers fight for every square millimetre.

Chiplets: splitting a big die Varsity

Instead of one 800 mm² die, make four 200 mm² **chiplets** and join them in one package. Each 200 mm² die has $N_{\text{gross}} = 306$ and $Y = e^{-0.2} = 81.9\%$, so 250.5 are good per wafer. That is 62.6 complete sets of four, so the silicon cost per set is $10\,000/62.6 = \text{US\$ } 160$, compared with US\$ 348 for the single big die. The saving must pay for more complex packaging, die-to-die links and extra test. Often it does, which is why AMD, Intel, Nvidia and others now build their biggest products from several dies.

Test, packaging and binning Varsity

Every die is tested on the wafer (wafer sort), and again after packaging (final test). Test time on expensive machines is a real cost per chip. For low-cost chips, the package and test can cost more than the silicon itself. **Binning** sorts working chips by how fast they run or how much power they use. A chip that cannot reach the top clock speed, or has one faulty core switched off, is sold as a cheaper model. Binning recovers yield and lets one design serve several price points.

ABOUT WAFER PRICES

Foundries do not publish wafer prices, so all public figures are Analyst Estimates. Morgan Stanley estimated about US\$ 20 000 per wafer for TSMC's "3 nm" class in 2024 (reported January 2025). Analysts expect roughly US\$ 28 000–30 000 for "2 nm". Large customers are believed to pay less than list. Mature-node wafers (for example 130–180 nm, as used by open shuttles) cost far less. At US\$ 20 000 per wafer, our 100 mm² example costs about US\$ 35 per good die before test and packaging.

KEY IDEA: WHAT MATTERS MOST

The two inputs that swing cost most are **die area** and **defect density**, because they enter both the number of dies and, through an exponential, the yield. Shrinking a design by 20 % or maturing the process to halve D_0 can save more than any price negotiation.

CHECK YOURSELF

- G12 Use Eq. 13.2 to count 50 mm² dies on a 300 mm wafer.
- G12 Find the Poisson yield of a 50 mm² die at $D_0 = 0.2 \text{ cm}^{-2}$.
- G12 With a US\$ 3 000 mature-node wafer, find the cost per good 50 mm² die using your answers above.
- Varsity Show that for small AD_0 the Poisson and Murphy models agree. (Hint: expand the exponentials.)
- Varsity Repeat the chiplet example with two 400 mm² dies. Is it better or worse than four 200 mm² dies?
- Varsity Why might a company sell a chip with one disabled core instead of scrapping it?

WATCH & EXPLORE

- ▶ [Asianometry](https://www.youtube.com/@Asianometry) — videos on yield, chiplets and the economics of advanced packaging. [youtube.com/@Asianometry](https://www.youtube.com/@Asianometry) Varsity
- [Silicon Analysts: TSMC wafer price estimates](https://siliconanalysts.com/market-data/wafer-price-tsmc) — collected analyst estimates by node, with caveats. siliconanalysts.com/market-data/wafer-price-tsmc Pro

IN PLAIN WORDS

Chips are made many at a time on a round wafer. To find the cost of one chip, divide the wafer's cost by the number of chips that work. Tiny defects ruin some chips, and big chips are more likely to catch a defect. So big chips cost much more than their size suggests. That is why companies now split big chips into several smaller pieces joined together. Every good chip is then tested and packaged, which adds more cost.

SOURCES

1. Dies-per-wafer and yield models: J. L. Hennessy & D. A. Patterson, *Computer Architecture: A Quantitative Approach*, 6th ed., Morgan Kaufmann, 2017, ch. 1; B. T. Murphy, "Cost-size optima of monolithic integrated circuits", *Proc. IEEE* 52, 1537 (1964); C. H. Stapper, *IBM J. Res. Dev.* 27, 549 (1983).
2. Wafer price estimates: Silicon Analysts, "TSMC wafer price by node" (citing Morgan Stanley via Tom's Hardware, Jan 2025), accessed 15 Oct 2026.

Fabs, Design Costs and Business Models

VARSITY

PROFESSIONAL

Why do only a handful of companies make the most advanced chips? Why do most chip companies own no factory at all? And when does it pay to make your own chip instead of using an off-the-shelf part or an FPGA? This chapter answers all three with arithmetic.

IN THIS CHAPTER YOU WILL LEARN

- the scale of fab investment (2025–2026 figures)
- depreciation and why empty fabs lose money
- NRE: the one-off cost of a chip design
- break-even volume and the FPGA-versus-ASIC decision
- shuttles: the low-cost entry route
- IDM, fabless, foundry and the rest of the value chain

The most expensive factories on Earth VARSITY

A leading-edge fab needs hundreds of lithography, etch, deposition and inspection tools, some costing hundreds of millions of US dollars each (Chapter 10), housed in vibration-free cleanrooms with their own water treatment and power supply. The numbers are national-budget scale:

Table 14.1 Fab spending, for scale

Item	Figure	Date	Label
TSMC capital expenditure, full year 2025	US\$ 40.9 billion	Jan 2026	COMPANY REPORT
TSMC 2026 capex guidance	US\$ 52–56 billion	Jan 2026	COMPANY GUIDANCE
TSMC planned US investment: six fabs, two packaging plants, one R&D centre	US\$ 165 billion	2025	COMPANY PLAN
One High-NA EUV scanner	about US\$ 380 million	May 2026	PRESS ESTIMATE

FOR SCALE

At roughly R16–17 to the US dollar (the 2026 range), TSMC's 2025 capital spending alone was about R660–700 billion: a large fraction of South Africa's entire national budget, spent by one company in one year on equipment and buildings.

Depreciation and utilisation

Equipment is **depreciated**: its cost is spread over its useful life (for fab tools, typically about five years) as an expense each year, whether or not the tools are busy. So a fab has a huge fixed cost per year. Spread it over the wafers actually produced and the cost per wafer depends strongly on **utilisation**, the fraction of capacity in use:

$$\text{cost per wafer} = \frac{\text{fixed cost per year}}{u \times \text{capacity}} + \text{variable cost per wafer} \quad (14.1)$$

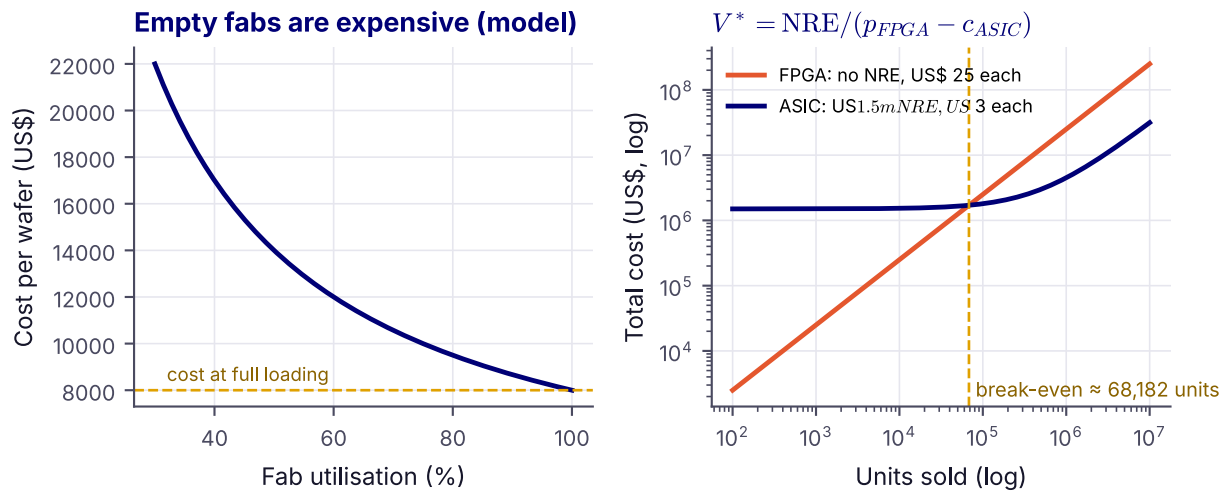


Figure 14.1 Left: Eq. 14.1 with an illustrative US\$ 6 000 fixed and US\$ 2 000 variable cost per wafer at full loading. At 50 % utilisation the cost per wafer rises from US\$ 8 000 to US\$ 14 000. Right: total cost of an FPGA-based product against a custom ASIC (illustrative numbers in the legend). The lines cross at the break-even volume.

This is why foundries fight to keep fabs full, why downturns hurt them so badly, and why they ask big customers for long-term commitments or prepayments before building new capacity.

Design economics: NRE and break-even VARSITY PRO

For a chip *designer*, the big fixed cost is **NRE** (non-recurring engineering): everything paid once, before the first chip is sold.

Table 14.2 What NRE is made of

Component	Notes
Engineering time	Architecture, RTL, verification (often the largest part), physical design, test development. Measured in engineer-years.
EDA tool licences	Commercial design tools are licensed per seat per year; open-source tools are free but cover mainly mature processes.
IP licences	Processor cores, interfaces, memories bought from others, often with royalties per chip sold.
Mask set	One mask per patterned layer. Cost rises steeply at advanced nodes with more layers and EUV masks.
Prototype wafers, packaging, test hardware	Plus re-spins if the first silicon has bugs.
Qualification	Reliability testing; for automotive parts this can take many months.

If a custom chip has NRE N and unit cost c , and the alternative (an existing chip or an FPGA) costs p per unit with no NRE, the custom chip wins above the **break-even volume**:

$$V^* = \frac{N}{p - c} \quad (14.2)$$

WORKED EXAMPLE FPGA or ASIC?

A product can use an FPGA costing US\$ 25, or an ASIC with NRE of US\$ 1.5 million and a unit cost of US\$ 3 (illustrative). Then $V^* = 1\,500\,000 / (25 - 3) = 68\,200$ units. Below that, the FPGA is cheaper overall. Above it, the ASIC wins, and by a lot at a million units: US\$ 4.5 million versus US\$ 25 million.

If the whole market is smaller than V^* , the ASIC is dead however elegant it is. Real decisions also weigh power, size, time to market and the risk of a failed first silicon.

Shuttles: the low-NRE entry route

A **multi-project wafer** (MPW, or shuttle) shares one mask set and wafer run among many designs. Each project gets a few dozen to a few hundred chips for a small share of the cost. On mature open processes this brings prototype costs down from millions to thousands or tens of thousands of US dollars. One shuttle service quoted US\$ 15 000 for 100 chips on SkyWater 130 nm (May 2025), and community programmes such as Tiny Tapeout sell small "tiles" for much less (Chapter 10). Shuttles do not deliver volume production, but they prove a design in real silicon cheaply.

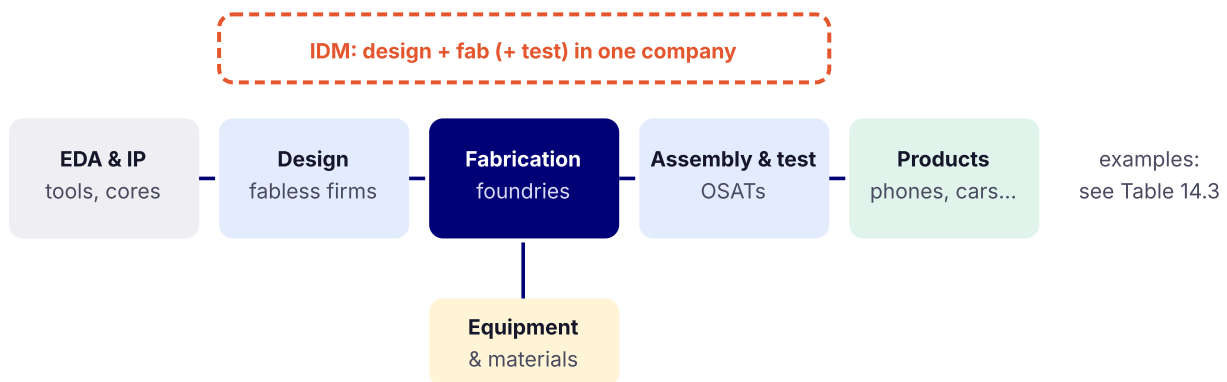
Who does what: business models VARSITY

Figure 14.2 The semiconductor value chain. Most chip companies specialise in one box; integrated device manufacturers span several. Equipment and materials suppliers feed the fabs.

Table 14.3 Business models with examples

Model	What they do	Examples
IDM	Design and manufacture their own chips	Intel, Samsung, Texas Instruments, Infineon, Micron
Fabless	Design only; outsource manufacturing	Nvidia, Qualcomm, AMD, Broadcom, MediaTek, Apple's chip unit
Pure-play foundry	Manufacture other companies' designs	TSMC, UMC, GlobalFoundries, SMIC
OSAT	Outsourced assembly (packaging) and test	ASE, Amkor, JCET
EDA	Design software	Synopsys, Cadence, Siemens EDA
IP licensing	Sell reusable designs (cores) for royalties	Arm
Equipment	Make the tools in fabs	ASML, Applied Materials, Lam Research, Tokyo Electron, KLA
Design services	Design chips for customers who lack the team	Many small and mid-size firms worldwide

The fabless–foundry split, pioneered when TSMC was founded in 1987, is why a start-up can design a chip without building a factory. Specialisation also concentrates profit: TSMC reported a 59.9 % gross margin on US\$ 122 billion of revenue in 2025, showing how valuable being the best manufacturer has become.

KEY IDEA

Capital decides who can manufacture. Volume decides who can design. Small companies enter where both are low: design services, IP, FPGA-based products, and chips on mature processes made through shuttles, growing towards volume only when customers commit.

CHECK YOURSELF

1. **Varsity** Using Eq. 14.1 with the Figure 14.1 numbers, find the cost per wafer at 70 % utilisation.
2. **Varsity** A design has NRE of US\$ 400 000 and a unit cost of US\$ 1.20; the off-the-shelf alternative costs US\$ 3.50. Find V^* .
3. **Varsity** Classify: Nvidia, TSMC, Intel, ASML, Arm, ASE.
4. **Pro** Why might a fabless company pay a foundry in advance for capacity?
5. **Pro** Give two non-cost reasons to choose an ASIC over an FPGA, and two to choose the FPGA.

WATCH & EXPLORE

- ▶ [Asianometry](https://www.youtube.com/@Asianometry) — the history of TSMC, the fabless model and fab economics. [youtube.com/@Asianometry](https://www.youtube.com/@Asianometry) **Varsity**
- [Tiny Tapeout](https://www.tinytapeout.com) — see what a shared-wafer tile costs today with its price calculator. [tinytapeout.com](https://www.tinytapeout.com) **Varsity**
- [Semiconductor Industry Association \(SIA\)](https://www.semiconductors.org) — free factbooks and reports on the industry. [semiconductors.org](https://www.semiconductors.org) **Pro**

IN PLAIN WORDS

Chip factories cost tens of billions of dollars, and they lose money when they are not busy. So only a few giant companies make the most advanced chips. Most chip companies only design chips and pay a factory, called a foundry, to make them. Designing a custom chip also has a big start-up cost. It only pays off if you sell enough chips; otherwise a ready-made chip or a reprogrammable FPGA is cheaper. Shared wafer runs let small teams make test chips cheaply.

SOURCES

1. TSMC Q4 2025 earnings conference call, 15 Jan 2026 (2025 revenue US\$ 122 billion, gross margin 59.9 %, capex US\$ 40.9 billion; 2026 capex guidance US\$ 52–56 billion), as summarised by VectorShift research.
2. Arizona Technology Council, "TSMC to scale up to a gigafab cluster in Arizona", 28 Apr 2025 (US\$ 165 billion).
3. A. Shilov, "ASML's roadmap for chipmaking lithography tools examined", Tom's Hardware, 1 May 2026 (EUV tool prices).
4. N. Flaherty, eeNews Europe, 16 May 2025 (ChipFoundry SKY130 shuttle, US\$ 15 000 for 100 chips).
5. Exchange-rate range R15.73–R17.19 per US\$ in 2026: ValutaFX mid-rates as compiled in Naple research note RA-005 (Oct 2026).

PART 3

The Semiconductor Market

How big the industry is, who leads it, where chips are made, the politics around them, the AI boom, and where South Africa and Africa fit, including how you can start a career.

- 15 The Global Chip Market and Who Leads It
- 16 Supply Chains, Geopolitics and the AI Boom
- 17 South Africa, Africa and Your Career in Chips



The Global Chip Market and Who Leads It

GRADE 11

GRADE 12

VARSITY

In 2025 the world bought about US\$ 792 billion of semiconductors. In 2026 that figure is on track to roughly double. This chapter sizes the market, breaks it into its parts, and introduces the companies at the top, with dated sources for every number.

IN THIS CHAPTER YOU WILL LEARN

- how big the market is, 2017–2026
- the main product categories and which are growing
- where chips are sold, by region
- the top 10 chip companies of 2025
- why one foundry makes most advanced chips
- how to read market data critically

How big is the market? G11



Figure 15.1 Global semiconductor sales, US\$ billion. 2021–2025 from SIA annual releases (WSTS data); 2017–2020 from earlier SIA releases (recalled, verify before citing); 2026 is the WSTS forecast published in August 2026. Notice the dip in 2019 and 2023 (cycles, Chapter 12) and the AI-driven surge in 2025–2026.

To make US\$ 791.7 billion feel real: it is roughly twice South Africa's entire annual economic output (GDP) at 2025 exchange rates. Yet a single phone chip may cost only tens of dollars. The market is huge because of volume: hundreds of billions of chips a year.

ABOUT THESE NUMBERS

Three bodies publish headline figures, and they differ slightly because they count differently. **SIA** reported 2025 sales of US\$ 791.7 billion (+25.6 %). **WSTS** gave US\$ 795.6 billion (+26.2 %). **Gartner** estimated vendor revenue of US\$ 793.4 billion (+21.0 %, preliminary). The 2026 forecast changed fast: WSTS forecast US\$ 1 511 billion in June 2026, then raised it to US\$ 1 655 billion (+108 %) in August after first-half sales of US\$ 702 billion. Treat 2026 as a forecast FORECAST, not a fact.

What is sold: product categories G11 G12

Table 15.1 Product categories used by WSTS

Category	What it includes	Example
Logic	Processors for AI and graphics, custom chips, programmable logic	GPUs, AI accelerators, FPGAs
Memory	DRAM (including HBM) and NAND flash	Phone storage, server memory
Micro	Microprocessors, microcontrollers, DSPs	PC CPUs; the MCU in a prepaid meter
Analog	Power management, amplifiers, data converters, interfaces	Phone charging chips
Discrete	Individual diodes and transistors, including power devices	Inverter power MOSFETs
Optoelectronics	LEDs, laser diodes, image sensors, optocouplers	Camera sensors, LED lights
Sensors	Pressure, motion, magnetic and other sensors	Phone accelerometer

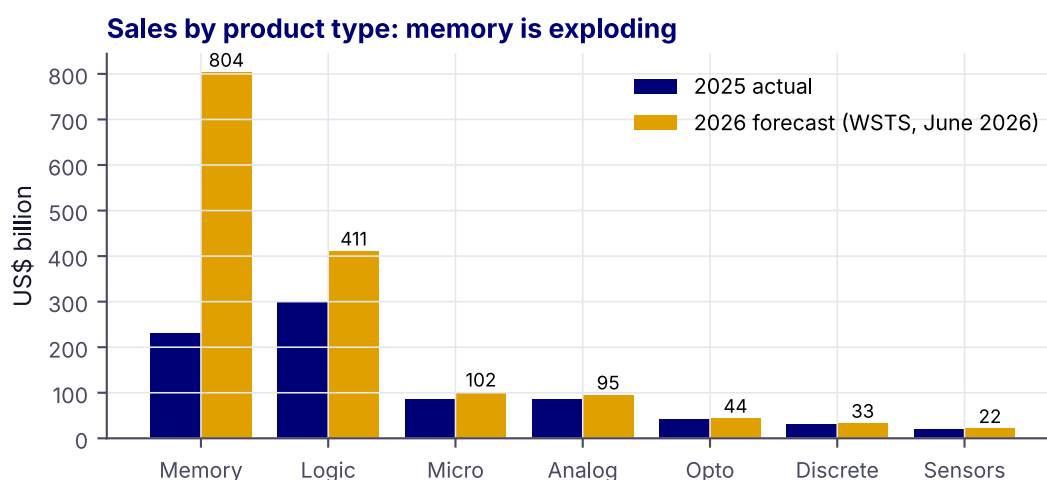


Figure 15.2 WSTS Spring 2026 forecast (published 2 June 2026), US\$ billion. Logic and memory dominate. The forecast memory jump (+250 %, raised to about +300 % in August 2026) comes mainly from soaring DRAM and HBM prices driven by AI data centres. Analog, discrete, opto and sensor chips grow at a normal pace.

Where chips are sold G12

Table 15.2 Sales by region (WSTS, US\$ billion; region where the chips are delivered, not where they are designed)

Region	2025 actual	2025 growth	2026 forecast (June)
Asia Pacific (incl. China, Taiwan, Korea)	439.7	+30.3 %	823.9
Americas	256.5	+31.4 %	543.7
Europe	54.7	+6.7 %	86.6
Japan	44.7	-4.3 %	57.1

Most chips are sold into Asia because that is where phones, PCs and servers are assembled, even when the end customer is elsewhere. Africa is not shown separately: its direct chip purchases are tiny, because most chips arrive inside finished products (Chapter 17).

Who leads: the companies G11 VARSITY

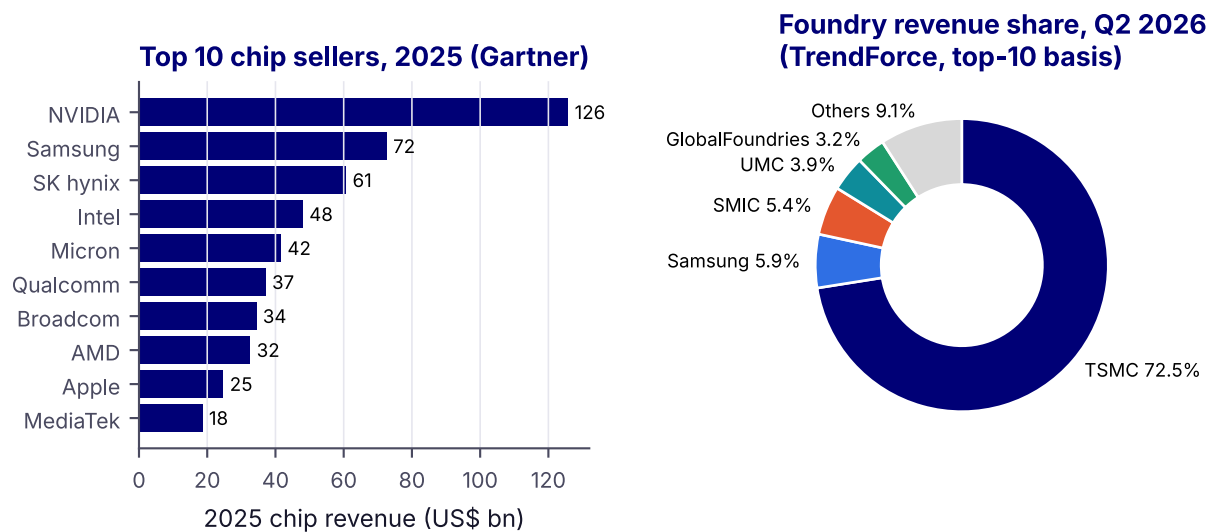


Figure 15.3 Left: top 10 chip sellers by 2025 revenue (Gartner, preliminary, January 2026). Right: foundry revenue shares in Q2 2026 among the top 10 foundries (TrendForce, September 2026).

Three stories stand out:

- **Nvidia** became the first company to sell more than US\$ 100 billion of chips in a year (US\$ 125.7 billion in 2025, 15.8 % of the market), almost entirely AI and data-centre processors. It is **fabless**: TSMC manufactures its chips.
- **Memory makers surged.** SK hynix rose to number 3 on high-bandwidth memory (HBM) for AI. HBM passed US\$ 30 billion in 2025, 23 % of the DRAM market (Gartner).
- **Intel**, the long-time leader, fell to 6 % market share, half its 2021 level (Gartner), while investing heavily to rebuild its manufacturing.

In manufacturing, **TSMC** took 72.5 % of foundry revenue in Q2 2026 (about US\$ 40.2 billion in the quarter), with its most advanced capacity fully booked and its "2 nm" process earning revenue for the first time. Samsung (5.9 %) and China's SMIC (5.4 %) followed. Gartner estimates that AI chips (processors, HBM and networking) made up nearly one-third of all 2025 chip sales, and expects them to exceed half by 2029.

KEY IDEA

The market is concentrated at each step: one company dominates advanced manufacturing (TSMC), one dominates AI processors (Nvidia), three make nearly all DRAM (Samsung, SK hynix, Micron), and one makes all EUV scanners (ASML). Concentration brings efficiency, but it also brings risk, which Chapter 16 explores.

READING MARKET DATA CRITICALLY

Always ask: **who** measured it (association, analyst, company)? **What** was counted (vendor revenue, or sales by delivery region)? **When** (actual or forecast; which revision)? And is growth from more **units** or higher **prices**? The 2026 surge is driven mainly by memory prices, which can fall as fast as they rose.

CHECK YOURSELF

1. **G11** By what percentage did global chip sales grow from 2024 (US\$ 630.5 billion) to 2025 (US\$ 791.7 billion)?
2. **G11** Which two product categories are largest?
3. **G12** What share of 2025 sales did Nvidia, Samsung and SK hynix have together?
4. **G12** Why is Asia Pacific the largest region even though many chip companies are American?
5. **Varsity** Why do SIA, WSTS and Gartner give slightly different 2025 totals?
6. **Varsity** Give one reason the 2026 forecast could turn out too high.

WATCH & EXPLORE

- [WSTS market releases](https://wsts.org) — official forecasts and quarterly updates. wsts.org **G12**
- [Gartner: worldwide semiconductor revenue 2025](https://gartner.com/en/newsroom) — the top-10 vendor ranking. gartner.com/en/newsroom **Varsity**
- [TrendForce: foundry ranking 2Q26](https://trendforce.com/presscenter) — quarterly foundry shares. trendforce.com/presscenter **Varsity**
- ▶ [Asianometry](https://youtube.com/@Asianometry) — company histories of TSMC, Nvidia, Samsung, SK hynix and others. youtube.com/@Asianometry **G12**

IN PLAIN WORDS

The world spent about 790 billion US dollars on chips in 2025. In 2026 it may spend about twice that, mostly because AI data centres need huge amounts of memory and processing. The biggest seller is Nvidia, which designs AI chips but does not make them. TSMC in Taiwan makes most of the world's advanced chips. A few companies lead each part of the industry. Numbers from different sources differ a little, and forecasts change often, so always check who made a number and when.

SOURCES

1. SIA, "Global Annual Semiconductor Sales Increase 25.6% to \$791.7 Billion in 2025", 6 Feb 2026; SIA 2022 and 2023 annual releases; earlier years from prior SIA annual releases.
2. WSTS Spring 2026 forecast press release, 2 Jun 2026 (wsts.org); WSTS, "Global Semiconductor Market records exceptional growth in Q2 2026", Aug 2026.
3. Gartner, "Gartner Says Worldwide Semiconductor Revenue Grew 21% in 2025", 12 Jan 2026 (preliminary).
4. TrendForce, foundry revenue ranking 2Q26, 9 Sep 2026.
5. South African GDP comparison: World Bank / Stats SA GDP of roughly US\$ 400 billion (order of magnitude; verify current figure).

Supply Chains, Geopolitics and the AI Boom

GRADE 12

VARSITY

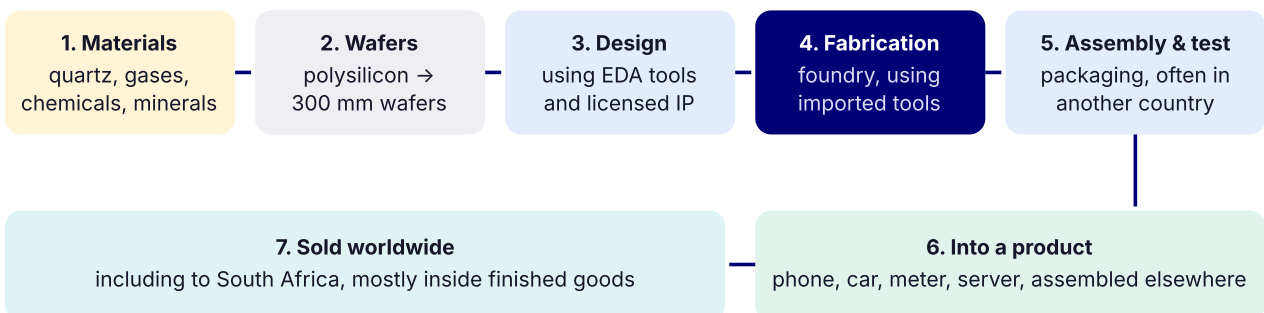
PROFESSIONAL

A single chip may cross borders a dozen times before it reaches your pocket. That makes chips efficient to produce and fragile to disrupt, and it has turned them into a matter of national strategy. Meanwhile, artificial intelligence is reshaping what the industry builds.

IN THIS CHAPTER YOU WILL LEARN

- the journey of a chip around the world
- chokepoints and single points of failure
- export controls, subsidies and tariffs (dated)
- why AI needs so many chips, and so much power
- the main arguments in the "AI bubble" debate
- how to keep geopolitical facts up to date

The journey of a chip G12



Typical pattern (varies by product): design in the USA, Europe or Asia; tools from the Netherlands, USA and Japan; fabrication in Taiwan, Korea, China or the USA; assembly and test in Southeast Asia or China.

Figure 16.1 A stylised journey. Each step is done where it is cheapest or where the expertise is, which is efficient but means a problem in one place can stop production everywhere.

Chokepoints Varsity

Table 16.1 Single points of failure in the supply chain (as of 2025–2026)

Chokepoint	Why it matters	Evidence
EUV lithography	Only one company (ASML, Netherlands) makes EUV scanners, needed for every leading-edge chip	Chapter 10
Leading-edge foundry	One company produces most advanced logic, mostly in Taiwan	TSMC 72.5 % of foundry revenue, Q2 2026 (TrendForce)
AI memory (HBM)	Three suppliers; capacity sold out during the AI boom	Gartner, Jan 2026
Advanced packaging	Packaging that joins AI processors to HBM has been a capacity bottleneck	Industry reports, 2024–2026
Critical minerals	Gallium and germanium (used in compound semiconductors and optics) are refined mostly in China, which has restricted exports since 2023	Policy announcements; verify current status
Mature-node supply	Cars, meters and appliances depend on older, cheaper chips; the 2020–2022 shortage hit these hardest	Chapter 12

Chips and national policy Varsity Pro

Because chips are essential to economies and militaries, governments now treat them as strategic. Four kinds of policy matter:

- **Subsidies** to attract fabs: the US CHIPS and Science Act (2022, about US\$ 52.7 billion for semiconductors); the EU Chips Act (2023, aiming to mobilise over €43 billion); large programmes in Japan, South Korea, China and India.
- **Export controls**: since October 2022 the United States has restricted exports of advanced AI chips and chipmaking tools to China, with several later updates. The Netherlands and Japan restrict some tool exports. China has restricted exports of some critical minerals.
- **Tariffs**: in January 2026 the US imposed a 25 % Section 232 tariff on a narrow category of advanced AI processors, with broad exemptions, and said wider semiconductor tariffs might follow depending on negotiations.
- **Localisation**: many countries want more of the supply chain inside their borders or among allies.

THE MAIN ARGUMENTS, PRESENTED EVENLY

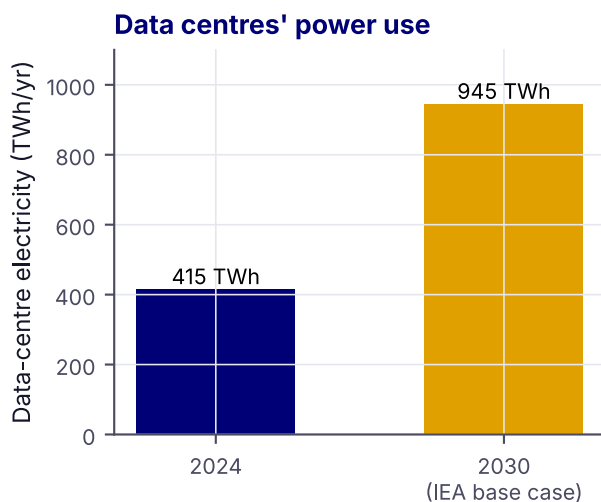
Supporters of subsidies and controls argue that concentration in a few places is a security and resilience risk, that leading-edge chips have military uses, and that the market alone underprices resilience. **Critics** argue that duplicating fabs raises costs for everyone, that controls can speed up rivals' self-sufficiency, that tariffs raise prices for consumers and manufacturers, and that subsidy races waste public money. Economists also disagree on how much resilience is worth paying for. This book takes no side. Look at the evidence as it develops.

DATES MATTER: VERIFY BEFORE RELYING

Policy changes faster than any book. Everything in this section is as understood in **October 2026**. Before using any of it for a decision, check the latest official sources: government gazettes, the US Bureau of Industry and Security, the European Commission and trade ministries. Older figures in this section (CHIPS Act, EU Chips Act, mineral controls) are widely reported but were not re-checked for this edition.

The AI boom Varsity Pro

Training and running large AI models is mostly **matrix multiplication**: billions of multiply-and-add operations that can run in parallel. GPUs and custom AI accelerators contain thousands of simple arithmetic units working at once. They are limited less by arithmetic than by how fast data reaches them, which is why AI systems use **HBM**: stacks of DRAM dies placed right next to the processor inside one advanced package.

**The numbers (dated):**

- AI processors passed US\$ 200 billion in 2025; AI chips overall were nearly a third of all chip sales (Gartner, Jan 2026).
- Gartner expected AI infrastructure spending to exceed US\$ 1.3 trillion in 2026.
- TSMC guided 2026 capital spending to US\$ 52–56 billion, up from US\$ 40.9 billion (Jan 2026).
- For scale, 945 TWh a year is roughly four to five times South Africa's annual electricity consumption.

Figure 16.2 Global data-centre electricity: about 415 TWh in 2024 (about 1.5 % of world electricity) and about 945 TWh projected for 2030 in the IEA base case, driven mainly by AI.

IS IT A BUBBLE? BOTH SIDES

The bull case: AI is a general-purpose technology. Demand for training and for running models (inference) keeps rising, and early revenue from AI services is growing, so heavy investment now builds lasting infrastructure. **The bear case:** spending is running far ahead of proven profits, much demand comes from a handful of very large buyers, memory and processor prices could fall sharply when new capacity arrives (as in past cycles, Chapter 12), and power and grid limits may slow deployment. History shows that genuine technologies can still have investment busts. Watch revenue, not just spending.

SOUTH AFRICAN CONTEXT

Data centres are being built in South Africa too, and they compete for the same constrained electricity grid. Power-electronics design, energy efficiency and renewable integration are areas where local engineering skills connect directly to the AI-chip story.

CHECK YOURSELF

1. **G12** List the seven steps in Figure 16.1 and give one country often involved in each.
2. **G12** What is a "chokepoint"? Give two examples from Table 16.1.
3. **VARSIITY** Why do AI chips need HBM?
4. **VARSIITY** By what factor does the IEA expect data-centre electricity to grow from 2024 to 2030?
5. **PRO** Write the strongest argument for and against export controls on chipmaking tools, in two sentences each.
6. **PRO** Name two indicators you would track to judge whether AI chip demand is sustainable.

WATCH & EXPLORE

- ▶ [The Ridiculous Engineering of the World's Most Important Machine](https://www.youtube.com/watch?v=MtUHjLxm3V0) — Veritasium: why EUV is a chokepoint. [youtube.com/watch?v=MtUHjLxm3V0](https://www.youtube.com/watch?v=MtUHjLxm3V0) **G12**
- [IEA: Energy and AI](https://www.iea.org/reports/energy-and-ai) — data-centre electricity, today and to 2030. [iea.org/reports/energy-and-ai](https://www.iea.org/reports/energy-and-ai) **VARSIITY**
- [White & Case: Section 232 tariff on advanced chips \(Jan 2026\)](https://www.whitecase.com) — a clear legal summary. [whitecase.com](https://www.whitecase.com) **PRO**
- ▶ [Asianometry](https://www.youtube.com/@Asianometry) — supply-chain and geopolitics explainers. [youtube.com/@Asianometry](https://www.youtube.com/@Asianometry) **VARSIITY**

IN PLAIN WORDS

A chip is made in steps across many countries: materials, wafers, design, factory, packaging. That keeps costs low, but if one step breaks, everything stops. Some steps rely on just one company or one country. So governments now pay to build chip factories at home, and some block sales of advanced chips and machines to rivals. People disagree about whether this is wise. AI is driving huge demand for chips and electricity. It may keep growing, or it may cool down. Rules change fast, so always check the latest news.

SOURCES

1. TrendForce, foundry ranking 2Q26, 9 Sep 2026; Gartner, worldwide semiconductor revenue 2025, 12 Jan 2026.
2. IEA, *Energy and AI*, executive summary, 2025 ([iea.org/reports/energy-and-ai](https://www.iea.org/reports/energy-and-ai)).
3. White & Case, "President Trump orders narrowly targeted 25% Section 232 tariff on certain advanced semiconductor articles", Jan 2026.
4. CHIPS and Science Act (US Public Law 117-167, Aug 2022); European Chips Act (Regulation (EU) 2023/1781); US BIS export-control rules of 7 Oct 2022 and later updates. Recalled; verify current status.
5. TSMC Q4 2025 earnings call, 15 Jan 2026.
6. South African electricity consumption of roughly 200–230 TWh a year: Eskom and Stats SA (order of magnitude; verify current figure).

South Africa, Africa and Your Career in Chips

ALL LEVELS

South Africa does not make advanced chips, but it buys billions of rands' worth inside phones, computers, cars and solar systems, and it has a real history of chip design. This chapter maps where Africa stands honestly, where the openings are, and how you can start a career, from Grade 10 onwards.

IN THIS CHAPTER YOU WILL LEARN

- what South Africa imports, and why most chips arrive "hidden"
- local and African semiconductor activity
- realistic opportunities for African companies
- the jobs in the chip industry
- subjects and degrees that lead there
- a 12-month self-study plan and a project ladder

What South Africa buys G10 G12

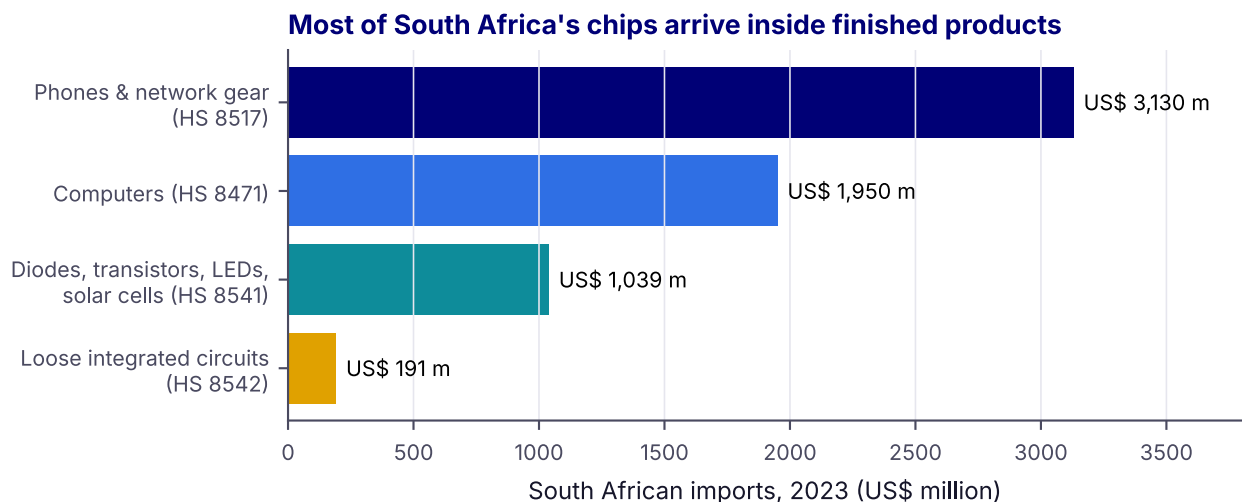


Figure 17.1 South African imports in 2023 (UN Comtrade data via TrendEconomy, compiled in Naple research note RA-005, Oct 2026). Loose integrated circuits were only about US\$ 191 million (roughly R3 billion), about 1.5 % of electrical-equipment imports. Phones, network gear and computers carried far more chip value inside them. The HS 8541 line is mostly solar cells and LEDs, about 87 % from China.

So most semiconductors enter South Africa **inside finished products**. The local market for loose chips is small, and the most-used chip types (memory, phone processors, general microcontrollers) are dominated worldwide by a handful of giant firms. That is the honest starting point for any African chip strategy.

Local and African activity G10 VARSIITY

Table 17.1 Examples of semiconductor-related activity (public information; verify current status)

Who / what	Where	What they do
Azoteq	Paarl, Western Cape (founded 1998)	Fabless design of capacitive and inductive sensing chips, sold worldwide
SAMES (historical)	South Africa	Designed energy-metering ICs used in electricity meters
Prepaid metering industry (e.g. Conlog)	Durban and elsewhere	Meter design and manufacture. The STS prepaid-token standard, now IEC 62055, began in South Africa and is reported in around 70 million meters in more than 40 countries (2024).
Universities	e.g. University of Pretoria (Carl and Emily Fuchs Institute for Microelectronics), Stellenbosch, UCT, Wits, UKZN	Microelectronics, RF, power-electronics and photonics research and teaching
Solar and inverter industry	Countrywide	System design, installation and some local electronics design
Morocco	Bouskoura	STMicroelectronics assembly and test operations
Egypt	Cairo	Chip-design centres of international firms and local design houses
Naple Semiconductor	Durban, KwaZulu-Natal	Digital chip design, FPGA prototyping and research (Chapter 11)

Where the openings are G12 PRO

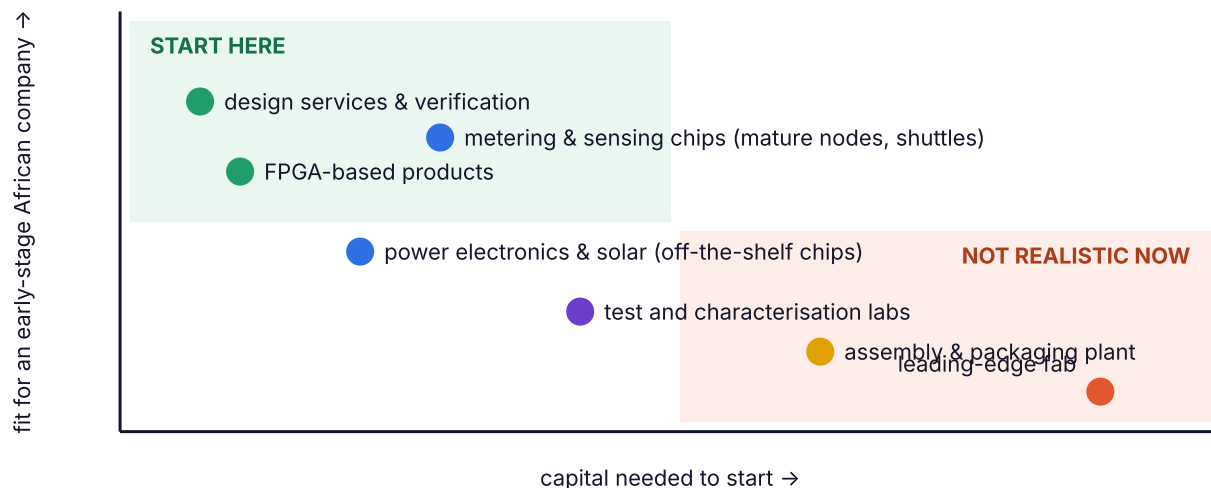


Figure 17.2 A qualitative opportunity map (positions are judgement, not data). Design, verification, FPGA products and chips on mature processes need skills more than capital, and match real local demand in metering, power, solar and sensing. Fabs need tens of billions of US dollars (Chapter 14).

KEY IDEA

Africa's best entry to the chip industry is **skills, not factories**. Design engineers, verification engineers, power-electronics engineers and test engineers can work from anywhere, for customers anywhere, and they are in short supply worldwide.

Jobs in the chip industry ALL

Table 17.2 Roles, and what each needs

Role	What you do	Core skills
Device / process engineer	Develop and improve transistors and fab steps	Physics, materials, chemistry, statistics (Chapters 4–10)
Digital design engineer	Write RTL for processors and accelerators	Logic, Verilog/VHDL, computer architecture
Verification engineer	Prove designs correct before tape-out	SystemVerilog, testbenches, Python; methodical thinking
Analogue / mixed-signal designer	Design amplifiers, converters, power and radio circuits	Circuit theory, device physics, SPICE
Physical design / layout engineer	Turn circuits into manufacturable geometry	EDA tools, timing, design rules
Test & product engineer	Test chips, find failures, improve yield	Instruments, statistics, scripting
Applications / field engineer	Help customers use chips in products	Electronics, communication
Industry analyst / strategist	Study markets, costs and policy	Economics, data analysis (Chapters 12–16)

From school to a career G10 G11



Figure 17.3 A typical pathway. Check each university's current admission requirements: engineering programmes usually require Mathematics and Physical Sciences with high marks.

A 12-month self-study plan

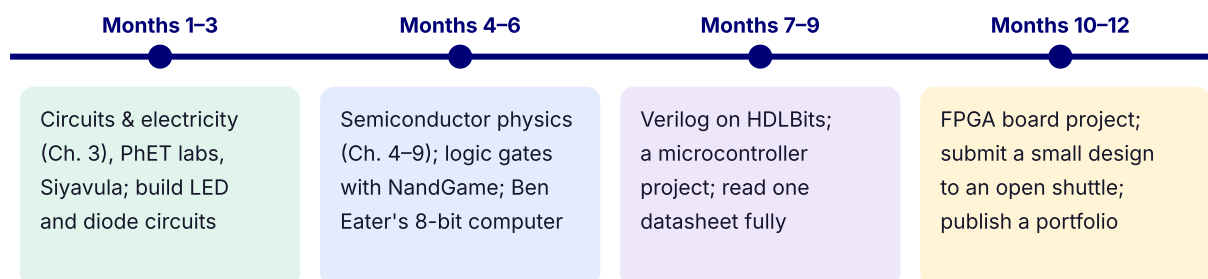


Figure 17.4 One year, about 5–8 hours a week. Each stage ends with something you built and can show. Adjust the pace to your level; Grade 10 learners may spend the whole year on the first two stages.

The project ladder

1. **LED and resistor** on a breadboard: Ohm's law and the diode knee (Chapters 3, 7).

2. **Logic chips:** build a half adder from 74-series gates (Chapter 8).
3. **Microcontroller** (for example an Arduino-class board): read a temperature sensor and correct a measurement, as in Chapter 11.
4. **FPGA:** put your Verilog counter (Chapter 10) on a low-cost board and drive LEDs.
5. **Silicon:** submit a tiny design to Tiny Tapeout or a similar open shuttle.

HOW TO READ A DATASHEET

Every chip comes with a datasheet. Read in this order: (1) the **features** list on page 1; (2) **absolute maximum ratings**: never exceed these; (3) **recommended operating conditions**; (4) **electrical characteristics** tables, always noting the test conditions (temperature, supply voltage), because a number without its conditions is meaningless; (5) typical-performance graphs; (6) the **pin-out** and package; (7) the application circuit. Practise on the datasheet of any common part, such as a voltage regulator or op-amp.

A WORD TO SOUTH AFRICAN LEARNERS

You do not need to live in Silicon Valley or Taiwan to start. Every tool in this plan is free or cheap and works on an ordinary laptop. What the industry lacks worldwide is people who understand both the physics and the economics, the two halves of this book. Start now, build things, and write down what you learn.

CHECK YOURSELF

1. **G10** Why are most of South Africa's chips "hidden" in imports?
2. **G10** Which two school subjects matter most for an engineering degree?
3. **G11** Name three roles in the chip industry and one skill each needs.
4. **G12** Using Figure 17.1, how many times larger were phone and network-gear imports than loose-IC imports?
5. **Varsity** Using Chapter 14's break-even idea, explain why design services are a better first business than a fab.
6. **Pro** Pick one opportunity from Figure 17.2 and list the three facts you would need to check before investing.

WATCH & EXPLORE

- [Siyavula](#) — free CAPS Mathematics and Physical Sciences practice, Grades 10–12. [siyavula.com](#) **G10**
- [NandGame](#) — build a computer from logic gates. [nandgame.com](#) **G11**
- [Ben Eater: Build an 8-bit computer](#) — the classic hands-on series. [eater.net/8bit](#) **G12**
- [HDLBits](#) — learn Verilog by doing. [hdlbits.01xz.net](#) **Varsity**
- [Tiny Tapeout](#) — make your first real chip. [tinytapeout.com](#) **Varsity**

IN PLAIN WORDS

South Africa buys lots of chips, but mostly hidden inside phones, computers and solar products. We do not make advanced chips here, and building a chip factory is far too expensive for now. But South Africans have designed chips before, and design needs skills more than money. Jobs in chips include designing, testing, checking and selling them, and studying the business. To start: take Mathematics and Physical Sciences, study engineering or physics, and build small projects, from an LED circuit to your own chip on a shared wafer.

SOURCES

1. South African trade data (HS 8541, 8542, 8471, 8517; 2023): UN Comtrade via TrendEconomy, as compiled in Naple research note RA-005, "The Chip South Africa Needs", Oct 2026.
2. Azoteq (founded 1998, Paarl), SAMES metering ICs, Conlog and STS (IEC 62055; about 70 million meters in 40+ countries): company websites, Enlit and News24 reports (2021–2024), as compiled in RA-005; verify current status.
3. University of Pretoria, Carl and Emily Fuchs Institute for Microelectronics (up.ac.za); STMicroelectronics Bouskoura site (st.com). Verify current programmes.

Resource Library

The best free places to keep learning, sorted by level. Every link was checked on 15 October 2026. Where a specific video could not be pinned down reliably, we link to the creator's channel; search it by topic.

TIP: LEARNING WITH VIDEOS

Watch once for the big picture. Watch again with this book open and pause to redo each calculation. Then explain the idea out loud to someone else in two minutes. If you can't, rewatch the part you skipped.

Start here G10 G11

Resource	Type	Creator	Why use it
Siyavula siyavula.com	Textbooks, practice	Siyavula (SA)	Free CAPS Physical Sciences and Mathematics, Grades 10–12
Circuit Construction Kit: DC phet.colorado.edu/en/simulations/circuit-construction-kit-dc	Simulator	PhET, Univ. of Colorado	Build circuits and measure current and voltage in a browser
Semiconductors phet.colorado.edu/en/simulations/semiconductor	Simulator (Java)	PhET	Doping, electrons, holes and p–n junctions
Branch Education youtube.com/@BranchEducation	Video channel	Branch Education	Animated explainers; search "How are Microchips Made?"
The Ridiculous Engineering of the World's Most Important Machine youtube.com/watch?v=MiUHjLxm3V0	Video	Veritasium	The EUV machines that print modern chips

Grade 12 and first steps in electronics G12

Resource	Type	Creator	Why use it
NandGame nandgame.com	Interactive game	NandGame	Build a computer from NAND gates, step by step
Build an 8-bit computer eater.net/8bit	Video series, kit	Ben Eater	A working computer from logic chips on breadboards
Asianometry youtube.com/@Asianometry	Video channel	Asianometry	Deep dives on chip history, fabs, companies and geopolitics

Varsity: courses and free textbooks VARSITY

Resource	Type	Creator	Why use it
Modern Semiconductor Devices for Integrated Circuits chu.berkeley.edu	Textbook (chapter PDFs)	Chenming Hu, UC Berkeley	Clear device physics from the co-inventor of the FinFET
6.012 Microelectronic Devices and Circuits ocw.mit.edu	Course notes, problem sets	MIT OpenCourseWare	Devices to circuits, with exams and solutions

Designing chips yourself Varsity Pro

Resource	Type	Creator	Why use it
HDLBits hdlbits.01xz.net	Exercises	HDLBits	Learn Verilog with instant automatic checking
Tiny Tapeout tinytapeout.com	Shuttle programme	Tiny Tapeout	Get a small design made on a shared chip (SKY, IHP, GF shuttles; check the price calculator)
wafer.space GF180MCU runs crowdsupply.com/wafer-space	Shuttle programme (paid)	wafer.space	Crowd-funded open-PDK manufacturing runs

Economics and market data Varsity Pro

Resource	Type	Creator	Why use it
Semiconductor Industry Association semiconductors.org	Data, reports	SIA	Monthly and annual sales, free factbooks
World Semiconductor Trade Statistics wsts.org	Data, forecasts	WSTS	The industry's official statistics and forecasts
Gartner newsroom gartner.com/en/newsroom	Press releases	Gartner	Annual top-10 vendor rankings
TrendForce press centre trendforce.com/presscenter	Press releases	TrendForce	Quarterly foundry and memory rankings
Energy and AI iea.org/reports/energy-and-ai	Report	International Energy Agency	Data-centre electricity, now and to 2030
TSMC wafer price estimates siliconanalysts.com	Data (partly paid)	Silicon Analysts	Collected analyst wafer-price estimates, with caveats

BOOKS WORTH BUYING OR BORROWING

C. Miller, *Chip War* (2022): the history and geopolitics of chips, very readable. C. Hu, *Modern Semiconductor Devices for Integrated Circuits*. S. M. Sze & K. K. Ng, *Physics of Semiconductor Devices*: the professional reference. N. Weste & D. Harris, *CMOS VLSI Design*: chip design. Check your university library or the National Library of South Africa.

LAST CHECKED

All links were checked on 15 October 2026. Websites, prices and programmes change; if a link breaks, search for the title and creator.

Formula Sheet and Constants

The key equations of the book in one place, with the chapter where each is explained.

Constants (SI, 300 K unless stated)

Quantity	Value	Quantity	Value
Elementary charge q	1.602×10^{-19} C	Planck constant h	6.626×10^{-34} J·s
Boltzmann constant k	1.381×10^{-23} J/K = 8.617×10^{-5} eV/K	Speed of light c	2.998×10^8 m/s
Thermal voltage kT/q	25.85 mV	Permittivity of free space ϵ_0	8.854×10^{-12} F/m
Si band gap E_g	1.12 eV	Si relative permittivity	11.7
Si intrinsic n_i	$\approx 1 \times 10^{10}$ cm $^{-3}$ (9.65×10^9)	SiO $_2$ relative permittivity	3.9
Si lattice constant	0.357 nm	Si atom density	5.0×10^{22} cm $^{-3}$
μ_n, μ_p (lightly doped Si)	$\approx 1400, \approx 470$ cm ^{2} /V·s	1 eV	1.602×10^{-19} J

Electricity (Ch. 1, 3)

Current, energy, power	$I = \frac{Q}{\Delta t}, \quad W = QV, \quad P = VI = I^2 R$
Ohm's law, resistivity	$V = IR, \quad R = \rho \frac{L}{A}$
emf and internal resistance	$\varepsilon = I(R + r)$

Bands, carriers, doping (Ch. 4–5)

Photon energy	$E = hf = \frac{hc}{\lambda}, \quad E(\text{eV}) \approx \frac{1240}{\lambda(\text{nm})}$
Intrinsic carriers	$n_i = \sqrt{N_C N_V} e^{-E_g/2kT}$
Mass-action law; doping	$np = n_i^2, \quad n \approx N_D \text{ (n-type)}, \quad p \approx N_A \text{ (p-type)}$
Fermi level	$E_F - E_i = kT \ln(n/n_i)$
Fermi–Dirac function	$f(E) = \frac{1}{1 + e^{(E-E_F)/kT}}$

Transport (Ch. 6)

Drift; conductivity	$v_d = \mu E, \quad \sigma = q(n\mu_n + p\mu_p), \quad \rho = 1/\sigma$
Diffusion current; Einstein	$J_n = qD_n \frac{dn}{dx}, \quad \frac{D}{\mu} = \frac{kT}{q}$
Diffusion length; Hall voltage	$L = \sqrt{D\tau}, \quad V_H = \frac{IB}{qnt}$

Junctions and light (Ch. 7)

Built-in potential	$V_{bi} = \frac{kT}{q} \ln \frac{N_A N_D}{n_i^2}$
Depletion width	$W = \sqrt{\frac{2\epsilon_{Si}(V_{bi} - V)}{q} \left(\frac{1}{N_A} + \frac{1}{N_D} \right)}$
Diode equation	$I = I_s (e^{qV/nkT} - 1)$
Solar cell fill factor, efficiency	$FF = \frac{P_{max}}{V_{oc} I_{sc}}, \quad \eta = \frac{P_{max}}{P_{light}}$

Transistors and chips (Ch. 8–10)

Threshold voltage	$V_T = V_{FB} + 2\phi_F + \frac{\sqrt{2\epsilon_{Si} q N_A (2\phi_F)}}{C_{ox}}, \quad C_{ox} = \frac{\epsilon_{ox}}{t_{ox}}$
MOSFET current	$I_D = \mu_n C_{ox} \frac{W}{L} [(V_{GS} - V_T)V_{DS} - \frac{1}{2}V_{DS}^2]; \quad I_{D,sat} = \frac{1}{2}\mu_n C_{ox} \frac{W}{L} (V_{GS} - V_T)^2$
Subthreshold swing	$S = m \frac{kT}{q} \ln 10 \geq 59.5 \text{ mV/dec (300 K)}$
Dynamic power	$P = \alpha C V_{DD}^2 f$
Rayleigh resolution	$CD = k_1 \frac{\lambda}{NA}$
Mass balance (Ch. 11)	$A_m \frac{dh}{dt} = Q_{in} - Q_{out}, \quad t_w = \frac{A_m(H - h_a)}{Q_{in}}$
Speed of sound in air	$c(T) = 331.3 \sqrt{1 + T/273.15} \text{ m/s}$

Economics (Ch. 12–14)

Average cost	$\bar{c} = \frac{F}{Q} + c$
Wright's law	$C_n = C_1 n^{\log_2 r}$
Dies per wafer	$N \approx \frac{\pi(d/2)^2}{A} - \frac{\pi d}{\sqrt{2A}}$
Yield (Poisson, Murphy)	$Y = e^{-AD_0}, \quad Y = \left(\frac{1 - e^{-AD_0}}{AD_0} \right)^2$
Cost per good die	$\frac{\text{wafer cost}}{N \times Y}$
Break-even volume	$V^* = \frac{\text{NRE}}{p - c}$

Glossary

Short definitions of the terms used in this book. The number in brackets is the chapter where the term is explained.

Acceptor

A dopant atom (e.g. boron) with one fewer valence electron than silicon; it creates a hole. (5)

ASIC

Application-specific integrated circuit: a custom chip made for one design. (10, 14)

Band gap (E_g)

The energy range with no allowed electron states between the valence and conduction bands; 1.12 eV in Si. (4)

Binning

Sorting working chips by speed or power and selling them as different products. (13)

Break-even volume

The number of units at which a custom chip becomes cheaper overall than the alternative. (14)

Built-in potential

The voltage across a p–n junction in equilibrium, created by the depletion-region charge. (7)

Chiplet

One of several smaller dies combined in a single package to act as one large chip. (13)

Cleanroom

A room with filtered air and controlled conditions where chips are made. (10)

CMOS

Complementary MOS: logic built from paired nMOS and pMOS transistors, using little standby power. (8)

Conduction band

The band of energies in which electrons are free to move and carry current. (4)

Defect density (D_0)

Killer defects per cm^2 on a wafer; sets yield. (13)

Depletion region

The zone near a junction emptied of mobile carriers, leaving fixed ions. (7)

Die

One chip cut from a wafer, before packaging. (10, 13)

Diffusion

Movement of carriers from high to low concentration. (6)

Diode

A device that conducts easily in one direction only, usually a p–n junction. (7)

Donor

A dopant atom (e.g. phosphorus) with one more valence electron than silicon; it gives a free electron. (5)

Doping

Adding small amounts of impurity atoms to control a semiconductor's conductivity. (5)

DRAM

Dynamic random-access memory: one transistor and one capacitor per bit; must be refreshed. (8)

Drift

Carrier motion caused by an electric field, $v = \mu E$. (6)

EDA

Electronic design automation: the software used to design chips. (10)

Electron-volt (eV)

Energy gained by an electron through 1 V: 1.602×10^{-19} J. (4)

EUV

Extreme-ultraviolet lithography at 13.5 nm wavelength. (10)

Fab / foundry

A chip factory; a foundry makes chips designed by other companies. (14)

Fabless

A company that designs chips but has no factory. (14)

Fermi level

The energy where the chance of a state being occupied is one half; the electrons' "water level". (5)

FinFET / GAA

Transistors whose gate wraps three sides (fin) or all sides (gate-all-around) of the channel. (9)

Flip-flop

A circuit storing one bit, updated on a clock edge. (8)

FPGA

Field-programmable gate array: a reconfigurable chip that can be set up as any digital circuit. (10)

HBM

High-bandwidth memory: stacked DRAM placed beside a processor, used in AI chips. (8, 15)

High-k dielectric

Gate insulator with a higher permittivity than SiO_2 (e.g. HfO_2), reducing leakage. (9)

Hole

A missing electron in a bond, behaving as a positive mobile charge. (2, 5)

IDM

Integrated device manufacturer: designs and makes its own chips. (14)

Integrated circuit (IC)

Many components made together on one piece of semiconductor. (1)

Intrinsic

Pure, undoped; intrinsic carrier concentration n_i . (5)

Lithography

Printing patterns onto a wafer with light through a mask. (10)

Majority / minority carrier

The more / less numerous carrier type in doped material. (5)

Mobility (μ)

How easily a carrier moves in a field, in $\text{cm}^2/\text{V}\cdot\text{s}$. (6)

Moore's law

The observation that transistor counts per chip double about every two years. (9)

MOSFET

Metal-oxide-semiconductor field-effect transistor: the voltage-controlled switch in almost every chip. (8)

MPW / shuttle

A wafer run shared by many designs to split mask costs. (10, 14)

n-type / p-type

Semiconductor doped so electrons / holes are the majority carriers. (5)

Node name

A marketing label for a process generation (e.g. "3 nm"); not a physical dimension. (9)

NRE

Non-recurring engineering: one-off costs of a chip design (engineering, tools, masks). (14)

OSAT

Outsourced semiconductor assembly and test. (14)

PDK

Process design kit: the rules and models needed to design for a factory's process. (10)

Photovoltaic effect

Generation of current by light in a p-n junction. (7)

RTL

Register-transfer level: a hardware description of a chip in an HDL such as Verilog. (10)

Saturation velocity

Maximum drift speed of carriers in strong fields, $\sim 10^7$ cm/s in Si. (6)

SRAM

Static RAM: six-transistor memory cell, fast, used in caches. (8)

Subthreshold swing

Gate voltage per tenfold change in off-state current; ≥ 60 mV/dec at 300 K. (9)

Tape-out

Sending a finished chip layout to the factory. (10)

Threshold voltage (V_T)

Gate voltage at which a MOSFET's channel forms. (8, 9)

Transistor

A semiconductor device that switches or amplifies current. (1, 8)

Utilisation

Fraction of a fab's capacity in use; drives cost per wafer. (14)

Valence band

The band of energies of the bonding electrons; full at 0 K in a semiconductor. (4)

Verilog

A hardware description language for digital circuits. (10)

Wafer

A thin slice of single-crystal silicon (today usually 300 mm) on which chips are made. (10)

Yield

Fraction of dies on a wafer that work. (13)

Answers to Check Yourself

Worked answers. Try each question properly before you look. Numbers are rounded sensibly; small differences from your answer are fine if your method is right.

Chapter 1 · What is a semiconductor?

1. Conductors: copper, aluminium, silver. Insulators: glass, rubber, plastic (PTFE). Semiconductors: silicon, germanium (also GaAs, GaN, SiC).
2. Ohm-metre ($\Omega \cdot \text{m}$). From $\rho = RA/L$: $\Omega \times \text{m}^2 \div \text{m} = \Omega \cdot \text{m}$.
3. $R = 1.7 \times 10^{-8} \times 2 / (0.5 \times 10^{-6}) = 0.068 \Omega$.
4. Doping: transistors and diodes. Electric field: MOSFET. Light: solar cell or LED. Temperature: thermistor.
5. Binary information uses two states, 0 and 1. Switches that control other switches can be combined into logic that computes and memory that stores. A tiny, fast, cheap, low-power switch therefore lets you build a computer.
6. Copper about $10^{-8} \Omega \cdot \text{m}$; glass 10^{10} – $10^{14} \Omega \cdot \text{m}$: about 18 to 22 orders of magnitude.

Chapter 2 · Atoms, bonds and the silicon crystal

1. 14 protons, 14 electrons, 4 valence electrons.
2. It has four valence electrons and needs eight in its outer shell. Sharing one electron with each of four neighbours achieves this.
3. Extra electron: phosphorus or arsenic (group 15). Missing electron: boron, aluminium or gallium (group 13).
4. $N = 8 / (0.5658 \times 10^{-7} \text{ cm})^3 = 4.42 \times 10^{22} \text{ cm}^{-3}$.
5. The freed electron becomes a mobile negative carrier, and the vacancy it leaves in the bond (the hole) can move as neighbours fill it. That makes two carriers from one broken bond.
6. Grain boundaries and disorder scatter and trap carriers, lowering mobility and lifetime and making devices unpredictable. Transistors need a perfect, uniform crystal.

Chapter 3 · Electricity essentials

1. Q per second $= 1.0 \times 10^{-3} \text{ C}$; number of electrons $= 1.0 \times 10^{-3} / 1.602 \times 10^{-19} = 6.24 \times 10^{15}$ per second.
2. $Q = I\Delta t = 0.30 \times 600 = 180 \text{ C}$. $W = QV = 180 \times 3.0 = 540 \text{ J}$.
3. Conventional current is drawn from the positive terminal round the circuit to the negative terminal. Electrons, being negative, actually move the other way in a metal wire. Circuit diagrams, formulas and datasheets use conventional current.
4. (a) $10 + 20 + 20 = 50 \Omega$. (b) $1/R = 1/10 + 1/20 + 1/20 = 0.2$, so $R = 5.0 \Omega$.
5. $R = \rho L/A = 1.68 \times 10^{-8} \times 200 / (1.0 \times 10^{-6}) = 3.36 \Omega$. $P = I^2 R = 8.0^2 \times 3.36 = 215 \text{ W}$, wasted as heat in the cord (and the appliance gets $8.0 \times 3.36 = 27 \text{ V}$ less).
6. Power saved $= 60 - 9 = 51 \text{ W}$. Energy saved $= 0.051 \text{ kW} \times 5 \text{ h} \times 30 = 7.65 \text{ kWh}$. Value $= 7.65 \times \text{R } 3.50 \approx \text{R } 26.80$ per month.
7. $I = V/R = 8.0/10 = 0.80 \text{ A}$. Lost volts $= 9.0 - 8.0 = 1.0 \text{ V} = Ir$, so $r = 1.0/0.80 = 1.25 \Omega$.
8. $v_d = I/(nAq) = 5.0 / (1.8 \times 10^{29} \times 2.0 \times 10^{-6} \times 1.602 \times 10^{-19}) = 8.7 \times 10^{-5} \text{ m/s}$, under a tenth of a millimetre per second.
9. A voltmeter is connected in parallel; a high resistance means it draws almost no current and so does not change the circuit it measures. An ammeter is connected in series; a low resistance means it adds almost no extra resistance (and drops almost no p.d.), so the current is unchanged.

Chapter 4 · Energy bands and the band gap

1. $1.12 \times 1.602 \times 10^{-19} = 1.79 \times 10^{-19} \text{ J}$.
2. Every state is occupied, so no electron can move to a slightly higher-energy state to gain momentum from the field. Equal numbers of electrons move each way, so the net current is zero.
3. $E = 1240/530 = 2.34 \text{ eV}$. Absorbed by silicon ($2.34 > 1.12$) but not by GaN ($2.34 < 3.4$).

4. $E_g \approx 1240/630 = 1.97$ eV (for example AlGaInP).
5. Glass (SiO_2) has a gap of about 9 eV, larger than any visible photon (1.65–3.3 eV), so visible light passes through. Silicon's 1.12 eV gap is below all visible photon energies, so it absorbs them all.
6. Silicon's gap is indirect. An electron must change crystal momentum to recombine, which needs a phonon. This slow process loses to non-radiative recombination, so little light comes out. Absorption of high-energy photons is still strong.

Chapter 5 • Electrons, holes and doping

1. A hole is a missing electron in a bond. When neighbouring electrons fill it one after another, the vacancy moves the opposite way to the electrons, exactly like a positive charge.
2. No. Each free electron is balanced by a fixed positive donor ion, so the material is neutral overall.
3. $p \approx N_A = 5 \times 10^{17} \text{ cm}^{-3}$; $n = n_i^2/p = 10^{20}/5 \times 10^{17} = 200 \text{ cm}^{-3}$.
4. In a semiconductor, heating creates many more carriers (n_i rises exponentially), which outweighs the mobility loss. In a metal the number of carriers is fixed and huge, and more lattice vibration only increases scattering.
5. Germanium: a smaller E_g gives a larger n_i (Eq. 5.1). At high temperature, intrinsic carriers swamp the doping and leakage rises, so Ge devices stop working properly at lower temperatures than Si.
6. $E_F - E_i = 0.02585 \ln(10^8) = 0.48$ eV.

Chapter 6 • Drift and diffusion

1. $v = \mu E = 1400 \times 100 = 1.4 \times 10^5 \text{ cm/s} = 1.4 \text{ km/s}$.
2. Electrons in the conduction band have a smaller effective mass and scatter less, giving higher mobility than holes in the valence band.
3. Drift: people walking downhill because of gravity (a push). Diffusion: perfume spreading through a still room (from crowded to sparse).
4. $\rho = 1/(1.602 \times 10^{-19} \times 10^{16} \times 429) = 1.46 \Omega \cdot \text{cm}$.
5. $D_p = 0.02585 \times 429 = 11.1 \text{ cm}^2/\text{s}$; $L_p = \sqrt{11.1 \times 2 \times 10^{-6}} = 4.7 \times 10^{-3} \text{ cm} = 47 \mu\text{m}$.
6. n-type: electrons are the majority carriers.

Chapter 7 • The p–n junction

1. Fixed ionised dopant atoms (negative acceptors on the p side, positive donors on the n side). Mobile

carriers have diffused across and recombined, so it is "depleted" of free carriers.

2. Near-zero current in reverse; a sharp exponential rise in forward bias around 0.6–0.7 V for silicon (see Figure 7.3).
3. $R = (5.0 - 3.0)/0.015 = 133 \Omega$. Use the next standard value up, 150Ω (about 13 mA).
4. In a bridge, current passes through two conducting diodes in series on each half-cycle, each dropping about 0.7 V.
5. $V_{bi} = 0.02585 \ln(10^{34}/10^{20}) = 0.83$ V.
6. $I = 1 \text{ mA} \times e^{0.12/0.02585} \approx 104 \text{ mA}$.
7. $60 \times 0.70 = 42$ V at standard temperature; somewhat higher on a cold morning. At $+40^\circ\text{C}$, each cell loses about $40 \times 2 \text{ mV} = 0.08$ V, giving $60 \times 0.62 \approx 37$ V.

Chapter 8 • The transistor

1. Source, drain and gate. The gate oxide insulates the gate from the channel, so the gate controls the channel by its electric field and draws almost no steady current.
2. The gate voltage at which a conducting channel forms. Like a tap handle: below a certain turn nothing flows; past it, water (current) flows, more the further you turn.
3. In either steady state one transistor of the pair is off, so there is no direct path from V_{DD} to ground; only tiny leakage flows.
4. 00 → sum 0, carry 0; 01 → 1, 0; 10 → 1, 0; 11 → 0, 1.
5. Doubling $V_{GS} - V_T$ multiplies I_D by 4. Halving L doubles I_D .
6. $P = 12 \times (0.8/1.0)^2 \times (1.5/2) = 5.8$ W.
7. DRAM stores charge on a capacitor that leaks away, so it must be rewritten periodically. SRAM's cross-coupled inverters actively hold their state while powered.

Chapter 9 • Inside the MOSFET

1. Bands bend down at the surface until E_i is below E_F ; electrons collect in a thin layer right at the oxide interface (Figure 9.1, right).
2. $\phi_F = 0.477$ V so $2\phi_F = 0.95$ V; $Q_{dep}/C_{ox} = 0.33$ V; $V_T - V_{FB} = 1.28$ V. The depletion-charge term grew most (from 0.10 to 0.33 V).
3. Integrate $I_D dy = W \mu_n C_{ox} (V_{GS} - V_T - V) dV$ from $y = 0$, $V = 0$ to $y = L$, $V = V_{DS}$: $I_D L = W \mu_n C_{ox} [(V_{GS} - V_T)V_{DS} - V_{DS}^2/2]$.
4. $T = 358$ K: $S_{min} = \ln 10 \times kT/q = 71$ mV/dec. Hotter chips need more gate swing per decade, so at fixed

V_T the off-current is higher; n_i and junction leakage also rise.

- V_T could not fall without raising off-state leakage (the 60 mV/dec limit), so V_{DD} stopped falling. Without lower voltage, power per transistor fell less than density rose, so power density increased.
- $EOT = 2.5 \times 3.9/25 = 0.39$ nm.
- The gate surrounds the channel, so it controls the channel potential far more strongly than the drain does. The drain's field cannot reach in to lower the source barrier.

Chapter 10 · How chips are made and designed

- Quartz sand → metallurgical-grade Si (arc furnace) → purified polysilicon (Siemens process) → single-crystal ingot (Czochralski) → sliced, polished wafers.
- Particles far larger than chip features would short or break circuits; one particle can kill a die.
- $CD = 0.35 \times 13.5/0.33 = 14.3$ nm; High-NA: $0.35 \times 13.5/0.55 = 8.6$ nm, about 40 % smaller.
- Specification → RTL → (verification) → synthesis → place and route → sign-off → tape-out.
- Reset loads a start value; on each enabled clock edge, if (count $\neq$ 0) count $\leq$ count - 1; .
- Examples: no NRE and no mask risk; fast time to market; can fix bugs after shipping; volume too low to justify an ASIC.

Chapter 11 · Case study: Raven-01

- $A_m = \pi(0.6)^2 = 1.13$ m²; volume = 1.70 m³; $t = 1.70/0.0005 = 3\,390$ s $\approx$ 57 minutes.
- Overflow would follow within minutes to hours, but repairs take days, so a warning that late cannot prevent a spill. Detecting partial blockage gives days of warning.
- $c(0) = 331.3$ m/s; $c(35) = 351.9$ m/s. Error at 1.5 m $\approx 1.5 \times (351.9/331.3 - 1) = 0.093$ m = 93 mm.
- At steady state $Q_{in} = K(1 - b)\sqrt{h^*}$, so $h^* = (Q_{in}/K(1 - b))^2 = h_0^*/(1 - b)^2$.
- A bug or misunderstanding in one implementation is unlikely to be repeated identically in an independent one, so any mismatch exposes an error.
- Both rain and a downstream blockage raise the level. Comparing several nodes tells them apart: rain raises all of them together, while a blockage raises the blocked manhole and those upstream but lowers flow downstream.

Chapter 12 · Why chips are an economics story

- Fixed: mask set, design engineers' salaries.
Variable: wafers, test time per chip, packaging per chip.
- At 100 000 units: US\$ 101 each. At 10 million: US\$ 2 each.
- Unit 256 is 8 doublings: $0.9^8 = 0.43$ of the first unit's cost.
- Chip capacity takes years and billions to add, while demand changes in months. Bread supply adjusts quickly and cheaply.
- $791.7/526.8 - 1 = 50\%$.
- Example: an ASIC with US\$ 2 million NRE saving US\$ 1 per unit needs 2 million sales; if the whole market is 500 000 units it can never break even.

Chapter 13 · The cost of a chip

- $70\,686/50 - 942.5/10 = 1\,413.7 - 94.2 \approx 1\,319$ dies.
- $AD_0 = 0.5 \times 0.2 = 0.1$; $Y = e^{-0.1} = 90.5\%$.
- $3\,000/(1\,319 \times 0.905) = \text{US\$ } 2.51$ per good die.
- For small $x = AD_0$: $e^{-x} \approx 1 - x$, and $(1 - e^{-x})/x \approx 1 - x/2$, so Murphy gives $(1 - x/2)^2 \approx 1 - x$. Both are $\approx 1 - x$.
- Two 400 mm² dies: 143 gross, $Y = e^{-0.4} = 67\%$, 95.9 good, 47.9 sets; US\$ 209 per set. Better than one 800 mm² die (US\$ 348), worse than four 200 mm² dies (US\$ 160).
- It still works at lower spec, so it earns revenue instead of being scrapped. This raises effective yield and serves cheaper market segments.

Chapter 14 · Fabs, design costs and business models

- $6\,000/0.7 + 2\,000 = \text{US\$ } 10\,571$ per wafer.
- $V^* = 400\,000/(3.50 - 1.20) \approx 174\,000$ units.
- Nvidia: fabless. TSMC: foundry. Intel: IDM. ASML: equipment. Arm: IP licensing. ASE: OSAT.
- To secure scarce capacity: foundries build only with commitments, and prepayment shares the capital risk.
- ASIC: lower power, smaller size, lower unit cost at volume, harder to copy. FPGA: no NRE, reprogrammable, faster to market, lower risk.

Chapter 15 · The global chip market

- $791.7/630.5 - 1 = 25.6\%$.
- Logic and memory.

3. $15.8 + 9.1 + 7.6 = 32.5\%$.
4. WSTS counts sales by delivery region, and most electronics are assembled in Asia.
5. Different methods: association sales data versus analyst vendor-revenue estimates, different company coverage, and preliminary versus final figures.
6. Memory prices could fall as new capacity arrives; AI spending could slow; forecasts made in booms are often revised down.

Chapter 16 · Supply chains, geopolitics and the AI boom

1. Materials (e.g. China, Japan); wafers (Japan, Korea); design (USA, Taiwan, Europe); fabrication (Taiwan, Korea, USA, China); assembly and test (Malaysia, China, Taiwan); product assembly (China, Vietnam, India); sale (worldwide).
2. A step that depends on very few suppliers, so a disruption there stops everything. Examples: EUV scanners (ASML); leading-edge foundry (TSMC).
3. AI accelerators are limited by how fast data reaches their arithmetic units; HBM stacks memory right beside the processor for very high bandwidth.
4. $945/415 \approx 2.3$ times.
5. For: denies rivals military-relevant capability and protects security. Against: speeds up rivals' self-sufficiency, costs exporters revenue, and fragments the market.
6. Examples: revenue from AI services; hyperscaler capex guidance; memory and GPU prices; utilisation of installed AI capacity; power availability.

Chapter 17 · South Africa, Africa and your career

1. Most chips are built into phones, computers, cars and solar equipment assembled abroad, so they are imported as finished goods, not as loose chips.
2. Mathematics (not Mathematical Literacy) and Physical Sciences.
3. Examples: verification engineer, SystemVerilog; analogue designer, circuit theory; test engineer, statistics.
4. $3\,130/191 \approx 16$ times.
5. A fab needs tens of billions of US dollars of fixed cost and enormous volume to break even. Design services need mainly skilled people and tools, so they break even at a far smaller scale.
6. Example (metering chips): the real annual volume of meters, the price customers pay per chip, and the NRE and unit cost of the chip.

About Naple Semiconductor



Naple Semiconductor is the chip-design arm of Naple Corporation, a South African company based in Durban, KwaZulu-Natal, founded by Esethu Tshintsholo.

We design digital chips from code, prototype and test them on FPGAs, and work towards silicon through shared-wafer runs on mature processes, the route Chapters 10, 11 and 14 describe. Our research work, such as the Raven-01 case study in Chapter 11, aims at problems that matter locally: infrastructure, utilities and the environment.

We wrote this guide because South Africa needs more people who understand both halves of the chip story: the physics that makes chips possible and the economics that decides which ones get built. If this book helps one learner choose Mathematics and Physical Sciences, build a first circuit, or design a first chip, it has done its job.

USE THIS GUIDE

Share it freely for non-commercial teaching, with attribution. Teachers are welcome to print chapters for classes.

HELP US IMPROVE IT

Found an error, a broken link, or something unclear? Tell us. Corrections will go into the next edition, along with chapters we left out of this shorter first edition: device fabrication in more depth, power and RF devices, memory technologies, packaging and reliability, and more on the economics of the AI era.

Contact: naplecorp@gmail.com

IN PLAIN WORDS

Naple is a small chip-design company in Durban. We made this free book so that more young South Africans can learn how chips work and how the chip business works, and maybe build chips themselves one day.