INDUSTRIES OF INDIA

# THE TRANSISTOR

*The device that taught matter to compute*
*- and remade the world around it*



KETINENI RAMARAO

# T H E

# T R A N S I S T O R

*The device that taught matter to compute - and remade the world around it.*

*Ketineni Ramarao*

Reference numbers in square brackets resolve in the companion volume, The Transistor - Source Companion.

# Contents


PART I - THE DEVICE 4

Chapter 1 - What a transistor is 5

Chapter 2 - The semiconductor idea 8

PART II - THE WORLD BEFORE 11

Chapter 3 - The vacuum tube age 12

Chapter 4 - The solid-state clue 15

PART III - THE INVENTION 18

Chapter 5 - Bell Labs and the problem 19

Chapter 6 - December 1947 22

Chapter 7 - Shockley’s junction 25

Chapter 8 - The announcement and the license 28

PART IV - THE MATERIAL 31

Chapter 9 - Germanium first 32

Chapter 10 - The case for silicon 34

Chapter 11 - Purity as an industry 37

PART V - THE MANUFACTURE 40

Chapter 12 - Printing the crystal 41

Chapter 13 - The planar process 43

Chapter 14 - Many on one slice 45

Chapter 15 - The MOS turn 47

Chapter 16 - Printing smaller 49

PART VI - THE PEOPLE AND COMPANIES 51

Chapter 17 - Shockley Semiconductor and the eight 52

Chapter 18 - Fairchild 54

Chapter 19 - The Texas way 56

Chapter 20 - Intel and memory 58

Chapter 21 - Japan learns the transistor 60

Chapter 22 - The Valley becomes silicon 62

PART VII - SCALING 64

Chapter 23 - Moore’s observation 65

Chapter 24 - Dennard’s rules 67

Chapter 25 - The node race 69

Chapter 26 - The factory that prints 71

Chapter 27 - The equipment makers 73

PART VIII - WHAT TRANSISTORS DO 75

Chapter 28 - Amplifying 76
Chapter 29 - Switching 78
Chapter 30 - Remembering 80
Chapter 31 - Power and light 82

PART IX - THE BUSINESS 84
Chapter 32 - The cycle 85
Chapter 33 - The foundry idea 87
Chapter 34 - The cost of the next step 89

PART X - POWER 90
Chapter 35 - The island 91
Chapter 36 - The chip war 92
Chapter 37 - The supply web 93

PART XI - LIMITS 95
Chapter 38 - The physics wall 96
Chapter 39 - After planar 98
Chapter 40 - Different switches 100

PART XII - THE CONNECTED SYSTEM 101
Chapter 41 - Energy and environment 102
Chapter 42 - The connected system 103
Chapter 43 - The long arc 105
Chapter 44 - A transistor glossary 107

# PART I - THE DEVICE

# Chapter 1 - What a transistor is

By one careful industry estimate, humanity had manufactured about 13 sextillion transistors by the late 2010s, and the total has been climbing ever faster since - by the early 2020s the industry's fabs were printing on the order of 10^21 transistors a year, more in that one year than in all the years before 2017 combined, by one industry estimate [S6]. A sextillion is a one followed by twenty-one zeros. Written out, the estimate is 13,000,000,000,000,000,000,000 individual devices, and a historian at the Computer History Museum, David Brock, has judged the transistor - specifically its most common modern form, the MOS transistor - the most frequently manufactured human artifact in history [H10]. Not the nail, not the brick, not the ball bearing. A device most people have never seen, because the ones that run the modern world are far too small to see.

One current example fixes the scale. The most powerful computer processors sold for artificial-intelligence work in 2025, Nvidia's Blackwell family, pack 208 billion transistors onto a pair of silicon dies the size of large postage stamps [I3]. Every one of those 208 billion devices does a job that, within living memory, was done by a glass bottle the size of a man's thumb, glowing hot and burning out on a schedule measured in months. The industry that makes them booked $791.7 billion in sales in 2025, its best year on record, and its trade association projected roughly a trillion dollars for 2026 [M1]. All of it rests on one idea, demonstrated once, in a laboratory in New Jersey, in the week before Christmas 1947 [H1].

Strip away the numbers and the device itself is disarmingly simple to describe. A transistor is a valve for electric current with no moving parts. An ordinary water valve controls a large flow with a small effort: a hand on a tap starts, stops or throttles a stream that the hand alone could never carry. A transistor does the same with electrons. A small electrical signal applied to one terminal of the device starts, stops or throttles a much larger current flowing through the other two. Because the controlling agent is itself electric, the valve can be worked by another circuit rather than by a hand, and because nothing inside moves, it can be worked very fast, very often, for a very long time.

From that single property, the whole of electronics descends, because the valve can be used in exactly two ways, and those two ways are the roots of every electronic system ever built.

The first way is amplification. Feed the transistor a weak, wobbling signal - the faint voltage induced in a radio antenna, the whisper picked up by a microphone, the dim pulse arriving through a long telephone line - and the large current it throttles can be made to follow the small one, wobble for wobble, but with far more power behind it. The output is the input made stronger. Before the transistor, this trick belonged to the

vacuum tube, the glowing glass valve that built radio, television, long-distance telephony and the first electronic computers, and the transistor was invented quite deliberately as the tube's solid-state replacement [H5][H10].

The second way is switching. Drive the control signal fully off or fully on and the transistor stops being a proportional valve and becomes a gate: shut or open, blocking current or passing it. A gate that is shut can stand for a zero; a gate that is open can stand for a one. Wire gates together so that some gates control others and you can build circuits that compare, remember, add and choose - the elementary operations of logic. Everything a computer does, from weather forecasting to the generation of this sentence, reduces at the bottom of the machine to vast armies of transistors opening and closing. A modern processor switches its transistors billions of times per second, which is why a machine with no moving parts can seem to think quickly.

One device, two jobs. Amplification serves the world of signals - radio, audio, sensing, communications. Switching serves the world of logic - computing, memory, control. The transistor is unusual among inventions in being indispensable to both, and that double usefulness is the first clue to its reach. The same object sits in a hearing aid and in a data centre, in a car's ignition controller and in the guidance computer of a rocket.

The word that early writers reached for was "solid," and it is worth pausing on, because everything the transistor displaced was in some way hollow. The vacuum tube worked because electrons boiled off a hot filament and crossed an empty space inside a glass envelope. It needed the heat, the vacuum and the glass, and it paid for them in bulk, power draw and fragility. The transistor's controlling dance of electrons happens entirely inside a solid crystal. No filament burns, so nothing burns out on a filament's schedule. No vacuum leaks. No warm-up. The device is cold, instant and small. That is why a device that began life barely able to outperform a good tube ended up replacing tubes almost everywhere: not because the first transistors were better, but because solidity removed the ceilings that heat and glass had imposed on electronics, and left engineers free to chase size, cost and reliability instead.

It is still possible to buy a single transistor - a speck of semiconductor in a metal can or a plastic blob, with three wire legs - and hobbyists and power engineers do. But the transistor's real habitat today is the crowd. Since 1959, the manufacturing art has consisted of printing not one device at a time but whole circuits at once: thousands, then millions, then billions of transistors patterned together on a wafer of silicon and wired into processors, memories and sensors on the same piece of crystal [H4]. The discrete transistor with three legs is the ancestor preserved in amber. The living species is the integrated circuit, and the living species exists in numbers that strain language.

The object itself has changed almost beyond recognition across its life. The first transistor was a crude thing: a sliver of the element germanium, a plastic wedge, and two gold foil contacts pressed against the crystal's surface a hair's breadth apart. It was built by hand, and its makers were delighted when the voltage on one contact modulated the current through the other by a factor of up to a hundred [H1]. A modern transistor is not assembled but printed, layer by layer, on a wafer of ultra-pure silicon, and its critical dimensions are measured in atoms. The most advanced production lines in the world, such as the TSMC fab in Kaohsiung, Taiwan, that began volume output of 2-nanometre-class chips in late 2025, now build transistors as stacks of semiconductor sheets a few atoms thick, each wrapped on all sides by its control electrode [I1][I2]. Between the gold foil of 1947 and the wrapped nanosheets of 2025 runs one continuous story of materials, machines, companies and people, and that story is this book.

A word about what the book is. It is not a textbook of semiconductor physics, though the physics is here, told in plain language, because without it nothing that follows makes sense. It is not a history of computers, though computers keep walking through the story, because the transistor remade them. It is the biography of one device - where it came from, what it is made of, how it is manufactured, who built the industry around it, what that industry costs and earns, where power over it now sits, and where its limits lie. The transistor is small. The industry it opened is one of the largest and most concentrated on earth. Understanding the second begins with the first, and the first begins with a puzzle about crystals that took physicists a century to solve.

# Chapter 2 - The semiconductor idea

Nature sorts materials, so far as electricity is concerned, into two great classes with a vast emptiness between them. At one extreme stand the metals - copper, silver, gold - through which electric charge flows almost as freely as water downhill; a good conductor carries roughly one mobile charge for every atom it contains. At the other extreme stand the insulators - glass, rubber, dry air - in which mobile charge is almost entirely absent, so that current simply cannot pass. The Nobel committee's spokesman in 1956, reaching for a picture of the gulf, asked his audience to imagine all the Easter pilgrims packed into St Peter's Square: that is a conductor. Then imagine the Antarctic with one solitary traveller: that is an insulator. And between those extremes, he said, lies a third family of materials, with carriers about as numerous as the dockers in a harbour on the day a loaded freighter arrives. That family is the semiconductors [H7].

The family embarrassed physics for the better part of a century, because semiconductors did not behave. Metals were obliging: their resistance rose smoothly as they warmed. But in 1833 Michael Faraday recorded an "extraordinary case" - crystals of silver sulfide whose conduction increased with temperature, the opposite of copper [H2]. In 1874 the German physicist Ferdinand Braun, later a Nobel laureate for wireless telegraphy, wrote the first description of a still stranger effect: at the point where a metal tip touched a crystal of galena, a lead ore, current flowed freely in one direction and not in the other [H2]. A contact that conducts one way only is a rectifier, and a rectifier is precisely what a radio receiver needs to extract sound from a radio wave. By 1901 the radio pioneer Jagadis Chandra Bose had patented the semiconductor crystal rectifier as a detector of radio waves [H2], and the "cat's whisker" detector - a fine wire pressed against a crystal, adjusted by hand until a sensitive spot was found - carried the first decade of wireless broadcasting. Nobody could say why it worked. The crystal set was an appliance built on an unexplained anomaly, and everyone who twiddled the whisker knew the anomaly was temperamental.

Explanation had to wait for quantum mechanics, and the wait lasted into the 1930s. The quantum picture, in plain terms, is this. Electrons in a crystal are not free to possess any energy whatever; they are confined to permitted bands of energy, separated by gaps they may not occupy. In a metal, the topmost occupied band is only partly filled, so electrons can accelerate freely and carry current. In an insulator, that band is completely full and the next empty band lies across a wide forbidden gap; a full band cannot carry current, and the gap forbids promotion, so the material is electrically inert. A semiconductor is an insulator with a narrow gap - narrow enough that heat, light or a small impurity can lift a few electrons across it, or leave a few empty places behind in the full band. In 1931 Alan Wilson at Cambridge published the quantum theory that made this picture precise, "The Theory of Electronic Semi-

Conductors," and seven years later Boris Davydov in the Soviet Union, Nevill Mott in Britain and Walter Schottky in Germany independently worked out the theory of how a semiconductor rectifies [H2]. The anomaly had become a science.

The practical miracle came from the impurities. A perfectly pure semiconductor, silicon say, has very few carriers and is little use. But dissolve a trace of the right foreign atom into the crystal and everything changes. Each atom of phosphorus lodged in a silicon crystal brings one electron more than the silicon lattice needs, and that spare electron is easily set roaming: the material becomes an n-type semiconductor, n for negative carriers. Each atom of boron brings one electron too few, and the missing electron - the "hole" - behaves, astonishingly, like a mobile positive charge in its own right, wandering the crystal as neighbouring electrons shuffle into the vacancy. The guest atom of boron, as the Nobel audience heard it, manages to be a positive carrier "by stealing an electron which his host, silicon, had kept locked up" [H7]. The concentrations involved are almost absurdly small: a few parts in a hundred thousand make a good semiconductor [H7]. The electronics of the twentieth century is, in a sense, the art of adding almost nothing to silicon, very exactly.

Now put the two doped kinds side by side. Where an n-type region meets a p-type region, the mobile electrons and holes near the boundary recombine and vanish, leaving a thin zone stripped of carriers - a zone current cannot easily cross. But the zone is asymmetric. Apply a voltage one way across the junction and it thins, carriers flood across, current flows; reverse the voltage and the depleted zone widens, and the current is blocked. The junction conducts one way only. It is a rectifier with no filament, no vacuum and no glass, and it is the fundamental building block of all semiconductor devices. The man who first saw one plainly was Russell Ohl, an electrochemist at Bell Telephone Laboratories, who in the mid-1930s was studying silicon rectifiers as detectors for radar. On 23 February 1940 Ohl tested a small silicon slab that behaved bizarrely: its current jumped when bright light fell on it, and different regions of the crystal gave opposite electrical effects under his probe. With his colleague Jack Scaff he traced the behaviour to a seam in the crystal where impurities of the two kinds met - phosphorus on one side, boron on the other - and in doing so named the regions n-type and p-type and identified the boundary between them as the p-n junction [H8]. Ohl had stumbled, in the same accident, on the photovoltaic effect that now powers solar cells. His employers patented the light-sensitive device in 1941 [H8].

One capability remained missing, and its absence is the reason the transistor was not invented in 1940. A rectifier has two terminals; it passes current or it blocks it, but nothing controls it from outside. The dream, obvious to everyone in the art, was a solid-state triode: a crystal device with a third electrode by which a small signal could command the main current, the way the grid of a vacuum tube commands the flow

from cathode to plate. Attempts to control semiconductor rectifiers with an extra electrode run back into the 1930s, and they all failed [H7]. A Polish-American physicist, Julius Lilienfeld, had even patented designs for a three-electrode semiconductor amplifier as early as 1926 - devices meant to be controlled by an electric field - but there is no evidence he ever built a working unit, and with the materials of the day he almost certainly could not have [H2][H10]. The crystal kept its secrets. Something at the semiconductor's surface, nobody knew quite what, swallowed the control signal every time.

That something - the physics of surfaces, and the physics of the thin, carrier-starved zone at a junction - was the puzzle that the war interrupted and that a laboratory in Murray Hill, New Jersey, would take up again in 1945 with unusual determination. The science was ready. The motive was ready too, because by then an entire industry had grown old and fragile on the vacuum tube, and the people who ran the world's largest telephone network wanted out. Their story comes next; but it begins, properly, with the tube itself.

# PART II - THE WORLD BEFORE

# Chapter 3 - The vacuum tube age

In a large room at the Aberdeen Proving Ground in Maryland, in the years after the Second World War, stood the most complex electronic machine yet built by anyone, anywhere. ENIAC - the Electronic Numerical Integrator and Computer - had been constructed at the University of Pennsylvania between 1943 and 1945 to calculate artillery firing tables for the United States Army, and it ran on vacuum tubes: almost eighteen thousand of them, glowing in racks that lined the walls [H11]. It was the first large-scale computer to operate at electronic speed, unhampered by mechanical parts, and for a decade it may have performed more calculations than all of humanity had managed before it [H11]. It was also, by construction, a reliability nightmare. With eighteen thousand tubes in circuit, ENIAC lost roughly one tube every day or two, and finding the failed one among the racks was the central fact of its operators' lives. Its engineers fought back with discipline: they ran the tubes gently, well below their rated voltages, tested components relentlessly, and drilled the maintenance team until a dead tube could be located and replaced in about fifteen minutes [H11]. That was what triumph looked like at the ceiling of the vacuum tube age: a machine of genius, kept alive by a fire brigade.

The tube had ruled electronics for four decades, and it had earned the throne. Its story begins inside an ordinary light bulb. In 1883, experimenting with his incandescent lamps, Thomas Edison sealed a metal probe into a bulb alongside the glowing filament and noticed that current would cross the empty space between them - but only in one direction [H13]. He patented the arrangement as a voltage indicator, saw no use for it, and moved on. The "Edison effect" waited twenty years for John Ambrose Fleming, a British engineer who had studied under Maxwell and served as scientific advisor first to Edison's London lighting company and then to Marconi's wireless company. In October 1904, hunting for a better detector of radio waves, Fleming remembered the one-way bulbs, had them retrieved from storage, and found they converted the oscillating current of a radio signal into a direct current a headphone could use. Within a month he had patented his "oscillation valve" [H5] [H13]. The diode valve was a detector, not an amplifier; it could hear a signal but not strengthen it.

The amplification that turned the valve into the master device of its century arrived two years later, from an American with a Yale doctorate, a flair for self-promotion and a genuine instinct for the main chance. Lee de Forest had spent years groping for a better wireless detector when, working from copies of Fleming's valves, he tried inserting a third electrode between the filament and the plate - not a solid sheet, which would have blocked the current, but a piece of wire bent back and forth like a miniature gridiron. The shape gave the component its name: the grid. Bulbs made to this design were delivered to de Forest on 25 November 1906, and when he finally

tested them on the last day of that year, they showed definite amplification [H13]. His patent application for the three-electrode “Audion” - submitted in October 1906 in an early form and completed in January 1907 - described a “device for amplifying feeble electrical currents” [H12][H13]. A small voltage on the grid now controlled a large current flowing through the valve. Feed it a faint radio signal and out came the same signal, stronger. Every electronic system built in the next fifty years rested on that trick.

The early Audions were wretched things - temperamental, inconsistent from one example to the next, short-lived, their behaviour clouded by residual gas. The engineering that tamed them came from the industrial laboratories, above all from Irving Langmuir at General Electric, who from 1909 subjected the triode to painstaking study: he moved the grid closer to the cathode to increase gain, perfected durable cathodes, and in 1915, simultaneously with Walter Schottky at Siemens in Germany, added a second “screen” grid to create the tetrode and push valves to higher frequencies [H13]. Progress was repeatedly choked by litigation. De Forest, Fleming, Marconi’s company, AT&T, GE and others fought over the valve patents for years; the litigation grew so destructive that in 1920, at the American government’s request, the combatants pooled their patents in a cross-licensing scheme administered around the new Radio Corporation of America, which proceeded to extract a royalty of 7.5 percent of the sale price of every radio [H14]. The pattern is worth remembering: control of a fundamental device patent, and the terms on which it is licensed, decided who profited from an entire industry. The transistor’s owners would face the same choice a generation later and choose differently.

Between the wars the tube built the electronic world. Broadcasting networks filled the airwaves; valve amplifiers in repeater stations carried telephone conversations across continents and, by mid-century, under oceans; talking pictures, television, radar and the first electronic calculators all ran on thermionic valves. Production rose to match: American factories turned out 108 million tubes in 1940 and 139 million in 1945 [H14]. Western Electric, the manufacturing arm of the Bell System, made over one and a half million tubes in 1944 and had millions more built to its designs by others [H14]. The tube was a mature, ubiquitous, profitable component - and it was flawed in ways that no refinement could remove, because the flaws were physical. A tube works by boiling electrons off a hot filament into a vacuum. It therefore consumes power continuously, takes time to warm up, runs hot, and is built of thin wires suspended in glass. Filaments evaporate and fail; glass breaks; vacuums leak. A radio with five tubes needed a repair every year or two as a matter of course. A machine with eighteen thousand tubes needed one every day.

The telephone network felt the ceiling first and hardest, because it was the largest electronic system in existence. Long-distance calls faded with distance unless they were re-amplified along the way, and by the 1940s the Bell System’s trunk routes

were studded with repeater stations full of tubes, each one consuming power, generating heat and drifting towards failure. Switching, meanwhile, still belonged to electromechanical relays - banks of clattering metal contacts that wore, pitted and stuck. Bell’s engineers could see the shape of the coming network, with its rising traffic and its appetite for automatic electronic switching, and they could see that a future built from hot glass bottles and moving contacts would drown in its own maintenance [H14]. The corporation that employed the world’s best communications physicists had the strongest commercial motive in the world to find something better.

This was the ceiling that the late 1930s had reached. Electronics was limited not by what engineers could conceive but by what tubes could survive. Every ambitious system - bigger computers, denser telephone exchanges, airborne radar - multiplied tube counts and thereby multiplied failures, heat and power draw in proportion. Inside the Bell System, whose business was reliability on a continental scale, the conclusion was drawn early and bluntly. Looking back, an internal AT&T history put it this way: the transistor “was not an accident,” because the company “had determined prewar that the evolving telephone system would be untenable with vacuum tubes” [H14]. The search for a solid-state alternative was not a hobby of curious physicists. It was a strategic program, funded by the largest electronics customer on earth, aimed from the start at the tube’s replacement.

# Chapter 4 - The solid-state clue

The idea that a crystal might do a tube's work was not born at Bell Labs, and it was not born in 1945. It had been visible, in fragments, for two decades - in patents nobody could reduce to practice, in rectifiers that worked without anyone fully knowing why, and in the stubborn memory of the cat's whisker, the crystal detector of radio's first decade, which had done a valve's job with a lump of ore before being swept aside by the Audion. The solid-state clue was always there. The problem was that nobody could make it amplify.

The most tantalizing fragment belonged to Julius Edgar Lilienfeld, a Polish-American physicist who between 1925 and 1928 filed patents in Canada and the United States describing three-electrode semiconductor devices meant to amplify [H2][H10]. His designs proposed controlling the current through a semiconductor film with an electric field applied across an insulating layer - in structure, startlingly close to the field-effect transistor that dominates electronics today. But there is no evidence that Lilienfeld ever built a functioning unit, and with the semiconductor materials of the 1920s it is unlikely that anyone could have [H10]. His patents sank without trace into the literature. Decades later, when credit was being weighed, John Bardeen himself acknowledged that Lilienfeld "had the basic concept of controlling the flow of current in a semiconductor" [H10]. The concept had preceded the materials that could honour it by a full generation.

Meanwhile a humbler solid-state industry was quietly thriving on rectification alone. Copper-oxide and selenium rectifiers - stacks of treated metal discs that passed current one way only - were manufactured in quantity from the 1930s for battery chargers, power supplies and instrument circuits. They were crude, their physics contested, and they worked. At Bell Telephone Laboratories, where the rectifier art was familiar, the obvious question hung in the air: if a semiconductor junction can decide the direction of a current, can a semiconductor device be made to control its size? In the late 1930s the laboratory put the question to one of its brightest theoretical recruits. William Shockley was charged with finding a solid-state replacement for the triode valve [H10]. With Walter Brattain, an experimentalist of legendary touch, he tried the field-effect route on the copper-based rectifier materials of the day: apply a field to the surface of the semiconductor, they reasoned, and the current inside should respond. Nothing responded. The surface of the semiconductor swallowed the field every time, and nobody knew why [H10].

Then the war intervened, and the war changed everything except the question. Radar was the decisive electronic technology of the conflict, and radar had a quiet secret: at the microwave frequencies where it worked best, the vacuum tube was useless as a detector. The electrons inside a tube could not respond fast enough, and the tube's

own structure added too much electrical inertia. The only devices quick enough were crystal rectifiers - the cat's-whisker principle reborn, a sharpened metal point pressed against a sliver of semiconductor inside a cartridge small enough to lose in a pocket. Their low capacitance made them natural microwave detectors [H16]. The RAF's early sets had flown with English "red-dot" silicon detectors [H16], and as radar became the war's electronic front, millions of crystal rectifier diodes were fabricated for Allied receivers [H15]. The demand dragged semiconductor materials science out of its doldrums and made it a war industry.

The transformation centred on two elements and a handful of laboratories. Under the wing of MIT's Radiation Laboratory, silicon was purified at the University of Pennsylvania under Frederick Seitz and at DuPont, reaching 99.999 percent purity by the war's end [H15]. Germanium fell to Purdue University, where the department head Karl Lark-Horovitz had won an Office of Scientific Research and Development grant in March 1942 for an improved point-contact rectifier [H16]. Lark-Horovitz chose germanium on the eminently practical ground that its low melting point would be kinder to crucibles than silicon's, then imposed a regime of purify-first, dope-second: the material had to be made as pure as humanly possible before its behaviour could mean anything. A single graduate student, Randy Whaley, was handed the crystal-growing problem and the supply line - germanium was an unimportant byproduct of the Eagle-Picher Company's tin production - and coaxed the raw dioxide through hydrogen reduction and long vacuum baking into respectable crystals [H16]. The program's first triumph belonged to Seymour Benzer, whose germanium rectifiers withstood reverse voltages above a hundred volts, solving the burnout problem that had plagued crystal detectors; the device was patented by Purdue, mass-produced at Bell Laboratories, and went into radar sets across the Allied inventory [H16].

The rectifier cartridge itself deserves a glance, because it is the transistor's closest ancestor. Inside its ceramic body sat a sliver of silicon or germanium and a fine tungsten wire, the whisker, pressed against the crystal's face; at the contact point, nature provided the one-way junction that let the diode strip a signal from a microwave carrier. Every one of the millions of cartridges shipped to the war [H15] was, in embryo, half of a transistor: a semiconductor, a contact, a controlled junction. What it lacked was the second contact and the third terminal - the means of control.

Two quieter results of the war years mattered as much as any device. At Bell Labs, the metallurgist Jack Scaff and the chemist Henry Theurer systematically mapped how specific impurities made silicon and germanium into n-type or p-type material - phosphorus and the other fifth-column elements donating electrons, boron and the third column making holes [H15]. And everyone involved absorbed a lesson that the pre-war art had not known: that a semiconductor's behaviour is written by parts-per-million contaminants, and that control of purity is control of the device. It was a measure of the starting point that Frederick Seitz's encyclopedic Modern Theory of

Solids, published in 1940, did not mention silicon or germanium at all [H16]. By 1945 the elements had a literature, a measured energy gap, a map of their carriers - and a purpose.

When peace came, the pieces lay scattered on the bench waiting to be assembled. There was a working quantum theory of semiconductors, fifteen years old [H2]. There was the p-n junction, observed and patented [H8]. There were newly purified crystals and the wartime art of handling them. There was a documented failure - the field that would not penetrate the surface - with nobody yet explaining it. And there was a customer: the Bell System, whose leadership had concluded before the war that the telephone network's future could not be built from vacuum tubes [H14]. In 1945 Mervin Kelly, the laboratories' executive vice president, reassembled Bell's peacetime research program and put the solid-state question at its centre. William Shockley returned from wartime assignments to organize a new semiconductor group, and among the men he gathered were Brattain, the master of the delicate experiment, and a new recruit, a theorist named John Bardeen [H1]. The group's brief was to understand semiconductors. Its prize, understood by everyone in the building, was the amplifier that the clue had been promising for twenty years. One more detail completes the scene, and it is the kind of detail on which the history of technology turns: when the moment came, in December 1947, the slab of germanium in Walter Brattain's apparatus had been grown at Purdue [H15][H16].

# PART III - THE INVENTION

# Chapter 5 - Bell Labs and the problem

The transistor was conceived by a committee, in the best sense of that abused word: a deliberate assembly of talent, given a problem and the means to attack it. The committee's chairman, in effect, was Mervin Kelly, and the institution was Bell Telephone Laboratories, the research arm of the American Telephone and Telegraph Company - then the largest corporation on earth, owner of most of the telephones in the United States and of the manufacturing company, Western Electric, that equipped them. To understand why the transistor was born there and then, it helps to understand Kelly, because the transistor's existence is the result of a research director changing his mind.

Kelly was a tube man. He had joined Bell before the First World War and risen through the vacuum-tube department; when he was made director of research in 1936, his professional life had been spent inside the thermionic world [H20]. His mandate was the telephone network's future, and the future he could see was troubled. Telephone traffic was growing; the network's amplifiers and switches were tubes and electromechanical relays; and both technologies were reaching the limits of their potential [H17]. Through the late 1930s Kelly poured money - by his own later accounting, hundreds of thousands of dollars - into making better tubes, and got nowhere worth the price [H20]. The failure taught him a lesson that industrial research directors rarely get to learn so cleanly: that an incumbent technology can be improved expensively past the point of diminishing returns while the replacement sits unexplored in the next field. Kelly began, in the late 1930s, to bet on solid state.

Bell Laboratories itself was half the invention's parent. It occupied a purpose-built campus at Murray Hill, New Jersey [H7], and it was financed by a regulated telephone monopoly whose revenues let it practice a kind of patience that ordinary companies cannot afford: research programs measured in decades, aimed at problems the network would not feel for years. Kelly embodied the style. Colleagues remembered him in conferences, eyes closed, leaning back, apparently asleep - until somebody made a mistake, whereupon he would snap awake and nail the error with a question [H20]. He ran a laboratory that could afford to let a theorist think for a year about why a field would not penetrate a crystal.

His first move was people. Solid-state physics barely existed as a discipline in America, and Bell did not have the staff for it, so Kelly started hiring the young Ph.D.s who were just then making it one [H20]. From MIT in 1936 he recruited a theoretical physicist named William Shockley - London-born, Palo Alto-raised, Caltech and MIT trained, brilliant, restless and, by the private testimony of colleagues, "ill-tempered, spoiled, almost uncontrollable" as a child and never entirely cured of it [H21]. Gordon Moore, who later worked for him, said Shockley "could see

electrons”; Charles Townes, who won a Nobel Prize for the laser, delivered the most quoted verdict: “He understood everything but people” [H21]. Shockley was not yet thirty. In 1939 he wrote a sentence in his laboratory notebook that records the program’s birthday: “It has today occurred to me that an amplifier using semiconductors rather than vacuum is in principle possible” [H21]. With Walter Brattain, a Bell experimentalist since 1929 - born in Amoy, China, raised on a Washington State ranch, a man whose hands could make recalcitrant apparatus confess - Shockley tried to build the amplifier. He failed, for reasons nobody yet understood [H10][H21]. The effort was real, not a sketch: Brattain and Shockley wound coils, evaporated films and measured nothing happening, patiently, in every configuration they could think of. The lesson they could not yet read was that the failure itself was the discovery waiting to happen.

The war suspended the hunt and, as the last chapter described, armed it. When Bell Labs regrouped for peace in August 1945, Kelly reorganized the entire laboratory and created a Solid State Physics group, with Shockley and the chemist Stanley Morgan in charge, and a semiconductor subdivision under Shockley [H17]. It was, not incidentally, the same building in which Russell Ohl had found the p-n junction five years earlier and in which Jack Scaff and Henry Theurer had mapped the doping of silicon and germanium during the war [H8][H15]. The new group did not have to look far for its materials or its precedents. Into it went Brattain, the device physicist Gerald Pearson, the chemist Robert Gibney; and onto its rolls came a new hire, John Bardeen - a Wisconsin-born theorist with a Princeton doctorate, a child prodigy who had skipped three grades, and arguably the most quietly formidable mind in the building [H17][H21]. The group’s charter was fundamental understanding, on the sound industrial logic that the quantum theory of semiconductors, laid down by Wigner, Seitz and others in the 1930s, had never yet been turned into a device [H17]. Kelly suspected solid-state science might yield the amplifier the telephone system needed, and he knew that if it did, the military - entering a Cold War and funding anything promising - would be a second customer [H17].

By every account the group was extraordinary. “I cannot overemphasize the rapport of this group,” Brattain recalled. “We would meet together to discuss important steps almost on the spur of the moment of an afternoon... one person’s remarks suggesting an idea to another... when we got to the place where something needed to be done, experimental or theoretical, there was never any question as to who was the appropriate man in the group to do it” [H17]. Shockley’s contribution, at first, was to preside without smothering; one historian has wondered whether he “had enough brains to leave them alone” deliberately or by temperament [H17].

The group’s immediate target was amplifiers for the telephone plant, but nobody in the room believed the stakes ended there. A solid-state amplifier - sturdier, smaller, more efficient than any tube - would serve every system the Bell System owned and

many it did not, and as the Cold War began, military agencies were investing lavishly in exactly such long-horizon research [H17]. Kelly had positioned his people where corporate need, scientific readiness and government money all converged.

The first order of business was the old failure. Shockley asked Bardeen to find out why his 1939 field-effect design had not worked [H21]. Bardeen’s answer, reached theoretically on 19 March 1946, is one of the great diagnoses in the history of physics: the design should have worked in the bulk of the material, but electrons at a semiconductor’s surface are not free like electrons inside it - they are caught in what he called surface states, a thin trap of immobile charge that shields the interior from any outside field [H21]. The field-effect amplifier had failed because the crystal’s own skin was eating the field. By early 1947 Bardeen, Brattain, Pearson and Gibney had demonstrated the barrier’s existence experimentally [H21]. The project now had an enemy with a name. Its work - the secret priority, known inside the labs as “The Surface State Job” - was to find a way through or around the barrier [H21]. What happened when they breached it occupied a single astonishing month, and it deserves a chapter of its own.

# Chapter 6 - December 1947

The month that made the transistor began, as experimental physics often does, with a suggestion over apparatus. On Monday, 17 November 1947, Robert Gibney proposed that Brattain try applying a voltage between a metal plate above a slab of germanium and a contact beneath it, so that a strong electric field would stand perpendicular to the crystal's surface - and that he add a drop of liquid electrolyte where the contacts touched [H21]. The electrolyte was the trick: it could neutralize the surface states, discharge Bardeen's trap, and let the field finally reach into the crystal. It worked. There was a measurable field effect at last, the first time anyone had pushed an outside field into a semiconductor and seen the current inside respond [H21].

Four days later, on 21 November, Brattain - following Bardeen's suggestion that he probe the surface with a sharp metal point surrounded by the electrolyte - had a functioning amplifier [H21]. It was a pitiful amplifier: it worked only at very low frequencies, it depended on a drop of liquid clinging to a crystal, and it was in no sense a practical device. But it amplified. The surface barrier had been breached, and the group knew the rest was now engineering of a kind they were very good at. What followed was two weeks of blackboards and lab benches, long hours and feverish pace, that Shockley later named "the Magic Month" [H21].

The decisive step was getting rid of the electrolyte - and it was taken through an insight of Bardeen's and a piece of watchmaker's improvisation by Brattain. Bardeen calculated that if two contacts were placed extremely close together on the germanium's surface, the effect they had found would be greatly enhanced: the second contact could inject the carriers that lifted the blockade the first contact created [H21]. Extremely close meant closer than any two wires could be soldered. Brattain's solution has become the most famous piece of bench craftsmanship in electronics. He cemented gold foil around the edge of a plastic wedge and, with a razor blade, slit the foil at the wedge's point with surgical precision, producing two gold contacts separated by about the width of a sheet of paper [H21]. On the afternoon of Tuesday, 16 December 1947, he attached a spring to press the wedge's point firmly against the germanium surface and wiggled the assembly until it came alive. His own understated record: "I had an amplifier with the order of magnitude of 100 amplification, clear up to the audio range" [H21]. The input voltage on one gold contact modulated, a hundredfold, the current through the other. The solid-state amplifier existed, and this time there was no electrolyte, no tube, no filament - just crystal, gold and physics.

What was happening inside the crystal took the group some months to untangle, and the untangling belongs to the next chapter. But the working picture, as it emerged, is worth fixing here because it is the birth certificate of all later transistor action. The germanium in Brattain's apparatus was n-type, carrying its current in electrons. Held

at the wrong voltage, its contact with a metal point became a rectifier: a carrier-starved zone formed that choked the current off. The second gold point, placed a paper's width away, was the key that opened the choke. When it was driven positive, it sprayed positive holes into the crystal, and around the holes the negative carriers gathered again: the blockade dissolved, current flowed, and because it took only a few holes - a small, cheap injection - to release a large electron current, the small signal on the second contact commanded the large one through the first [H7]. The Nobel committee's speaker, explaining it nine years later, compared it to bribing a picket line with a handful of gold coins [H7]. The point-contact transistor was, at heart, a rectifier whose blockade could be dissolved on demand - and that is what amplification is.

Driving home that evening, Brattain told his carpool what had happened and swore them to secrecy: "Bardeen and I produced an amplifier using the field effects at very low frequencies... the next night, I swore them all to secrecy" [H21]. Bardeen, a man who almost never brought work home, remarked to his wife Jane as she peeled carrots in the kitchen, "We discovered something today." "That's great," she answered, peeling on; she learned later what the something was [H21]. The two men agreed on the necessary next step, and Brattain recorded it in five words: "We should tell Shockley what we did today" [H21].

That telling changed the laboratory's internal weather permanently, and the next chapter must deal with it. For now, the sequence of events: Shockley - whose own account admitted the news "provoked conflicting emotions in me. My elation with the group's success was balanced by the frustration of not being one of the inventors" - recognized immediately what he was looking at and arranged a demonstration for Bell's executives on the afternoon of Tuesday, 23 December [H21]. The point-contact device was wired to a microphone and headphones. Brattain's notebook records that the circuit "was actually spoken over and... could be heard and seen on the scope presentation" [H21]. Nobody remembers what was said. Shockley called the demonstration "a magnificent Christmas present" [H1][H21].

The device the executives saw was, as hardware, almost absurd: a sliver of germanium - wartime-purity material that traced back to the Purdue program [H15] [H16] - a polystyrene wedge, two strips of gold foil and a spring. Everything important about it was invisible: the dance of electrons and holes at the crystal's surface, where one contact injected the positive carriers that unlocked the current flowing to the other. It had no moving parts, no vacuum, no heat. It drew a whisper of power. And its amplification, roughly a hundredfold, already spanned the full audio range that a telephone network cared about [H21].

What it did not yet have was a name. Bell's patent attorneys began documenting the invention within days of Christmas, and the naming problem went around the building. The device worked by transferring an electrical signal across a resistor-like

element - “trans-resistance” was the physicist’s description of its action - and it was John Pierce, a colleague with a gift for words who would later father the communications satellite, who contracted the idea into “transistor” [H1][H21]. The name was almost too modest. A resistor is a passive lump; the transistor was a resistor that answered to a signal. Within a few years the word would name an industry; within a few decades, the most numerous manufactured object in human history [H10].

The laboratory understood at once that it had to move from one hand-built wedge to a manufacturable product, and that the distance between those two things was great. The point-contact transistor was delicate and temperamental: its behaviour depended on the precise pressure and condition of the gold contacts, no two units behaved quite alike, and nobody could yet say whether it could be produced in quantity at tolerable cost [H3]. The announcement went out anyway, because Bell was a public institution as well as a private one. On 30 June 1948, at a press conference in New York, the laboratory showed the transistor to the world, with a spokesman’s carefully measured claim that it “may have far-reaching significance in electronics and electrical communication” [H1][H21]. The world was underwhelmed. The New York Times carried the news the next day on its “News of Radio” page, below the announcement of a soap opera’s new sponsor [H21]. It is one of history’s small jokes that the most consequential invention of the twentieth century debuted beneath a detergent advertisement - and one of its large lessons that nobody, not even the men holding the device, could yet see what they had. What they had was not, in truth, the transistor that would conquer the world. That one existed only in Shockley’s head, and he was already at home, writing.

# Chapter 7 - Shockley’s junction

William Shockley spent the Christmas season of 1947 in a state that his colleagues came to recognize as dangerous. He had built the group, set the problem, and suggested the field-effect line of attack; and the device that resulted had been made by two other men, in a series of experiments he had not directed day to day, using an effect he had not predicted [H3][H21]. His response revealed both the worst and the best of him. First came the worst: he called Bardeen and Brattain in separately, Brattain recalled, and told each of them “that sometimes the people who do the work don’t get the credit for it” - to which Brattain answered, “Oh hell, Shockley, there’s enough glory in this for everybody” [H21]. In the laboratory’s publicity machine, Shockley became the transistor’s visible spokesman; an order came down that no pictures were to be taken of Bardeen and Brattain without him [H21]. Then came the best. Shockley went home, and for a month he worked alone in a sustained burst of theoretical fury that stands comparison with any in the history of physics [H3][H21].

He emerged with two things. The first was the correct theory of what Bardeen and Brattain had actually made. The point-contact transistor, Shockley argued, did not work by the surface field effect everyone had been hunting. Its secret was that the injecting contact pushed holes not along the surface but into the body of the germanium - “minority carrier injection,” carriers of the opposite sign driven into the bulk crystal, where they lived briefly, drifted, and unlocked the main current [H3]. If he was right, the whole game changed: the device’s heart was not the treacherous surface but the interior, and the interior could be engineered. On 16 February 1948 the physicist John Shive settled the matter with a beautiful experiment: he made a germanium sliver with point contacts on opposite faces - so that no surface path between them existed - and it still showed transistor action [H3]. Holes really were flowing through the bulk. Shockley was right - and the consequence was liberation. If the working current travelled inside the crystal rather than along its skin, then the device no longer lived or died by the condition of a surface that a fingerprint could change.

The second product of the month was the design that followed, set down in Shockley’s notebook on 23 January 1948: the junction transistor [H3]. Instead of two point contacts pressed onto a surface, imagine the p-n junction itself - Ohl’s junction, the grown-in boundary between doped regions - as the working element, and build a sandwich: a thin layer of p-type crystal between two thicker n-type slabs. The thin middle layer, the base, becomes the control electrode. Feed it a small current of holes and it modulates the great electron current crossing the sandwich from one outer slab, the emitter, to the other, the collector [H3]. Every contact is a robust junction inside the crystal, not a delicate point on its skin. Shockley saw at once that such a device would be quieter, more predictable and - the property he cared about most, watching

the point-contact units come off the bench each with its own personality - manufacturable [H3]. He filed for a patent in June 1948 [H3]. The bipolar junction transistor, as the world came to call it, is the form in which the invention first conquered industry; the design principles in that 1948 notebook still govern the bipolar transistors made today [H3].

Shockley did not keep the idea to himself. In 1949 he published the full theory of p-n junctions and junction transistors in the Bell System Technical Journal, and in 1950 he gathered the whole science into his book Electrons and Holes in Semiconductors, which became the field's first textbook [H3]. The publication policy was deliberate: Bell's management had decided that the transistor was too big to hoard, and the reasons for that decision, and its consequences, are the subject of the next chapter.

Between the design and the device, however, stood the hardest material problem the young art had yet faced. A junction transistor required a crystal in which the doping changed, precisely, twice, over a distance of thousandths of an inch - n, then an exquisitely thin p, then n again - inside a single piece of germanium of wartime purity or better. No such crystal existed. The man who forced the issue was Gordon Teal, a Bell Labs chemist who argued, to a largely deaf management, that the whole field was building on mud: that polycrystalline germanium, full of grain boundaries and stray impurities, could never yield reliable junction devices, and that large single crystals were the only foundation [H9]. With minimal support, Teal built his own crystal-growing apparatus with the mechanical engineer John Little and the technician Ernest Buehler, adapting a 1917 method of the Polish chemist Jan Czochralski: dip a seed crystal into molten germanium and withdraw it slowly, and the melt freezes onto the seed as one continuous crystal [H9]. Shockley, who had not listened, later called Teal's achievement "the most important scientific development in the semiconductor field in the early days" [H9].

From Teal's crystals came the device. His colleague Morgan Sparks found that he could create the junctions during the growth itself: by dropping pellets of impurity into the melt in sequence - first a p-type dose, then an n-type one - he grew a crystal with a thin p-layer frozen between n-type regions, a grown junction [H9]. In April 1950 Sparks and Teal produced the first grown-junction transistors; within a year they surpassed the best point-contact devices [H9]. Bell Labs announced the junction transistor to the world on 4 July 1951, at a press conference starring Shockley [H9] [H21]. The choice of Independence Day was presumably accidental.

The Americans were not, it turned out, alone. In Paris, at a Westinghouse subsidiary's laboratory, the German physicists Herbert Mataré and Heinrich Welker - working from the same wartime germanium rectifier tradition - independently developed their own point-contact transistor in 1948 [H2]. Their "transistron" briefly went into French telephone equipment. The parallel invention confirmed what the Bell group already

suspected: the time for the device had come, the physics was abroad in the world, and credit would belong to whoever built not just the first device but the working industry.

The invention's human cost was by then already being paid. Bardeen, his working relationship with Shockley poisoned, left Bell Labs in 1951 for the University of Illinois - a loss Kelly later conceded he had seen coming and failed to prevent [H20]. There Bardeen did the work on superconductivity that brought him a second Nobel Prize in Physics, in 1972 [H22]. Brattain stayed at Bell, increasingly detached from Shockley's orbit, content to let the devices speak for the man who had made them work. The group that had made the point-contact transistor did not survive its success. On 10 December 1956, however, the three men stood together in Stockholm one last time: Shockley, Bardeen and Brattain, sharing the Nobel Prize in Physics "for their researches on semiconductors and their discovery of the transistor effect" [H6]. The point-contact device had won them the prize. The junction design - conceived in one month of jealous solitary labour - was winning the world.

# Chapter 8 - The announcement and the license

Every industrial civilization gets a few decisions that compound for a century. The Bell System made one of them in the early 1950s, and it was not a technical decision: it was to give the transistor away - not for nothing, but on terms so gentle that the decision effectively seeded an industry rather than defending a property.

The context matters. AT&T in those years was the largest private corporation on earth: 746,000 employees, revenues of $5.3 billion - nearly two percent of the gross domestic product of the United States [P1]. Its research arm employed six thousand people, a third of them professional scientists and engineers, about one percent of the entire American technical workforce [P1]. It was also, for exactly those reasons, permanently in the sights of the antitrust authorities. The Justice Department filed its great suit against the Bell System in January 1949, six months after the transistor's announcement [P1]. The suit sought, among other things, to split Western Electric, the manufacturing arm, from the parent. The transistor thus spent its infancy inside a corporation fighting for its structure, and the fight shaped the child's upbringing.

Bell's first instinct had been conventional. The laboratory patented the work thoroughly - the solid-state group filed 166 patents around the transistor, 110 of them published by 1952 [P2] - and Western Electric tooled up to manufacture. The first commercial transistors were the point-contact Type A units, produced in their many thousands despite their delicate construction [H1]; by mid-1952, Western Electric's plant was making more than 6,000 devices a month, most bound for telephone switching and for the first market that would pay a premium for smallness and thrift, hearing aids [H21]. The Sonotone 1010 of December 1952, the first commercial hearing aid to use a transistor, sold for $229.50 and still kept two tubes alongside its single transistor - a hybrid of two eras in one pocket-sized case [H2][H21][S8]. Within two years the hybrid was extinct: by 1954, 97 percent of new hearing aids used transistors alone, Raytheon rode that volume to become the world's largest transistor manufacturer from 1952 to 1955, and the price of its little CK722 transistor fell from $7.60 in early 1953 to 99 cents by 1956 - a seven-fold collapse in under four years, the first taste of a price curve the world would come to live on [S8]. The hearing-aid makers understood what they were buying: the transistor's small size and miserly power draw were worth a premium wherever portability and battery life ruled, and the lesson - find the customers who will pay for the new device's advantages while it is still expensive - became the industry's standard playbook for every technology that followed [H2].

Then came the decision. Inside Bell, the argument was led by Jack Morton, the engineer running the semiconductor development program, and his own statement of it survives: "We realized that if this thing was as big as we thought, we couldn't keep

it to ourselves and we couldn't make all the technical contributions. It was in our interest to spread it around. If you cast your bread on the water, sometimes it comes back angel food cake" [H18]. There was hardheadedness under the folksiness. AT&T was a regulated monopoly whose business was telephone service, not components; every improvement other companies made in transistor technology was an improvement Bell could use or buy; and a licensing policy generous enough to disarm the antitrust charge of hoarding was worth more than any royalty stream [P1]. In September 1951 Bell held a first symposium, aimed principally at military users, to tell three hundred scientists and engineers what the transistor could do [H18][H19]. What it could not tell them in a lecture was how to build one. For that, Bell set a price: a $25,000 advance against royalties, and twenty-six companies, American and foreign, paid it [H18].

The $25,000 was a non-refundable advance against future royalties - five percent of net selling price at first, cut to two percent in 1953 - and Bell waived royalties entirely for hearing-aid applications, in homage to Alexander Graham Bell [S7]. It was real money in 1952, deliberately so: it filtered for companies serious enough to build, and it bought not just patent rights but the teaching. What Bell was selling, at a price any large electronics firm could pay, was the contents of its own learning curve - the zone refining, the crystal growing, the junction forming, the surface preparation, the circuit lore [H19]. An economist reading the arrangement today recognizes it as an unusually pure case of technology transfer: not the sale of a device, but the sale of the ability to make one [P2].

In April 1952 the licensees came to Murray Hill for the Transistor Technology Symposium - over a hundred engineers from forty companies, worked day and night through the laboratory's accumulated know-how, and walked through Western Electric's new transistor plant at Allentown, Pennsylvania, to see it done [H19]. The roster mixed the titans with the unknowns: General Electric and RCA sat alongside a small Dallas geophysical-instrument firm called Texas Instruments and a young Tokyo company named Sony [H19]. Everything taught was bound into volumes - Transistor Technology, later published commercially - that the industry came to know, with exasperated affection, as "Ma Bell's Cookbook" [H18][H19]. Economists who have since measured the aftermath find the effects plainly visible in the record: the spread of transistor technology into other fields, and the spillover of inventions built on it, jumped after 1952, and the jump tracks the licensing-and-teaching program rather than any reduction in royalties [P2]. The transistor diffused because Bell taught, not merely because Bell permitted.

The contrast with the last time a foundational electronic device had been born is stark. When the vacuum tube was young, its patents became weapons: years of litigation, a stagnant art, and finally a government-brokered pool under RCA that taxed every radio sold at 7.5 percent [H14]. The tube's owners had extracted rent; the transistor's

owner chose to multiply builders. A generation later, when a federal consent decree in January 1956 formally required Bell to license its entire pre-1956 patent portfolio - 7,820 patents - royalty-free to all comers, and confined the company to the telecommunications business, the decree mostly ratified what the transistor program had already demonstrated: that the Bell System had learned to prefer a big industry it did not control to a small one it did [P1]. The licensees did not disappoint. Within a few years the Dallas firm would bet its future on silicon and build the first commercial silicon transistors; the Tokyo firm would build a radio that changed what the world expected of Japan; and a procession of engineers who learned the art at Bell's knee, or at a licensee's, would carry it west to California. Those are the stories of Parts V and VI. The decree's shadow side - the rule barring Bell from general business - would echo strangely through the rest of this book: the institution that invented the transistor, and later much of the computer's basic art, was forbidden to sell computers. But that is a later chapter's irony. In 1952, with the cookbook handed out and the licensees scattering home to Dallas and Tokyo and a hundred other places, the transistor ceased to be a laboratory's possession and became an industry's.

It is worth pausing on what was being freed. Only about 31 percent of Bell's patents concerned telecommunications at all; the rest were scattered across thirty-four other technical categories, from materials processing to measuring instruments to semiconductor devices themselves [P1]. The consent decree therefore did not merely open a telephone company's toolbox; it opened one of the largest concentrations of industrial science ever assembled to any American business that wanted a look.

Which raises the question the next part of this book must answer: an industry making what, exactly, out of what? The transistor of 1952 was germanium, grown a crystal at a time, expensive and touchy. The road from there to silicon wafers printed by the billion begins with the material itself.

# PART IV - THE MATERIAL

# Chapter 9 - Germanium first

For the first six years of its commercial life, the transistor was a germanium device, and it is worth asking why, because the answer explains both the element’s reign and its fall. Germanium did not win because it was better. It won because it was ready: the wartime radar programs had already taught American laboratories to purify it [H15], its modest melting point - 937 degrees Celsius, against silicon’s punishing 1,415 - made it tractable in the crucibles and furnaces of the day [T1], and the first working device happened to be made of it. Technologies are often standardized by the accident of which material was on the bench when the breakthrough came, and so it was here: germanium’s head start, earned in radar, became the transistor industry’s first foundation.

The germanium years have a texture of their own, half laboratory, half workshop. The point-contact transistors of 1948 to 1951 were essentially handmade: a sliver of crystal, a wedge, gold contacts, a spring, adjusted until the unit came alive, each one a little different from its brothers [H1]. The grown-junction transistor that replaced them was more dignified but no less artisanal in its way. A seed crystal dipped into a crucible of molten germanium was withdrawn slowly, growing a single crystal behind it like a candle pulled from wax; at precisely timed moments, pellets of impurity - first one type, then the other - were dropped into the melt, freezing the n-p-n sandwich into the crystal’s length [H9]. The crystal was then sliced along its axis and the thin base layer contacted by hand. Yields were modest, uniformity was a hope rather than a specification, and the thin base layer - on which speed depended - was bounded by what a man watching a crucible could time.

The industry’s first great transfer of this art came in the spring of 1952, when Bell gathered its licensees for a symposium and laid the whole practice open: the crystal growing, the junction forming, the handling tricks, the testing. The two-dozen-odd companies that paid the license fee did not merely buy a patent position; they bought a craft, demonstrated bench by bench, that would have taken each of them years to reconstruct alone [H18]. Within months the first commercial germanium devices were on the market - hearing aids first, then telephone equipment - and in 1953 an experimental transistorized computer prototype demonstrated in compact form the size and power savings that the tube could never offer [H2].

Yet it worked, and it sold. Through the early and middle 1950s, germanium transistors went into hearing aids, into the first pocket radios, into telephone equipment, into military systems and into the first transistorized computers, where their advantages over tubes outweighed their vices. Raytheon, a tube company that had pivoted hard, stood as the largest supplier of transistors to the open market [T2]. The germanium

transistor was, for its moment, a triumph: it proved the device, created the market, and trained the first generation of semiconductor engineers.

Germanium had a second, quieter weakness: it is scarce. The element does not occur in minable concentrations; it arrives in the world as a trace byproduct of other metals' refining - the wartime program's germanium had come from the Eagle-Picher Company as an incidental output of its tin production [H16]. An industry built on germanium was an industry built on the crumbs of other industries, with supply set by someone else's demand for zinc and tin. Nobody in 1952 knew how large the semiconductor market would become, but anyone who priced the raw material could see that "rare byproduct" was not the foundation for a mass industry. That argument, like the temperature argument, pointed at the same abundant alternative.

Nor did the art stand still while it waited for the answer. In 1954, following their work using high-temperature diffusion to make solar cells, Bell's Charles Lee and Morris Tanenbaum applied diffusion to transistor making: instead of freezing the junctions into a growing crystal, they drove impurity atoms into a finished wafer in a furnace, letting heat do with precision what a man's timing in front of a crucible could not [H2]. Diffused-base germanium transistors reached frequencies the grown-junction art could not touch. The technique was the germ of everything that followed: once junctions could be printed into a wafer's face with furnace precision, the wafer - not the crucible - became the factory floor.

Its limits were constitutional, and everyone in the industry knew them. Germanium's narrow band gap - the property that made it so easy to coax into conducting - made its devices leaky and heat-shy. A germanium transistor in the "off" state still passed a treacherous trickle of current, an anathema for computer logic, where thousands of off devices must stay decisively off [T2]. Worse, the element's operating range ran from 0 to 70 degrees Celsius [T2]: fine for a hearing aid in a shirt pocket, hopeless for a missile's electronics bay, an aircraft at altitude, a desert installation, or a machine room full of glowing racks. The military, the era's most demanding customer, wanted ruggedness germanium could not offer. The computer makers wanted logic that stayed off. And the engineers who understood materials knew that forty-some elements to the left of germanium in the periodic table there was nothing better, but directly above it - same column, same family - sat an element with none of these vices and one overwhelming practical problem. The problem was making it pure, and the element was silicon.

# Chapter 10 - The case for silicon

Silicon sits directly above germanium in the periodic table: same column, same family, the same four outer electrons that make both elements semiconductors. Everything else about the two elements could hardly differ more. Germanium is a rare trace of other men's mines [H16]; silicon makes up 27.7 percent of the Earth's crust by mass, second in abundance only to the oxygen it is usually married to [T4]. It is the stuff of sand and quartz. If the semiconductor industry was going to grow into a mass industry - and by 1954 its leaders were beginning to believe it might - the material argument for silicon was the argument of a man choosing to build with stone instead of ivory.

The physics argued the same way, and more urgently. Silicon's band gap - the energy an electron needs to break free and conduct - is wider than germanium's, and the consequences run straight through the device's behavior. In germanium, ordinary warmth is nearly enough to shake carriers loose, which is why a germanium transistor leaks current in the "off" state and fails outright in heat, restricted to a range of 0 to 70 degrees Celsius [T2]. In silicon, the carriers stay put. Silicon devices showed far less of the leakage that was anathema to computer logic, and once DuPont began supplying high-purity semiconductor-grade silicon, they could operate from 55 degrees below zero to 125 above [T2]. For the military customer - whose electronics flew in jets, rode in missiles, and sat in deserts - that range was not a specification but a passport.

The numbers behind the gap tell the story precisely. At room temperature, the thermal energy jostling every atom in a crystal amounts to about 26 millielectron-volts. Germanium's band gap is about 0.67 electron-volts; silicon's is about 1.12 [P3]. Both gaps dwarf the thermal energy - yet conduction in a semiconductor is an exponential affair, in which carriers are freed in proportion to an exponential of the gap, and germanium's smaller barrier lets through, at room temperature, roughly two thousand times more stray carriers per cubic centimeter than silicon's [P4]. That exponential is the whole germanium problem in one line: leakage, noise, and heat sensitivity are not defects of germanium manufacturing but consequences of its gap, and no amount of engineering could repeal them.

And silicon carried one more gift, less obvious in 1954 than the other two and in the long run the most valuable of all. Expose a silicon surface to oxygen and it grows its own skin: a thin, dense, adherent layer of silicon dioxide - glass, in essence - that is an excellent insulator, that clings to the crystal beneath it, and that can be opened selectively where a chemist wants it open and left closed where he does not [H4]. Germanium's oxide is water-soluble and worthless. Silicon's oxide would become the mask, the insulator, and the gate of nearly every transistor made since. The age of

silicon was made possible by purity and promised by abundance, but it was built, quite literally, on silicon's rust.

It is worth pausing on the material's first industrial patron. The silicon in those early devices did not come from a semiconductor company at all but from DuPont, the chemical giant, which had learned to make semiconductor-grade silicon for its own reasons and began supplying the trade [T2]. The pattern would repeat throughout the industry's history: behind the famous device makers stood quieter specialists in materials, gases, and machines, without whose patience the famous names would have had nothing to work with.

The problem was making the stuff pure enough to use. Silicon melts at 1,415 degrees Celsius and, molten, reacts with nearly every material one might melt it in, dissolving its own contamination [T1]. Purifying it demanded a technique that did not yet exist when the war ended - which is the subject of the next chapter - and for years the demand kept silicon a laboratory promise rather than a product.

The promise was kept first, quietly, at Bell Labs. In January 1954 Morris Tanenbaum, a young chemist there, fashioned the first working silicon transistor, a variation on the grown-junction method [T2]. Bell Labs did not pursue it commercially [T2]. The reasons were reasonable ones: Bell had an enormous installed base of germanium knowledge and equipment, its own business was telephony rather than military hardware, and the grown-junction silicon device was difficult to make. Reasonable reasons often read differently in retrospect.

The man who made silicon an industry had been Teal's colleague.

Teal's move deserves a second look, because it was the industry's second great transplantation. Shockley's departure from Bell had carried the transistor idea to California [M17]; Teal's carried the silicon idea to Texas. Bell Labs, the most generous laboratory in industrial history, was proving to be an equally generous exporter of its own people, and each export seeded a new industrial geography. The silicon industry was born in Dallas before it was born in Silicon Valley.

Gordon Teal - the same Teal whose single crystal had saved the junction transistor in 1950 [H10] - had grown silicon crystals at Bell as early as that year [T3], and in 1952 he had gone home to Texas, to Texas Instruments, then a small geophysical-instruments company with ambitions. There he hired a team under Willis Adcock and set them to work on the material Bell had set aside. On April 14, 1954, Teal walked into the office of TI's vice president Pat Haggerty with a working silicon transistor, built from DuPont's semiconductor-grade material [T3].

Then came the moment that announced the material to the world. On May 10, 1954, at the Institute of Radio Engineers' national conference in Dayton, Ohio, Teal sat through a late-afternoon session in which speaker after speaker dismissed the silicon

transistor as years away. When his turn came, he reached into his pocket: “Contrary to what my colleagues have told you about the bleak prospects for silicon transistors, I happen to have a few of them here in my pocket” [T3]. Texas Instruments, he announced, had them in production and for sale. It was a sensation [T2] - the more delicious because Tanenbaum’s quieter Bell device, four months older, was unknown to the room.

Texas Instruments dominated the silicon-transistor market for the next few years, cutting deeply into Raytheon’s position as the largest supplier to the open market [T2]. The customer that made silicon indispensable was the one germanium had failed: the military, and above all the missile and aviation programs, where heat, cold, and vibration were facts of life. By the end of the 1950s silicon was the industry’s preferred material [T2], and germanium - the radar veteran, the material of the first transistor, the foundation of the first industry - was on its way to becoming a specialty. The torch had been passed not to a better device but to a better element, and the industry would spend the next seventy years learning how much could be done with it.

# Chapter 11 - Purity as an industry

Every technology has a founding constraint - the thing that must be true before anything else can happen. For aviation it was lift; for the telegraph, insulation; for the transistor, from the first day to this one, it has been purity. The semiconductor effect lives in the balance between a crystal’s own atoms and the deliberate few added to it, and a single unwanted atom in the wrong place can drown the design. When the war began, the purest germanium on Earth was measured in grams and held in a handful of laboratories [H15]. Before the transistor could become an industry, purity itself had to become one.

The first great purification technology came from William Pfann at Bell Labs, and it was beautiful in the way of good physical ideas. Take a rod of germanium in a tube and pass it slowly through a heating coil, so that a short molten zone travels the rod’s length. Impurities, preferring the liquid, ride along in the molten zone and are swept to the far end; repeat the pass again and again, and each sweep carries more contamination to the tip, which is finally sawn off and discarded [T1]. Zone refining, as Pfann called it, drove germanium to impurity levels of one part in ten billion - a pinch of salt, as the Computer History Museum’s account puts it, in three freight cars of sugar [T1]. With material like that, the grown-junction transistor of 1951 became manufacturable, and the germanium industry had its foundation stone.

Silicon mocked the technique. Molten silicon reacts with nearly every crucible material, dissolving its own contamination as fast as the zone could remove it [T1]. The answer, worked out at Bell by Henry Theurer from 1952, was to abolish the crucible: hold a vertical rod clamped at both ends and float the molten zone in place by surface tension, a bead of liquid silicon hanging between solid ends, traveling up the rod with nothing touching it. Boron, the most stubborn contaminant, was stripped by passing steam through the melt, and by early 1955 float-zone silicon had broken below one part per billion [T1]. The same idea occurred independently in the same years to Paul Keck and Marcel Golay at the U.S. Army Signal Corps laboratories at Fort Monmouth, and to Rolf Emeis under Eberhard Spenke at Siemens in Pretzfeld, West Germany [T1] - an early sign that in this industry, the hard problems were few enough that the whole world found them at once.

Float-zone silicon was superb and boutique. To feed an industry, purity had to come from chemistry at chemical scale, and that meant a different route altogether: not moving impurities along a bar, but converting the silicon itself into something that could be distilled. The route begins with metallurgical-grade silicon - quartz rock reduced with carbon in an electric arc furnace, about 98 percent pure, an ordinary commodity of the aluminum and chemical trades [P5]. The crude metal is pulverized and reacted with hydrogen chloride in a fluidized bed, producing trichlorosilane, a

liquid that boils at ordinary industrial temperatures. Liquids can be distilled, and distillation is the chemical industry's oldest and most refined separation art: passed through column after column, the trichlorosilane leaves its boron, phosphorus, and metal contaminants behind. The purified vapor is then decomposed with hydrogen onto heated silicon rods, where pure silicon deposits atom by atom - chemical vapor deposition - growing thick rods of hyperpure polysilicon that are broken into chunks as feedstock for crystal growth [P5]. This Siemens process, steadily modified and improved, remains the backbone of the world's electronic-grade silicon supply [P5].

The numbers of the ladder are worth setting down, because they have no equal in any other industry. Metallurgical silicon: 98 percent - two contaminant atoms in a hundred. Solar-grade silicon: six nines, 99.9999 percent [P5]. Electronic grade runs from nine nines to thirteen nines; at eleven nines - 99.999999999 percent - one impurity atom is permitted among one hundred billion silicon atoms [M3]. The impurity budget for electronic-grade material is specified element by element in parts per billion: boron at a tenth of a part per billion, phosphorus at three-tenths, gold at a hundred-thousandth [P6]. It is routinely described as among the purest materials commonly made by humankind [P6], and the description is fair: the finest reagent chemicals of an ordinary laboratory would be reckoned dirt by a crystal grower.

The last step turns purity into architecture. A chunk of polysilicon feedstock is melted in a silica crucible under argon; a seed crystal, a thin wand of perfect silicon, is dipped into the melt and withdrawn with infinite slowness - a millimeter a minute, slower at the start - while seed and crucible counter-rotate. The melt crystallizes onto the seed, following its atomic pattern, and a single crystal rises from the crucible like liquid freezing upward. This is the Czochralski method, descended from a Polish chemist's 1916 accident with a pen and a crucible of molten metal, adapted for semiconductors by Teal at Bell [H9][M2]. A modern pull begins with a neck - the crystal drawn thin and fast so that dislocations, the crystal's own structural defects, grow out to the surface and die - and then widens to its full diameter and holds it, exactly, for the length of the pull [M2]. About 95 percent of the world's silicon single crystals are grown this way [M2].

The scale these crystals have reached would have seemed deranged to the men of 1950. Where Teal's laboratory boule weighed grams, a modern Czochralski crystal for 300-millimeter wafers runs two meters long and weighs over 250 kilograms, grown from a 300-kilogram melt over three to four and a half days of continuous, instrumented, human-supervised precision [M2]. The crystal is ground to diameter, sliced into wafers less than a millimeter thick, and polished to an atomic-level mirror. Around this need a specialist industry condensed: the polysilicon plants, the puller makers, and the great wafer houses - Shin-Etsu and SUMCO in Japan, Wacker's silicon division in Germany - whose names are unknown to the public and foundational to everything the public owns [M2]. A wafer from one of these houses is

the blank page on which every fab in the world writes. The transistor began as a problem of understanding materials; it became an industry only when a second industry arose to supply them, and that second industry - invisible, energy-hungry, and unimaginably exact - has never once been allowed to fail.

# PART V - THE MANUFACTURE

# Chapter 12 - Printing the crystal

The junction transistor of 1951 was a crystal-grower’s device: its junctions were frozen into a bar as the bar was drawn from the melt, one sandwich at a time, by a man with a crucible and a clock. It was precise the way violin-making is precise. Between 1952 and 1957, Bell Labs replaced that art with a different one, borrowed from older crafts - baking, painting, and printing - and in doing so invented the grammar in which every chip since has been written.

The first element was the furnace. In 1952 Calvin Fuller had shown that donor and acceptor atoms could be introduced into a semiconductor in a controlled way by surrounding it, at high temperature, with a vapor carrying the dopant: the atoms diffuse into the crystal’s surface, and the temperature sets the rate, and therefore the depth, with precision no crucible could match [M5]. In 1954 Charles Lee used the method to build the first diffused-base germanium transistor, with a cutoff frequency of 500 megahertz - ten times the best alloy transistors of the day [M5]. By 1955 Morris Tanenbaum and D. E. Thomas had produced an all-diffused silicon transistor, and the announcement, in the words of a later Bell Labs history, created great excitement throughout the infant industry [M5]. The wafer had replaced the crucible. Now the question was control: diffusion doped everything it touched, and a transistor needs its impurities placed.

The second element was discovered, like so much in this story, by accident. Early in 1955, Carl Frosch at Bell Labs was fighting a pitting problem: at diffusion temperatures above 1,000 degrees Celsius, the carrier gases eroded the silicon surface. Then the hydrogen gas carrying impurities through his furnace briefly caught fire. The accident flooded the chamber with water vapor, and when the wafers came out they were coated in a thin, glassy skin of silicon dioxide - and unharmed [M4]. Most accidents get cleaned up and forgotten. Frosch, working with his technician Lincoln Derick, spent the following months learning what the accident had made. The oxide film, they found, protected the surface perfectly during the high-temperature steps. More: it admitted some dopant atoms and barred others - boron and phosphorus, the workhorse dopants, could not pass [M4]. A silicon wafer wearing its own oxide was a wafer that could be doped selectively: cut a window in the glass, and only the exposed silicon receives the impurity. They published and patented the technique in 1957 [M4]. Conventional practice then required stripping the oxide off afterward; no one yet imagined leaving it on.

Frosch and Derick showed what the window meant for device structure: they built a transistor by double diffusion, both emitter and base contacts made at the surface in adjacent areas [M4] - a flat, two-sided device on a one-sided wafer, something no crucible-grown crystal could offer. And there was a quieter gift folded into the same

skin of glass. Bardeen’s surface states - the dangling bonds that had screened off every electric field and strangled the field-effect idea in 1945 - were tamed by the thermally grown oxide, as Mohamed Atalla and his colleagues at Bell would establish by the decade’s end: the oxide satisfied the surface’s dangling chemistry and quieted it [M5]. The accident that solved pitting had, in passing, reopened the door Shockley had walked through in failure fifteen years before.

The third element came from the printing trade. Windows in the oxide had to be cut somewhere precise, and in 1955 Jules Andrus and Walter Bond at Bell adapted photoengraving - the printer’s art of etching metal plates through light-hardened resists - to the purpose [M5]. Coat the oxide with a light-sensitive lacquer, expose it through a photographic mask, wash away the unhardened lacquer, and the oxide beneath can be etched open exactly where the mask decreed, with draftsman’s precision at microscopic scale. Photography had been miniaturizing images for a century; now it was pointed downward, at crystals. The mask, the resist, the exposure, the etch: the words are a printer’s, and they remain the industry’s words today.

One more property of the loop decided the industry’s economics, and it deserves to be stated plainly because everything later in this book follows from it. A photographic mask does not care how many times its pattern repeats. The same exposure that prints one window prints ten thousand, and the same furnace run that diffuses one wafer diffuses a boat of wafers. Printing is a batch art: the cost of the pattern is paid once, in the mask, and then divided among every device on every wafer the mask ever prints. The crucible had built transistors one at a time. The loop would build them by the million, at a marginal cost per device that fell toward the cost of the raw silicon itself. The industry’s famous deflation - the ninety-nine percent price collapse, the doubling curves, the disposable electronics of the later chapters - is all written into that one property of light.

Put the three together - diffuse, mask, print - and a new manufacturing logic appears. A pattern is designed once, at drawing-board scale. Light reduces it onto a wafer. Chemistry carves it. The furnace fills it. Repeat with a second pattern, aligned on the first, and structures of stacked n and p regions accumulate, all defined by drawings. Nothing in the loop depends on a craftsman’s hand except the drawings themselves. In 1955 and 1956 the loop’s pieces sat on benches at Bell Labs as separate marvels: Fuller’s diffusion, Frosch and Derick’s oxide, Andrus and Bond’s photoengraving [M5]. What they would become when assembled - and they would be assembled not at Bell but at a brash young firm in California - is the story of the next chapter: the process that made the transistor printable, and with it, made the transistor cheap.

# Chapter 13 - The planar process

The young company that assembled Bell's printing loop was Fairchild Semiconductor. Its founding is a drama of its own - William Shockley, freshly a Nobel laureate, leaving Bell Labs in 1955 to start Shockley Semiconductor in Palo Alto [M5]; the brilliant young staff he assembled; the abrasive management that drove eight of them, led by Robert Noyce, to resign together in 1957 and found Fairchild under the wing of an East Coast camera-and-instruments firm [M17][M5]. That story - the men, the betrayal, and the valley it started - gets its own chapter later in this book. What matters here is the company they built in 1957 and what it did first. They licensed Bell's techniques, bet on silicon and on diffusion, and made their name with the era's champion device: the double-diffused silicon mesa transistor. The mesa was built by diffusing base and emitter into the wafer's face, masking the precious active region with wax, and etching the surrounding silicon away with acid, leaving a flat-topped plateau - a mesa, in the geographer's sense - with the junctions inside it [M5]. The devices were fast and rugged, exactly what the aerospace and missile programs were buying; Fairchild's 2N1132 flew in the Minuteman program, and its gold-doped 2N706 of 1959 switched in sixteen nanoseconds, three times smaller than its predecessor [M9]. An order from IBM for diffused mesa transistors put the young firm into computers, and within two years of its founding Fairchild was profitable, with revenues over half a million dollars [M5].

And the mesa had a flaw that no one could fix. Etching the plateau exposed the junctions at the surface, naked to the world. Dust settled on them. Process residues clung to them. A fingerprint's oils drifted across them. The exposed junction was a standing invitation to contamination, and contamination meant devices that failed in the factory as yield losses and devices that failed in the field as reliability disasters [M5]. The industry's best transistor was an open wound. Conventional wisdom held that the oxide used to mask the diffusions had itself to be stripped away when the work was done, since no one could vouch for what it might contain [M6]. The junctions had to be bare. Everyone knew that.

Jean Hoerni did not believe it. A Swiss-born physicist among the eight founders, Hoerni had written an idea into his notebook in December 1957: leave the oxide in place. Let the glass skin that had masked the diffusions stay on the wafer forever, a permanent protective coat with the junctions sealed beneath it, and build the device so that everything - diffusions, contacts, all of it - happened from the one protected face. The result would be flat, so he called it planar [M6]. Fairchild, straining to get its first products into production, had no time for the idea in 1957. On January 14, 1959, Hoerni wrote it up as a patent disclosure - a document the historian Christophe Lécuyer has called the most important innovation in twentieth-century technology - and on March 12, 1959, working with Jay Last, he demonstrated the first planar

transistor to his colleagues [M9]. The demonstration had a showman's touch: to prove the oxide's protection, Hoerni spit on the unpackaged die, and the device worked on undisturbed [M9].

The planar transistor was not merely cleaner than the mesa; it was better. Leakage currents - the parameter that decides whether computer logic can tell off from on - fell dramatically [M6]. The flat face invited a new kind of contact, evaporated metal running across the oxide to the windows. And because every step now happened from one side of the wafer, through masks, under glass, the printing loop of the last chapter could be repeated as many times as a design required: oxide, pattern, diffuse, oxide, pattern, diffuse, layer upon aligned layer, a city built by successive printings [M5] [M6]. Fairchild put the 2N1613 planar transistor on the market in April 1960 and licensed the process across the industry [M6]. Within a few years the mesa was extinct, and the planar process - mutated, miniaturized, and elaborated beyond all recognition - remains the basic method of making every integrated circuit on Earth [M7].

It is worth being precise about what Hoerni had done, because it is the pivot on which this book turns. The transistor of 1947 proved that a crystal could amplify. The junction transistor of 1948 proved it could be designed. Silicon proved it could endure. Zone refining and the Siemens chemistry proved it could be fed. Diffusion, oxide masking, and photolithography proved it could be printed. Hoerni's oxide left-on proved that what was printed could be sealed - and sealing the surface turned out to be the last precondition for the step that made everything else matter. For if a wafer's face is flat, protected, patterned, and metallized, then there is no reason to print one transistor on it. You can print two, and connect them. You can print a circuit. In the same winter that Hoerni wrote his disclosure, his colleague Robert Noyce was writing one of his own [M9]. That is the next chapter.

# Chapter 14 - Many on one slice

By the late 1950s the transistor had a strange new problem: it worked too well. Single devices were small, fast, and reliable, and engineers could now design systems - computers, missile guidance, telephone switches - that called for thousands and tens of thousands of them. But a circuit is not its components; it is its components plus their connections, and the connections were still made by hand, one soldered wire at a time. Jack Morton, a Bell Labs vice president, gave the bottleneck its enduring name in 1958: the tyranny of numbers. “Such systems,” he wrote, “require hundreds, thousands, and sometimes tens of thousands of electron devices. Each element must be made, tested, packed, shipped, unpacked, retested, and interconnected one-at-a-time to produce a whole system” [M12]. Every wire was a cost, and every solder joint was a future failure. The industry was building cathedrals out of bricks that each had to be carried by hand.

The answers arrived, as answers in this industry tended to, twice at once.

The first came from a man alone in a laboratory. Jack Kilby had joined Texas Instruments in May 1958 and had asked to work on microminiaturization - the quest to shrink military electronics - but he quickly judged the company’s chosen approach, an Army-backed scheme of standardized plug-together modules, to be a dead end that did nothing about the interconnections themselves [M13][M11]. When TI emptied for its summer shutdown in July, Kilby, too new to have earned vacation, stayed behind with the problem. His solution reversed the industry’s instincts: instead of making components from different materials and wiring them together, make everything - transistors, resistors, capacitors, and the base itself - from a single piece of semiconductor [M10]. A monolith: one stone. On September 12, 1958, he demonstrated a working circuit of this kind to TI’s executives: a sliver of germanium carrying all the elements of an oscillator, joined by a web of fine gold wires that rose above the chip like guy-wires on a tent - the famous “flying wires” [M11]. It worked. Kilby, who knew what it looked like, remarked later that if he had known he would be showing it for the rest of his life, he would have “prettied it up a little” [M11]. His patent application, filed on February 6, 1959, described “a body of semiconductor material ... wherein all the components of the electronic circuit are completely integrated” [M13]. TI announced the “Solid Circuit” at the IRE show that March [M9].

The flying wires were the point and the problem. Kilby had proven the monolithic idea, but wires threaded by hand above the chip could never be a production process; his own company described the approach as impractical for mass production [M11]. Six months behind and half a continent away, Robert Noyce at Fairchild completed the idea. Prodded by a patent attorney to find further uses for Hoerni’s planar process,

Noyce wrote up, in late January 1959, a different route to the same destination: build the components into a silicon wafer by the planar method, and then print the wiring too - aluminum lines evaporated onto the protective oxide, connecting device to device through windows in the glass, the whole circuit defined by masks and made in the same batch process as the transistors themselves [M7]. No flying wires; no handwork; the interconnections were printed like everything else. Noyce filed his patent on July 30, 1959 [M7].

Making it work took Fairchild another year and a team. Under Jay Last, a group including Sam Fok, Isy Haas, Lionel Kattner, and James Nall built the first working monolithic circuits on May 26, 1960, isolating the components from one another by etching channels and filling them with epoxy; on September 27, 1960, they produced circuits with the cleaner method that became standard, electrical isolation by reverse-biased junctions [M8]. Fairchild announced its Micrologic family of integrated circuits in March 1961, with a photograph in LIFE magazine; Texas Instruments, which had adopted the planar process, announced its own Series 51 integrated circuits that October [M8]. The two companies sued each other over the patents for years, and eventually settled on cross-licenses; the courts would later side with Noyce, but by then the industry had long since made its own judgment [M7]. History credits both men, and correctly: Kilby for the first working circuit made of one piece of semiconductor, Noyce for the interconnection scheme that made such circuits manufacturable [M7]. When the Nobel Prize in Physics honored the integrated circuit in 2000, it went to Kilby alone, because Noyce had died in 1990 and the prize is not given posthumously; many believe Noyce would have stood beside him had he lived [M7].

The integrated circuit did not immediately look like a revolution. It was expensive, it was small in capability, and its first customers were, as ever, the military and the space program, who bought reliability and lightness at prices no civilian would pay. But the direction was set. Every discrete-transistor circuit was now, in principle, a candidate to be printed whole on one slice - and printing, as the industry was learning, is a process that only gets cheaper. The tyranny of numbers had met the batch, and the batch won.

# Chapter 15 - The MOS turn

There is an irony at the heart of this book, and it is time to acknowledge it. The transistor that won - the one that fills the world, the one inside the device on which you may be reading these words - is not the transistor Bardeen and Brattain demonstrated in December 1947, nor the junction transistor Shockley designed in 1948. It is the transistor Shockley failed to build in 1945: the field-effect device, in which a voltage on a gate, insulated from the crystal, controls the current through it [M14]. The idea was the oldest of them all - Lilienfeld had patented a version in 1926 [H2] - and it had waited thirty-five years, blocked by a single enemy: the surface states, the crystal's chemically ragged skin, which drank the applied field and left nothing to control the current beneath.

The man who tamed the skin was Mohamed "John" Atalla, an Egyptian-born engineer with a Purdue doctorate who joined Bell Labs and took up the least glamorous subject in the building: the silicon surface itself [M14]. Studying thermally grown silicon dioxide - the same accidental glass of Frosch and Derick - Atalla and his colleagues showed in 1959 that the oxide did more than protect: at the interface between silicon and its oxide, the surface states were markedly reduced, the dangling bonds chemically satisfied, the skin quieted [M14]. Then Atalla gave his new colleague Dawon Kahng, a Korean-born engineer fresh from his doctorate, the assignment: use the quiet surface to build the field-effect transistor. In 1960 they did. A metal gate on an oxide film on silicon - metal, oxide, semiconductor, MOS - and the field from the gate at last reached into the crystal and governed it [M14]. They announced the device at the IRE Solid-State Device Research Conference that year [M14].

Bell Labs yawned. The MOS transistor was slow, and the telephone system had no need of it; the laboratory that had invented the transistor declined to pursue its successor [M14]. Kahng saw what his managers did not. In a January 1961 internal memo he pointed out the device's "ease of fabrication and the possibility of application in integrated circuits" [M14] - one of the great quietly correct sentences in the history of technology. Others saw it too. At RCA, Karl Zaininger and Charles Meuller had built an MOS transistor in 1960, and Fred Heiman and Steven Hofstein followed with a sixteen-transistor MOS integrated circuit in 1962; at Fairchild, C. T. Sah built an MOS-controlled device [M14]. The first commercial MOS transistors reached the market in 1964 [M14].

The early MOS years were beset by a plague with a chemistry: the devices drifted, their thresholds wandering until circuits misbehaved. In October 1963, Ed Snow at Fairchild identified the culprit - sodium, common salt's metal, migrating through the oxide - and once the contaminant had a name it could be excluded, and reliable MOS became possible [M16]. The deeper advantage, though, was structural. A bipolar

junction transistor is a layer cake of alternating regions with current driven through its heart; an MOS transistor is simpler, smaller, and needs no input current to hold its state - a voltage on an insulated gate suffices. Smaller and lower-power: in an industry about to print millions of devices per chip, those were the two virtues that would decide everything. Over 99 percent of the microchips produced today are built from MOS transistors [M14].

The final refinement came almost immediately, from Frank Wanlass at Fairchild. Wanlass noticed that n-channel and its mirror - the p-channel device - could be paired: a complementary circuit in which one transistor of each polarity stood in each logic gate, arranged so that in either of the gate's two steady states, one of the pair is off. Current flows only while the gate is switching. In February 1963 Wanlass and Sah presented the scheme at the International Solid-State Circuits Conference; Wanlass filed his patent that June [M15]. Complementary MOS - CMOS - drew close to zero power at standby, and its early champions were exactly the customers who cared: RCA built CMOS logic and memory for an Air Force computer in 1965, and the configuration's first high-volume home was the digital watch, where a battery must last a year [M15]. Speed was not its early strength, and the fast machines stayed with power-hungrier technologies - until, as the following chapters describe, the shrinking of the transistor changed the balance and CMOS became, as it remains, the nearly universal fabric of logic [M15]. The field-effect idea had taken thirty-five years to arrive. It has not been displaced since.

# Chapter 16 - Printing smaller

The planar process made the transistor printable; the next half-century made the print finer. That refinement has a machinery of its own, and its history is the least known and most consequential in the industry: the story of the camera that shrinks. Every other advance in this book rides on it, because nearly every improvement in chips - speed, cost, memory, the microprocessor itself - has been, at bottom, an improvement in how small a pattern of light can draw.

The first cameras were not cameras at all. Through the 1960s the workhorse was the contact aligner: the mask, a glass plate bearing the pattern at chip scale, was pressed directly against the resist-coated wafer and flooded with ultraviolet light [L2]. Contact printing is honest, ancient technology - the same principle as a photographic contact print - and it has the vices of intimacy. Every touch risks scratching the resist, embedding dust, and wearing the mask; a ruined pattern means a ruined wafer, and the defects multiplied exactly as volumes grew [L1]. Kulicke & Soffa sold the first commercial contact aligners in 1965 and for a time held nearly the whole of the small market [L2]. The industry needed to stop touching.

The first escape was projection. In 1973 Perkin-Elmer introduced the Micralign, born of an Air Force research contract: a 1:1 scanner that threw the mask’s image onto the wafer through an elegant all-mirror optical system, with nothing touching anything [L2]. The effect on defect rates was dramatic - projection printing, combined with the newer positive photoresists, cut them sharply - and the Micralign made Perkin-Elmer the largest equipment firm in the industry [L2]. But the 1:1 approach carried a hidden ceiling. The mask had to be the same size as the wafer’s full pattern, and wafers were growing: from 50 to 75 millimeters and on toward six inches, dragging mask size, mask defects, and above all overlay - the alignment of each new layer upon the ones beneath, which must hold within a fifth of the smallest feature - toward impossibility [L1].

The second escape was reduction, and it changed everything. GCA, a company that had come to the industry through map-making and mask-making machines, adapted its photorepeater into a new kind of tool: the wafer stepper. Instead of printing the whole wafer at once from a mask its own size, the stepper printed one chip’s pattern from a reticle ten times larger than the final image, reduced through a precision Zeiss lens onto one patch of the wafer - and then stepped to the next patch, and the next, until the wafer was covered [L1]. The GCA DSW 4800 of 1978, the first successful stepper, worked at the mercury lamp’s g-line, 436 nanometers of violet light, with automated wafer handling and alignment [L1]. Reduction freed the mask-maker from wafer scale, freed the lens from covering everything at once, and let each exposure be individually aligned - but at a price in speed, since the wafer now took dozens of

exposures instead of one, and for years the stepper could not match the Micralign's throughput [L2][L3]. What decided the contest was memory: the 256-kilobit DRAM generation, with its relentless demand for finer lines and tighter overlay, made the stepper the mainstream tool, and the industry's trajectory was set [L2]. Nikon, prototyping its own stepper in 1978 under Japan's national VLSI program, and later Canon and a small Dutch firm called ASM Lithography, would carry the reduction principle forward [L2].

Notice what the stepper's logic implies. Resolution in optical printing is governed by the light itself: finer features demand shorter wavelengths, wider lenses, or tricks of interference and chemistry. The stepper architecture was a chassis onto which each of those could be bolted, generation after generation - shorter-wavelength lamps, then excimer lasers in the deep ultraviolet, then immersion fluids between lens and wafer, and finally, decades later, the extreme-ultraviolet machines that deserve their own chapter. The stepper also created, almost incidentally, the industry's division of labor: a specialized equipment industry, selling precision machines to any chip maker, so that manufacturing brilliance no longer had to be homegrown. From the 1980s on, the question "who can print the smallest transistor?" would be answered less often by the chip companies than by a handful of toolmakers in Veldhoven, Tokyo, and Silicon Valley. The camera had become an industry of its own - and, in time, the chokepoint of the whole world economy. But that is a later chapter; first, the people and the companies who built the industry on everything described so far.

# PART VI - THE PEOPLE AND COMPANIES

# Chapter 17 - Shockley Semiconductor and the eight

William Shockley left Bell Labs in 1955 to build something of his own, and he built it in California, the country of his upbringing - trading the security of the world's finest laboratory for the chance to be not a department head but a founder. In September 1955 he agreed with Arnold Beckman - the instrumentation industrialist whose company was prospering in Southern California - to establish the Shockley Semiconductor Laboratory as a Beckman Instruments division [M17]. He rented a building at 391 South San Antonio Road in Mountain View, an address then surrounded by orchards, and set out to recruit, in his own words, "the most creative team in the world for developing and producing transistors" [M17]. It is the great, bitter irony of this story that he succeeded.

The team he assembled was extraordinary: Gordon Moore, Robert Noyce, Jean Hoerni, Jay Last, Eugene Kleiner, Julius Blank, Victor Grinich, Sheldon Roberts - the eight names that would matter most - plus dozens of others who learned silicon and diffusion at his benches [M17]. In December 1956, a year into the venture, Shockley went to Stockholm to share the Nobel Prize in Physics with Bardeen and Brattain [M17]. He was forty-six, famous, richly backed, surrounded by talent, and located on the future's doorstep. Within a year of the Nobel it was rubble.

The failure was managerial, and it was total. Shockley fixated on a difficult four-layer diode he had conceived at Bell for telephone switching - a distant, demanding device - while his staff argued the company should pursue the immediate silicon-transistor market opening in front of it [M17]. He set up a separate, secret research group on his pet project, cutting his own engineers out of the loop [M18]. He was, by near-universal testimony of those who were there, suspicious and abrasive; the employees eventually went over his head and asked Beckman to appoint a different manager, and Beckman declined [M17]. At a May 1957 meeting, stung by Beckman's own proposals to control expenses, Shockley stood up and stormed out, shouting that if they did not like what he was doing, he could take the group and find support anywhere [M18]. Nobody followed him out of the room.

That summer, seven of the senior staff met with Alfred "Bud" Coyle and Arthur Rock, two young bankers at the New York firm of Hayden, Stone, about finding jobs elsewhere as a group. Coyle and Rock offered a different thought: don't find employers - find a backer, and found a company [M9]. Coyle reached Sherman Fairchild, the founder of Fairchild Camera and Instrument on Long Island, and in September the deal was struck: $1.38 million from Fairchild Camera to establish a new semiconductor company run by the eight [M18]. On September 18, 1957, they resigned - Shockley's notebook records it in four words, "Wed 18 Sep - Group resigns" [M18] - and on September 19 they signed the papers that created the

Fairchild Semiconductor Corporation [M19]. They are remembered, with the affection of an industry that loves its founding myths, as the traitorous eight.

They started in founders' homes, building probe stations and workbenches in Victor Grinich's garage, before leasing a 14,000-square-foot building at 844 Charleston Road in Palo Alto - later designated California Historical Landmark number 1000 [M9]. And they carried contraband more valuable than any equipment: while still at Shockley's, several of them had read the preprint of Frosch and Derick's oxide-masking paper and had begun experimenting with double diffusion and oxide work themselves [M18]. When the paper was published in September 1957, the Fairchild founders had a head start on nearly every competitor on Earth [M18]. Within a year they would ship the first commercially successful silicon mesa transistors, to IBM, for the B-70 bomber's guidance computer, and end 1958 - their first full year - in the black on sales over half a million dollars [M18].

Shockley Semiconductor never turned a profit. Beckman sold it to the Clevite Corporation in 1960, and Shockley himself went to Stanford as a professor [M17]. His later years brought notoriety of a different kind, far from semiconductors, and they belong to a sadder and separate story. What he built accidentally and lost quickly was the more consequential legacy: the eight, and then the dozens who followed them out the door, seeded not one company but a culture - brilliant staff, restless loyalty, and the conviction that the natural destiny of a good engineer is a startup. The walkout of September 1957 became the valley's creation myth and its operating procedure. Everything in the next five chapters descends from it.

# Chapter 18 - Fairchild

For one decade, Fairchild Semiconductor was the center of the semiconductor world, and the argument of this chapter is that it remains, in a real sense, the center: not because of what it sells - it long ago stopped mattering as a company - but because of what it started. The integrated circuit, the planar process, the venture-funded silicon startup, the industry's sales culture, and a genealogy of founders that includes Intel and AMD all trace to one leased building on Charleston Road [M9].

The decade began as the last chapter described: mesa transistors for IBM's B-70 program, profitability in the first full year, sales over half a million dollars in 1958 [M18]. Then came the stack of inventions already told - Hoerni's planar process in 1959, Noyce's monolithic circuit the same year, Last's Micrologic family in 1961 - and with them the market. The military and space programs of the 1960s paid premium prices for exactly what Fairchild sold: reliability in silicon. By the middle of the decade the division was one of the world's largest producers of silicon transistors and controlled more than 30 percent of the market for integrated circuits [M20]. In 1965 its research director, Gordon Moore, published the observation that would organize the industry's next half-century, of which much more later. Fairchild did not merely ride the transistor's golden age; in large part, Fairchild was the golden age.

The structure that made it great also unmade it. Fairchild Semiconductor was a division of Fairchild Camera and Instrument, headquartered in Syosset, New York, three time zones and a world of management culture away. The California division generated the profits - by 1967 more than half of corporate sales - while decisions about its investment, its capacity, and crucially its stock options were made by men in New York who visited once a year [M20]. In an industry where a good process engineer could walk across the street and double his prospects, Fairchild could not grant equity to keep its own [M20]. The attrition began early - Rheem Semiconductor took a general manager and eight senior men as early as 1959 [M9] - and it never stopped. Each departure carried a piece of the company's capability and a piece of its culture to a competitor that, being new, could offer the one thing Fairchild could not: a stake.

The reckoning came in 1967. An industry downturn, manufacturing difficulties, and market share lost to Texas Instruments' SN7400 TTL logic family combined to produce the division's first loss since its founding year [M20]. In March of that year Charlie Sporck, the manufacturing chief, resigned to run National Semiconductor, taking senior managers with him and, in the Chronicles' words, seriously damaging Fairchild's manufacturing capability [M20][M22]. Even in the bad year there were triumphs - the Univac order that September, three million custom circuits worth about $7 million, the largest the mainframe maker had ever placed [M20] - but the board

fired the corporate CEO, and then, in February 1968, removed his successor too. Noyce, passed over for the top job at the company his division was carrying, made his decision. He resigned on June 25, 1968, writing that he wanted to “get close to advanced technology again” [M20]. Moore followed on July 3; on July 18 the two incorporated a new company, soon renamed Intel, with money raised by Arthur Rock and with five of the eight Fairchild founders among its investors [M20]. Jerry Sanders, the brash marketing director, was pushed out by the new regime and in May 1969 co-founded Advanced Micro Devices [M20]. A team of Motorola executives under C. Lester Hogan arrived that August to run what was left [M20].

What was left was still formidable - the μA741 op-amp of 1968 became the most popular operational amplifier ever made, and Fairchild built the industry’s first standard MOS-LSI product and its first commercial silicon-gate MOS circuit in the same year [M20] - but the company never regained its position. Fairchild Camera sold it in 1979 to Schlumberger, the French oilfield-services giant, for $425 million; eight difficult years later, Schlumberger sold it to National Semiconductor - by then run, inevitably, by the Fairchild veteran Charlie Sporck - for $122 million, about half of what Fujitsu had offered before the U.S. government objected on national-security grounds [M21]. That intervention, in 1987, is worth a bookmark: it is an early glimpse of the geopolitics of Part X, in which chips stop being merchandise and become strategic territory.

Fairchild’s epitaph was written many times, most memorably by its alumni, who called themselves Fairchildren and filled the industry’s executive suites. But the better measure is a map. Draw the founding tree of Silicon Valley’s semiconductor companies - who left what to start which - and nearly every line passes through one node. The traitorous eight had walked out of Shockley in a single afternoon. Fairchild turned that act into a system: talent in, companies out, decade after decade. It was less a company, in the end, than a school that paid its teachers in stock it could not grant - and so the teachers left and founded the industry instead.

# Chapter 19 - The Texas way

If Fairchild is the industry's romantic legend, Texas Instruments is its argument that the legend is optional. TI had no walkout, no traitorous eight, no orchard myth. It had something rarer: a deliberate, decades-long corporate strategy, made by identifiable managers, that turned an oil-prospecting service company into the industry's transistor leader within a decade and its dominant logic supplier for two more. The transistor industry has two creation stories, and the Texas one is about choice rather than accident.

The company's origins could hardly be further from electronics. It began as a geophysical enterprise - Karcher and McDermott's seismic exploration firm, founded in Oklahoma in 1924 and reestablished in Dallas in 1930 as Geophysical Service Incorporated, selling reflection-seismograph surveys to oil companies [I5]. The war converted it: the same echo-reading art that found oil could find submarines, and GSI began manufacturing military electronics [I5]. In 1945 the company hired a young Navy procurement officer named Patrick Haggerty to run its laboratory and manufacturing division, and Haggerty arrived convinced that the firm's future was electronic [I5]. By 1951 the electronics business had so outgrown the geophysical one that the company renamed itself Texas Instruments [I5].

Haggerty's defining trait was impatience with other people's components. When Western Electric offered transistor licenses in 1951 at $25,000, Haggerty bought one in 1952 - a bet that the device was real before there was a market to prove it [I4]. Within months TI had its first transistor; by late 1953 it was mass-producing germanium devices under Mark Shepherd [I5]. Then came the two strokes already told in this book: hiring Gordon Teal away from Bell Labs, and through him the silicon transistor of 1954, which TI had in production while the rest of the industry was still dismissing the material [T2][T3]. Silicon gave TI the military's heat and altitude business, and by 1956 the company led the industry in transistor sales, with total revenues of $45 million against $3 million a decade earlier [I5].

But Haggerty's most instructive move was a radio. TI in 1954 had superb silicon transistors and almost no consumer market for any transistor at all. So on May 21, 1954, Haggerty ordered a research group to bring him a working transistor radio within a week. They returned in four days with a breadboard of eight transistors, and Haggerty committed $2 million - nearly a tenth of the company's revenue - to making it a product [I4]. Partnering with a small Indianapolis firm, IDEA, which would sell it under its Regency brand, TI drove the design down to four germanium transistors and announced the Regency TR-1 on October 18, 1954: the first commercial transistor radio, pocket-sized, in stores in New York and Los Angeles on November 1 [I4]. It sold 100,000 units in its first year. It was never very profitable, and within a few years

both companies had left the radio business [I4]. The radio was never the point. It was a demand pump - a proof to every radio maker in the world that transistors were ready, and an education for TI in high-volume manufacture. Haggerty estimated it pulled the consumer transistor market forward by two years [I4]. The device itself mattered less than the lesson: in this industry, you sometimes have to build the market before you can sell into it.

The same discipline made TI the king of standard logic. The industry standard of the 1960s and 1970s - the 7400 series of transistor-transistor logic, the NAND gates and flip-flops from which a generation of computers was assembled - was not a TI invention. TTL was patented by James Buie at Pacific Semiconductor in 1961, and the first commercial family was Sylvania's SUHL in 1963 [I6]. TI entered in 1964 with a military-grade family, the SN5400, and then made the decisive move: in 1966 it issued the SN7400 series in cheap plastic dual-inline packages for industrial customers [I6][I7]. Cheap, standardized, available, second-sourced - within a few years TI held over half of the world logic market, and when Fairchild pioneered more complex TTL functions, TI simply copied them into its own numbering and out-distributed the originator [I7]. The 7400 series became the industry's common parts catalog - TI made nearly a hundred different devices in the family - and remained the way computers were built until CMOS integration made whole systems fit on single chips [I6]. TI stayed the dominant and most profitable logic supplier into the mid-1970s, when the worldwide TTL market reached three-quarters of a billion dollars [I7].

Kilby's integrated circuit of 1958 had already guaranteed TI a place in this book; the silicon transistor and the pocket radio and the 7400 series gave it a place in every factory. The pattern across all four is the Texas way: license early, staff deliberately, industrialize ruthlessly, and when the market does not exist, build a radio. The valley got the glory. Dallas got the volume - and for twenty years, the two of them divided the transistor world between them.

# Chapter 20 - Intel and memory

The company Noyce and Moore founded in July 1968 was a bet on a specific technical judgment: that semiconductor memory - data stored in transistors rather than in the magnetic cores that then filled computer memory racks - was about to become practical, and that whoever industrialized it first would own a market measured in billions of bits [M20]. Arthur Rock raised the money; five of the other Fairchild founders invested; Andrew Grove, Fairchild's MOS chemist, joined early and eventually ran the place [M20][M16]. The plan was memory. What actually happened was memory, and then an accident that ate the world.

The enabling technology came from the last refinement of the MOS transistor: the silicon gate. At Bell Labs in 1967, Kerwin, Klein, and Sarace had replaced the transistor's metal gate electrode with polycrystalline silicon, a change that aligned the gate to its own channel automatically and made MOS devices faster, denser, and more reliable [M25]. Federico Faggin, working with Tom Klein at Fairchild, turned the idea into a product - the 3708 of 1968, the first commercial silicon-gate integrated circuit [M25] - and Intel seized on the process as the basis of its memory business: three to five times the speed in half the chip area of conventional MOS [M25]. Intel's first MOS product, the 1101 256-bit static memory, appeared in 1969 [M25].

The product that made the company was the 1103, and its target was one of computing's oldest components. Main memory in 1970 was magnetic core: millions of tiny ferrite rings, each threaded by hand on wires, each storing one bit. The 1103 packed 1,024 bits on one chip - a three-transistor cell idea, proposed by Bill Regitz at Honeywell, implemented and improved at Intel by Joel Karp, Ted Hoff, Bob Abbott, and Bob Reed - and sold it at one cent per bit [M23]. At that price, semiconductor memory crossed core's cost curve, and beginning in 1970 the 1103 swept the field; IBM committed to the new technology in its System/370 line in 1972 [M23]. The one-transistor cell patented by IBM's Robert Dennard, adopted at the sixteen-kilobit generation, carried DRAM to 64 kilobits from American and Japanese makers before the decade closed [M23]. Memory became the industry's volume engine - the product whose brutal price curve paid for each new generation of factories. It also became, in the 1980s, the battlefield on which the Japanese challenge was fought, which is a later chapter.

The accident began with a calculator. In 1969 the Japanese calculator maker Busicom asked Intel for a set of custom chips, and Ted Hoff, with Stanley Mazor, proposed instead something audaciously general: a small computer's entire central processing unit on a single chip, with the calculator's particular behavior supplied by a stored program rather than by fixed wiring [M24]. Faggin, who joined Intel in 1970, brought the silicon-gate mastery - and, with Busicom's Masatoshi Shima, squeezed the 2,300

transistors of the 4004 into a sixteen-pin package in 1971 [M24]. The 4004 was modest: four bits at a time, a calculator’s brain. But its principle was unlimited. The 8008 followed in 1972, and in 1974 the 8080 and Motorola’s 6800 established the microprocessor as a component any engineer could buy [M24]; the 6502 powered the first personal computers; and in 1981 the IBM PC would carry Intel’s 8088 into every office in the world [M24]. A computer’s central processor - the most complex artifact of the vacuum-tube age, the thing ENIAC had filled a room to be - had become a catalog item, printed by the million.

It is worth pausing on what Intel’s founding decade proves about the transistor’s logic. Memory and microprocessor were both consequences of the same property: printing scales. The batch economics of the planar process meant that anything that could be reduced to a repeated pattern of transistors would collapse in cost and explode in volume - memory first, because it is the most repetitive structure in computing, and then logic itself, reduced to a general-purpose pattern that software could specialize. Intel did not invent the integrated circuit. It was the first company built from the start on the assumption that the circuit would integrate everything - and it very nearly did.

# Chapter 21 - Japan learns the transistor

The transistor’s conquest of the world did not begin in the countries that invented it. It began in a country rebuilding from defeat, at a company its own government considered unqualified, making a product the inventors had not thought to make. This is the chapter where the transistor stops being an American story, and where its modern meaning - a consumer device, a global industry - first comes into view.

Japan in 1952 was a nation under reconstruction, its electronics industry a collection of tube makers supplying a domestic market in which a radio was still a piece of furniture the family gathered around [I10]. That year, Masaru Ibuka, the co-founder of a small Tokyo company called Tokyo Tsushin Kogyo - Totsuko - traveled to the United States and learned that Western Electric was licensing the transistor patents [I8]. Ibuka’s company made tape recorders; it had no semiconductor experience at all. What it had was a problem: success. The tape recorder had swelled Totsuko with talented engineers, and Ibuka needed a next mountain for them to climb [I9]. He decided the mountain would be the transistor, and the product would be a radio small enough to carry - a decision American engineers tried to talk him out of, explaining that transistors were suitable for hearing aids, not for the radio-frequency work a receiver demands [I9].

The obstacles were governmental as much as technical. The Ministry of International Trade and Industry initially refused permission: transistors were too advanced for a shabby little factory, and the big firms - Toshiba, Mitsubishi, Hitachi - were pursuing them through official technical-aid agreements [I8]. Western Electric, meanwhile, had been persuaded the other way by Totsuko’s persistent New York contact and by the evidence that this small company had developed a tape recorder entirely on its own [I8]. Akio Morita, Ibuka’s co-founder, traveled to America in August 1953 - his first trip abroad - to close the contract at the same $25,000 the other licensees paid, contingent on MITI’s approval, which finally came at the end of 1953 [I9][I8]. Totsuko was the first Japanese company to hold a Western Electric transistor license [I9].

What followed remains one of the great technology sprints. Kazuo Iwama, who had led the tape-recorder research, toured American transistor facilities and shipped home intelligence; the team exploited the published Bell literature - the same open textbooks and the zone-refining technique Bell had made public in 1952 - as a shortcut past problems that had stalled Japan’s best academic laboratories for years [I9]. By April 1954, six months after the contract, Totsuko was trial-manufacturing both point-contact and junction transistors [I9]. Then came the product education. The first radio, the TR-52 of early 1955 - the first article to carry the SONY logo - never reached the market: its stylish white grille warped away from the cabinet in the May heat, a

materials failure the company answered with a crash program in plastics [I8][I9]. Its successor, the TR-55, was shown in Tokyo in July 1955 and on sale that autumn - Sony’s own corporate history dates the launch to September 1955, though many secondary accounts say August [S1]. It was Japan’s first transistor radio, with five transistors made in-house, and the first product to reach the market carrying a new brand name - SONY - that the company would not adopt as its own name until 1958 [I9][I10][S1].

The TR-55 cost 18,900 yen, two and a half months of a graduate civil servant’s salary, and early buyers conceded the tube portables sounded better [I9]. It sold anyway, and then, in the summer of 1956, it exploded among young Japanese buyers, for the reason that would decide the transistor’s consumer future: running cost. A transistor radio sipped batteries where a tube portable drank them - Totsuko advertised battery costs one-twentieth of the tube sets’ [I9]. The transistor had found its first mass market not in the military systems that paid for its infancy, nor in the computers that justified its physics, but in the pocket of a teenager. Consumer electronics - cheap, portable, personal, global - was the demand the transistor had been waiting for, and a Japanese company found it first. In the following decades Japan would climb from licensed radios to memory-chip dominance, and the industry would learn that technology transferred is technology returned with interest. That climb, and the conflict it caused, belongs to the later chapters on business and power.

# Chapter 22 - The Valley becomes silicon

For most of the transistor's first decade, its map had two poles - Murray Hill and Dallas - and a scattering of outposts. By 1971 the map had a capital, and the capital had a name. How the Santa Clara Valley became Silicon Valley is not a story of a place choosing an industry. It is the story of an industry discovering that it needed a place - and of the few deliberate people who built one.

The place began as orchards. Into the 1950s the valley south of San Francisco was famous for fruit, the "Valley of Heart's Delight" [M26]. Its transformation rested first on a university with land and a dean with a doctrine. Frederick Terman of Stanford believed a great engineering school should grow great companies beside it: he had encouraged his students Hewlett and Packard to found their firm, lent them money, and arranged their first bank loan [M28]. As dean and then provost he institutionalized the doctrine in the Stanford Industrial Park - "our secret weapon," he called it - signing Varian Associates to the first ground lease in 1951, and by 1960 the park held more than forty companies on what had been farmland [M27][M28]. When Terman heard in 1955 that William Shockley was starting a transistor company, he wrote to point out the advantages of locating it near Stanford [M27]. The advantages were real, though not the ones anyone intended: what Shockley brought to the valley was not a successful company but a concentration of brilliant, soon-to-be-restless people.

The previous chapters have told what happened next: the walkout, Fairchild, and then Fairchild's own children. The pattern deserves a count. By 1980, more than fifty semiconductor companies in the valley could be traced directly to Shockley's original startup [M27]. Rheem, Amelco, Signetics, General Microelectronics, National under Sporck, Intel, AMD, and dozens more: each generation of the genealogy was founded by people who had learned the business inside the previous one, and each departure was financed a little more readily than the last, because the financiers had watched the earlier departures pay. A walkout that scandalized 1957 was, by 1970, simply how companies were born.

The third institution, after the university and the spin-off, was the capital. Arthur Rock, who had assembled the Fairchild financing from New York, moved west and backed Intel, becoming the model of a new profession [M19]. In 1972 two firms founded in the same year gave the profession its permanent local address: Kleiner Perkins - Eugene Kleiner, one of the eight, with Tom Perkins, an $8 million fund on Sand Hill Road, Noyce and Moore among its investors [M30] - and Sequoia Capital, founded by Don Valentine, a Fairchild and National sales veteran who evaluated markets rather than resumes and whose first fund backed Apple and Atari [M29]. By 1975 more than thirty venture firms were in the Bay Area [M31]. The valley had completed its loop: Stanford produced the engineers, the semiconductor companies

trained them, the walkouts freed them, and Sand Hill Road financed them - and the profits of each success refilled the machine.

The name arrived last, as names do. Through the late 1960s the people selling chips out of the valley had a slang name for it among themselves - “that’s what the guys call it,” a marketing man explained over lunch to Don Hoefler, a columnist for the industry’s weekly newspaper, Electronic News [M26]. Hoefler was finishing a three-part series on the semiconductor industry that had grown up around Fairchild, and on January 11, 1971, the series appeared under the headline “SILICON VALLEY U.S.A.” - the name’s first known appearance in print [M26]. Hoefler, reflecting a decade later, gave its rationale: “These revolutionary semiconductors are made in a valley, from silicon - not silicone, please - the second most-abundant chemical element (after oxygen) on Earth. How was I to know that the term would quickly be adopted industry-wide, and finally become generic worldwide?” [M26]

Generic worldwide: that is the measure of what the transistor built here. The phrase outgrew the place - “Silicon Valley is not a place but a state of mind,” as the saying attributed to various of its luminaries goes [M26] - and regions on every continent now brand their own valleys, glens, and corridors in imitation. The imitations mostly fail, which is instructive. The original was never an industrial policy or a real-estate scheme, though it had a research park; it was an ecology, assembled over twenty years by identifiable people making identifiable choices: Terman’s doctrine, Shockley’s recruiting, eight men’s resignation, Rock’s checkbook, Hoerni’s oxide, Noyce’s circuit, and ten thousand engineers who kept quitting comfortable jobs to start unreasonable companies. The transistor made the valley technically possible. The valley’s people made the transistor industrially inevitable. Between them they produced the thing this book’s next part takes up: an industry that learned to double on schedule.

# PART VII - SCALING

# Chapter 23 - Moore’s observation

In April 1965, the director of research at Fairchild Semiconductor published a three-page article in a trade magazine that organized the next sixty years. Gordon Moore’s “Cramming More Components onto Integrated Circuits” is famous now as the source of Moore’s Law, and misunderstood in a way worth correcting first, because the correction contains the whole point: it is not a law of physics, and Moore did not claim one. It is an observation about economics, made by a man trying to sell chips.

Moore’s construction was disarmingly simple. The cost of a circuit function, he argued, has a curious shape: the more components you integrate onto one chip, the cheaper each component becomes - up to the point where falling yields on the more complex chip begin to eat the savings. At any moment, then, there is a sweet spot: a complexity of minimum component cost [M33]. Moore plotted that sweet spot against time. He had five data points, from the planar transistor of 1959 to Fairchild’s current products in 1964, and through them he drew a straight line on a logarithmic scale: the economic optimum was doubling every twelve months [M32]. Extrapolating, he predicted that by 1975 the cost-optimal chip would carry 65,000 components [M33]. The article, published in Electronics on April 19, 1965, made the prediction to encourage Fairchild’s customers to design their next computers around chips that did not yet exist [M32]. It was, in part, a sales document - which turns out to be essential to understanding what followed.

Read closely, the article’s deeper claim is about causation. Moore did not say the doublings would happen by themselves. He identified the drivers - finer photolithography, larger wafers, better processes, and what he called “circuit and device cleverness” - and he noted that silicon would remain the industry’s material because of “the technology which has already evolved around it and its oxide” [M32] [M33]. The doublings were a forecast of what engineering effort would deliver, and by naming a number, the forecast recruited the effort. When Moore revisited the prediction at the 1975 IEEE International Electron Devices Meeting, he confirmed that the industry had hit his 65,000-component mark on schedule, and, folding in new data, he revised the cadence forward: a doubling every two years from then on [M32]. Carver Mead, the Caltech physicist, dubbed the cadence “Moore’s Law,” and a law it became in the only sense that mattered: the industry’s shared metronome [M32]. Roadmaps were written to it. Equipment makers tooled to it. Companies that fell behind it died. Moore had predicted the future, and the prediction had become a specification - a self-fulfilling prophecy, as he himself readily called it [M32].

A number of things commonly said about Moore’s Law are not true, and the record here is worth keeping straight. It was never about transistor size directly, and never about speed; it was about the count of components at minimum cost, an economic

quantity that rides on physics without being physics [M33]. It had an expiration date built into its logic from the start - Moore titled a 2003 address “No Exponential Is Forever: but ’Forever’ Can Be Delayed!” [M32]. And its two variants - one year, then two - are often mangled in retelling. What the observation actually said, decade after decade, was something subtler and more powerful than any fixed number: that in this industry, the future arrives on a schedule, and the schedule is set by whoever keeps pushing the cost curve down. The next chapter introduces the physicist’s half of the engine - the scaling rules that told engineers exactly how to shrink - and then we will follow the race those two documents started, all the way to its present, strained, astonishing state.

# Chapter 24 - Dennard’s rules

Moore told the industry what would happen. Robert Dennard told it how. In 1974, Dennard and his colleagues at IBM published a paper with an ungainly title - “Design of ion-implanted MOSFET’s with very small physical dimensions” - that contained the operating manual for the next thirty years of progress [P7]. The question it answered was practical: everyone could see that transistors would be made smaller, but smaller transistors misbehave. Shrink the channel and the electrical balance of the device breaks - the drain’s influence reaches where the gate’s should be, thresholds wander, the switch stops switching crisply. These short-channel effects were the wall of the early 1970s [P7]. Dennard’s group showed how to climb it, not by defeating the physics but by respecting it in a disciplined way.

The discipline is called constant-field scaling, and its beauty is that it is a recipe with one ingredient: proportion. Take a working transistor and shrink every dimension by the same factor - not just the channel length, but the oxide thickness, the junction depths, everything. Then shrink the voltages by the same factor, so that the electric fields inside the device stay exactly as they were. Finally, increase the doping by that same factor, to keep the charge landscape in proportion [P7]. Do all three together, and the scaled device is a perfect miniature of the original: same fields, same behavior, same reliability - only smaller. Dennard’s team proved the recipe in silicon, building working polysilicon-gate transistors with channels down to half a micron at a time when production devices were far coarser [P7].

Now follow the arithmetic, because the arithmetic is the modern world. Shrink dimensions by a factor of k and each transistor’s area falls by k squared - so each generation packs in roughly twice as many devices for a thirty-percent shrink. Delay falls by the same factor k, so circuits run faster. Power per device falls by k squared, because power scales with voltage squared. And here is the consequence on which three decades of industry rested: power density - watts per square millimeter of silicon - stays constant [G1]. Each generation delivered twice the transistors, running faster, at the same total heat. The chip did not get hotter as it got more powerful. That is why your telephone out-computes a 1975 room-sized mainframe while running cool in your pocket: for thirty years, miniaturization was a free lunch, and Dennard’s rules were the menu.

The lunch had a bill, deferred. The recipe requires shrinking voltages in proportion, and voltages cannot shrink forever: below a certain threshold, a transistor cannot reliably distinguish on from off, and leakage current - the same ghost that haunted germanium - returns in a new form, flowing even through devices that are supposed to be off. By the middle of the 2000s the industry could no longer scale voltage as the recipe demanded, and power density began its climb [G1]. The era that followed - the

multicore detour, the power wall, the desperate cleverness of the present decade - belongs to the chapters on limits. But between Dennard’s paper and that wall lay the longest run of compounding improvement any technology has ever sustained, and it is not exotic to say that the modern economy ran on it. Every forecast in this book’s remaining chapters - every fab, every market, every geopolitical anxiety - is downstream of two documents: a sales article from 1965 and a device-physics paper from 1974. One said shrink, and one said how.

# Chapter 25 - The node race

With Moore’s metronome and Dennard’s recipe in hand, the semiconductor industry converted progress into a procession. Its milestones are the “nodes” - the named generations of manufacturing technology - and their parade, with approximate years, runs like this: 10 microns in 1971, when the 4004’s transistors were carved; 3 microns in 1977; 1 micron in 1984; 800, 600, 350, 250, 180 nanometers through the 1990s; 130, 90, 65, 45 nanometers in the 2000s; 22, 14, 10, 7, 5, 3 nanometers in the 2010s and early 2020s [G2]. From first to last the printed feature shrinks by a factor of three thousand, and every step of it was won, not found. The node race is the industrial expression of everything the previous two chapters described: a global competition, run on a public schedule, in which falling behind by one generation could be fatal.

The schedule was made explicit. In the 1990s the industry formalized its self-fulfilling prophecy into an institution: the International Technology Roadmap for Semiconductors, a global consensus document, refreshed annually, that told every chip maker, equipment maker, and materials supplier what the industry would need years before it needed it [G3]. The ITRS recorded the actual cadence - a new node roughly every two years through the 1990s - and projected it forward, noting as the 2000s progressed that the cycle was stretching toward three years [G3]. This is one of the strangest documents in industrial history: competitors pooling their future plans so that their shared suppliers could build for them in time. It worked because Moore’s observation had made the future predictable, and it accelerated the very predictability it described.

A caution about names belongs here, because the node race’s later stages produced one of the industry’s durable confusions. Originally, a node’s number measured something real - the smallest feature, roughly the transistor’s gate length. That correspondence broke in the mid-1990s, and from then on the numbers became labels, not measurements [G2]. By the 2010s the divergence was farcical: one company’s “10 nanometer” process matched another’s “7 nanometer” in actual transistor density, and no transistor anywhere in a “3 nanometer” chip has any critical dimension of 3 nanometers [G2]. The race continued, but the scoreboard stopped measuring the track. What the labels still measure honestly is generation: who shipped first, at what density, at what yield, and at what cost.

For who shipped first became the only ranking that mattered. Through the 1970s and 1980s the leaders were the great integrated American firms - Intel, TI, Motorola - joined and periodically passed by the Japanese memory makers, with Europe’s Philips and Siemens and a dozen others always one generation behind the leaders and one generation ahead of extinction. The race’s cruelties were structural. Each new node demanded new equipment, new materials, and a new factory; the cost of staying in the

race rose with every lap; and the prize for leading - a few quarters of premium pricing before the pack arrived - had to be won again immediately. Companies that stumbled a generation rarely recovered their position; the list of former leaders is long and distinguished. And the race systematically concentrated: by the 2010s, only a handful of firms on Earth could afford a lap, and at the frontier, as the chapters on business and power will show, effectively one. The node race is the transistor’s promise turned into an arms race - and like every arms race, its deepest effect was on the number of players who could afford to run it.

# Chapter 26 - The factory that prints

Walk into a modern wafer fab and you will see almost no people. You will see a ballroom-sized cleanroom, lit yellow in the lithography bays because the photoresist must never meet the blue wavelengths that expose it, with rows of machines the size of trucks and, overhead, a ceiling railway of automated carriers - front-opening unified pods, FOUPs - shuttling sealed boxes of silicon from tool to tool on rails, thousands of moves an hour, untouched by human hands [G2]. The wafers inside are mirrors of pure crystal, 300 millimeters across, and each will pass through several hundred machines, accumulating layers, implants, etches, and printed patterns in a sequence that runs to a thousand-plus process steps over eleven to thirteen weeks - up to fifteen for the most advanced chips [G2]. The fab is the planar process of Chapter 13 grown to industrial maturity: the same loop of pattern, dope, deposit, etch, repeated dozens of times per wafer, at tolerances that would have seemed imaginary to the engineers who invented it.

Two disciplines rule the building. The first is cleanliness, and it is not hygiene but economics. A transistor a few dozen atoms across is killed by a speck of dust a hundred times smaller than what the eye can see, and a single killer defect anywhere on a chip costing hundreds of dollars destroys it. Modern cleanrooms hold the air around the wafers to particle counts thousands of times below a hospital operating room; the wafers themselves live in their nitrogen-filled FOUPs, a cleaner environment inside the clean environment, because even cleanroom air is dirty by the standards of a 3-nanometer transistor [G2]. The second discipline is yield: the fraction of chips on each wafer that work. Yield is the fab’s real product, and its most guarded secret. A new process may begin life throwing away most of every wafer; the months of diagnosis that raise yield from useless to profitable are where much of a fab’s true engineering happens, and they are why a rival can buy identical machines and still lose money printing the same chip.

Then there is the money. A state-of-the-art fab now costs roughly twenty to thirty billion dollars to build, about half of it machines [G4, G5, G6]; over a ten-year life, the full cost of owning one runs to thirty-five billion and up [G6]. The industry was projected to spend around 160 billion dollars on capital equipment in 2025 alone, and TSMC’s capital budget for 2026 approaches 56 billion dollars [G5, G4]. Inside the building, the star machines are the lithography systems: an entry extreme-ultraviolet scanner costs more than 200 million dollars, and the newest high-NA generation upwards of 350 million euros - over 400 million dollars - apiece, a price so steep that in 2026 even TSMC, ASML’s largest customer, announced it would defer adopting them through 2029 [G4, G5]. Compare these numbers with the industry’s origins: Fairchild’s entire first-year revenue was a rounding error in the deposit on one modern scanner, and the tool that prints a “3-nanometer” chip costs more than the total

capitalization of the companies that invented the integrated circuit. The next chapter looks at who builds these machines, because the answer to “why are there so few chipmakers left” is, in large part, “because there is essentially one company that can build the printer.”

# Chapter 27 - The equipment makers

Every machine in the last chapter’s cleanroom has a maker, and the list of makers has shrunk even faster than the list of chipmakers. Five firms now dominate the market for wafer-fab equipment - a hundred-billion-dollar-a-year business as of 2024: ASML of the Netherlands in lithography; Applied Materials of California across deposition, implantation, etch, and polish; Lam Research in etch; Tokyo Electron in coating, developing, and thermal processing; and KLA in the inspection and measurement tools without which nothing else can be controlled. Together they take about eighty-five percent of the top ten’s revenue [G9]. Below them sits a stratum of specialists and, notably, one rising Chinese maker, NAURA, the only Chinese firm yet in the global top ten [G9]. But one member of the five stands apart, because the story of modern chipmaking’s concentration is, more than anything, the story of ASML.

ASM Lithography began unpromisingly on April 1, 1984, as a joint venture between Philips - which had a stepper in its labs and no clear plan for it - and ASM International, a Dutch equipment entrepreneur’s firm. It opened in a leaky shed beside a Philips building in Eindhoven; forty-seven of the fifty Philips employees offered jobs in the venture dared to sign on [G7]. The first product, the PAS 2000, failed. The company survived on a 1986 partnership with the German optics house Carl Zeiss - a relationship that still defines it, since ASML’s machines are, at their core, vehicles for Zeiss’s lenses and mirrors - and broke through in the early 1990s with the PAS 5500 stepper, the platform that won the customers that made it profitable. By 2002 it was the world’s largest lithography supplier, past the once-dominant Japanese giants Nikon and Canon [G7, G8].

Its defining bet was the longest in the industry. In 1997 ASML started a program in extreme-ultraviolet lithography - printing with light of 13.5-nanometer wavelength, so short it is absorbed by air and glass alike, requiring vacuum vessels and mirrors instead of lenses. The science came from a consortium of American chipmakers and national laboratories, the EUV LLC, which ASML’s European consortium joined in 1999, accepting conditions that included sourcing most components for US-bound machines from US suppliers [G7, G8]. The machine that resulted is one of the most complex objects humans manufacture: its light source fires a high-power laser at droplets of molten tin fifty thousand times a second, hitting each droplet twice - once to flatten it, once to vaporize it into a plasma forty times hotter than the surface of the sun, which emits the EUV light the mirrors gather [G7]. ASML shipped its first pre-production EUV tool in 2010, acquired the San Diego light-source maker Cymer in 2013 to control the hardest problem, took billions in co-investment from Intel, Samsung, and TSMC - its own customers buying stakes to fund their supplier’s research - and reached reliable production machines only around 2018, thirteen years

later than its own engineers first predicted [G7, G8]. The first EUV-made commercial product reached consumers in a 2019 Samsung smartphone [G7].

The consequences define the present era. ASML is the sole source of EUV lithography on Earth - there is no second supplier, and none is coming soon - which makes a private Dutch company a chokepoint of the world economy, and makes export licenses for its machines an instrument of statecraft, as the chapters on power will describe. The equipment industry also completes the pattern this part of the book has traced from the start. Bell Labs gave the invention away; Fairchild industrialized it; the ecosystem separated design, manufacture, and equipment into distinct trades; and each trade concentrated until, at the sharpest edge, a single printer, a handful of fabs, and one indispensable foundry stood between the digital world and its hardware. The transistor democratized electronics. The business of printing it did the opposite.

# PART VIII - WHAT TRANSISTORS DO

# Chapter 28 - Amplifying

The transistor was invented to amplify, and most of the world’s transistors no longer do. That inversion is worth pausing on. The digital revolution recast the device as a switch, and the next chapter counts its conquest of logic. But the world itself is not digital. Sound, light, heat, pressure, radio waves, the flutter of a heart - every signal a machine senses arrives as a smoothly varying quantity, and every digital system lives embedded in that analog world, taking its inputs from it and returning its outputs to it. Between the two stand the analog transistors, doing the job Bardeen and Brattain’s device did on its first afternoon: making a small signal bigger.

The craft of analog design is older than the transistor and survived it nearly intact. Its foundation was laid at Bell Labs in the years before the invention: Harold Black’s negative-feedback amplifier of 1927, Harry Nyquist’s 1932 stability criterion, Hendrik Bode’s design mathematics of 1940 [N2]. Feedback - returning a fraction of the output to the input - is what turns a crude, drifting amplifier into a precise one, and it remains the central trick of the art. By 1947 the building block had a name, John Ragazzini’s “operational amplifier,” and by 1952 George Philbrick was selling a plug-in tube version, the K2-W, for about twenty-two dollars [N2]. The transistor’s first role in this world was simple substitution: germanium op-amps from 1958, silicon from 1960, each generation smaller, cooler, and more reliable than the last [N2].

The integrated circuit changed the economics. Widlar’s uA702 of 1964, the first op-amp on a chip, sold for three hundred dollars; his uA709 of 1965, “the first IC op-amp that works like an op-amp,” cut that to seventy and sold hugely [N2, N3]. When Fairchild declined his request for a raise, Widlar walked to National Semiconductor and designed the LM101; Fairchild’s answer came from a young engineer named David Fullagar, who saw that the LM101’s input stage was too sensitive to manufacturing variation, fixed it with a couple of extra transistors - the idea came to him, he said, driving to Tahoe - and produced the uA741 of 1968, the most successful op-amp ever made [N3]. Variants of it have sold in the billions and remain in production more than half a century later; the price of one original 702 now buys about seventeen hundred modern 741s [N3]. That longevity is the analog world’s signature. Digital chips are obsolete in years; a good analog design sells for decades, because nature’s signals do not scale with Moore’s law, and neither does the difficulty of handling them [N1].

Analog’s deeper claim on this book is epistemic. The analog business taught the industry that transistors are not abstract logic elements but physical devices, each one slightly different from its neighbor, each one a negotiation with noise, temperature, and the material’s moods. It bred its own legendary engineer-craftsmen - Widlar above all, with his axe and his sheep - and its own disciplines of layout and matching that

digital design would later have to relearn [N1]. And it never went away: every phone’s radio, every sensor, every battery charger and audio jack is a colony of analog transistors. The digital world won the headlines. The analog world kept the interface to reality - and reality, it turns out, is where the signals live.

# Chapter 29 - Switching

The transistor’s second life - the one that ate the world - began with a demotion. In a digital circuit the transistor no longer amplifies; it merely switches, full on or full off, and it spends its life being driven so hard that its subtleties vanish. This looked, to some early engineers, like using a violin as a hammer. It was in fact the deepest simplification in the history of electronics, and everything digital follows from it.

The reason is regeneration. An analog signal accumulates noise at every stage it passes through; a chain of amplifiers is a chain of accumulated distortions. A digital signal is different: it is allowed to be only one of two values, and at every gate it is restored - a slightly degraded “one” comes out a perfect one, a slightly corrupted “zero” a perfect zero. Noise never accumulates, because it is discarded at every step. This single property is why digital electronics could scale without limit while analog electronics could not, why a billion-transistor chip is buildable and a billion-stage amplifier is not, and why the world’s information moved, over the second half of the twentieth century, from continuous form into bits. The transistor made switching cheap; switching made error-free complexity possible; and complexity, reliably manufactured, is the whole story of computing.

The practical path ran through logic families - successive recipes for wiring transistors into gates. Resistor-transistor logic gave way to diode-transistor logic, and then, in the early 1960s, to transistor-transistor logic, TTL, which Texas Instruments turned into the 7400 series: a catalog of standard gates, flip-flops, and counters that became, for two decades, the default vocabulary of digital design [I6, I7]. A generation of computers was built by wiring these chips by the hundreds onto boards. Meanwhile MOS technology, slower but far denser and lower-powered, was climbing underneath: the chapters on Intel traced the result - memory chips that drowned core memory, and then the microprocessor, which put a whole computer’s logic on one die [M23, M24, M25]. CMOS, the pairing of complementary transistors that draws almost no power while holding still, completed the architecture in the 1980s and has carried it since [M15, M16]. Each family displaced its predecessor on the same grounds: more switches, less power, less money.

It is worth stating plainly what a modern chip is, because the numbers have ceased to be imaginable. A contemporary processor holds tens of billions of transistors, each switching billions of times per second, and a misbehaving handful would render it scrap. That such objects work at all - that they work so reliably we build civilization’s infrastructure on them - is the compound interest of regeneration, Dennard scaling, and fifty years of process discipline. The switch, the humblest thing a transistor can be, turned out to be the most consequential. Not because switching is clever, but

because it can be done without limit, without drift, and - at scale, printed by the trillion - essentially without cost.

# Chapter 30 - Remembering

A computer that cannot remember is a calculator with amnesia. The transistor’s conquest of computing required it to master storage as well as logic, and the memory story runs on its own track through this book’s history - with its own physics, its own champion company, and one of the industry’s great second acts.

When the integrated circuit arrived, the standard memory of computing was magnetic core: millions of tiny ferrite rings, each strung by hand on crossing wires, each storing one bit. Core memory was reliable and fast, but it was the most labor-intensive component ever mass-produced - the Apollo guidance computer’s memory was famously woven by textile workers - and its cost per bit could fall only so far. The semiconductor alternative was demonstrated by Robert Dennard at IBM in 1966: one transistor and one capacitor per bit, the transistor as the cell’s gatekeeper, the capacitor as the tiny bucket holding the charge that means one or zero [P7]. The dynamic random-access memory was born. Its catch was built into its physics: the charge leaks, so every cell must be read and rewritten thousands of times a second, forever - “dynamic” is the polite word for forgetting on a schedule.

Intel made DRAM an industry. Its 1103 of 1970, a kilobit of memory designed by Ted Hoff and process-engineered by Les Vadasz, was the first DRAM that worked well enough to replace core, and it did exactly that - becoming, by the early 1970s, the best-selling semiconductor device in the world and the engine of Intel’s early growth [M23]. DRAM’s one-transistor cell made it the densest circuit anyone could print, which made it the industry’s technology driver for two decades: memory makers led the node race, because their simple, repeating cells were the perfect vehicle for proving a new process generation. It also made DRAM the industry’s crucible: a commodity so brutally priced that it drove out the Americans in the 1980s, concentrated production in Japan and then Korea, and left the memory business a three-company oligopoly - a story the business chapters pick up.

The second memory revolution answered a different question: what remembers when the power is off? The answer began, improbably, with a bug. In 1969, Dov Frohman at Intel was asked to chase a reliability problem in the company’s first MOS memory; he found that charge was being trapped in the oxide, corrupting the device - and realized the defect could be turned into a feature: trap the charge deliberately, on a buried “floating” gate, and you have a memory cell that holds its bit for years with no power at all [N4, N5]. His EPROM of 1971, erased by ultraviolet light through a quartz window in the package, became the microprocessor’s essential companion and Intel’s most profitable product line into the mid-1980s [N4, N5]. At Toshiba’s labs in Kawasaki, Fujio Masuoka pushed the idea further: erase electrically, in bulk - “in a flash,” a colleague said, and the name stuck [N5]. His NOR flash appeared in 1984,

his denser NAND architecture in 1987, and by 1992 Toshiba was producing 16-megabit NAND chips [N5]. Toshiba, slow to see what it had, let others commercialize aggressively; Masuoka eventually sued his own employer for recognition. But flash remade the industry: it replaced film, then the hard disk in laptops and phones, and made the solid-state era possible.

Memory thus tells the transistor’s story in miniature twice over. DRAM is the triumph of the minimal cell - one transistor, one capacitor, scaled for half a century. Flash is the triumph of the engineered imperfection - a flaw in the oxide, domesticated into humanity’s cheapest way to remember. Both began as answers to narrow problems and ended as foundations of the digital world.

# Chapter 31 - Power and light

The transistor family has two brothers who left home early and built separate fortunes, away from the integrated-circuit mainstream. One tames electricity in bulk; the other converts it to and from light. Neither makes headlines like a processor, and between them they touch every electron in the modern economy.

The power transistor’s job is to be a valve for serious current - switching the hundreds of volts and amperes of motor drives, power grids, trains, and chargers, rather than the whispers of logic. The workhorse for the first decades was the thyristor, a four-layer latching switch descended from Shockley’s own device ideas, superb at conducting but clumsy to turn off. The breakthrough that made power electronics modern was, once again, the MOS gate: the power MOSFET, developed through Hitachi’s and Siliconix’s VMOS work in the 1970s and commercialized by International Rectifier’s HEXFET in 1978, put a voltage-controlled gate on a power switch [N7]. Then Jayant Baliga at General Electric fused the two lineages: the insulated-gate bipolar transistor, conceived in a 1977 patent disclosure, demonstrated in 1979, and released as a product in June 1983 - a device with the MOSFET’s easy gate and the bipolar’s brute conduction, which became the standard switch of high-power electronics [N6, N7]. Every variable-speed motor, every electric-vehicle inverter, every modern locomotive and solar inverter switches its power through IGBTs or their descendants. Baliga has argued, with some justice, that the energy saved by efficient semiconductor switching runs into the gigatons of avoided carbon - the transistor’s least visible and possibly largest environmental ledger entry.

The optoelectronic branch runs in the other direction: not controlling power but exchanging it for light. The silicon solar cell came first and from the same building as the transistor itself - Bell Labs, 1954, where Daryl Chapin, Calvin Fuller, and Gerald Pearson turned a serendipitous photocurrent in a doped silicon wafer into a cell that converted six percent of sunlight to electricity, demonstrated to the press by powering a toy Ferris wheel [N9]. Its first market was space: Vanguard 1 in 1958, and essentially every satellite since [N9]. Seven decades later its descendants are the cheapest source of bulk electricity humanity has ever built.

Light emission took a different material path. Silicon, it turns out, is a poor light emitter - its band structure makes electrons shed energy as heat rather than photons - so the light-emitting branch grew up on the compound semiconductors this book met in Chapter 9: the III-V alloys like gallium arsenide. In the autumn of 1962, within weeks of each other, Robert Hall at GE’s Schenectady lab made the first semiconductor laser (invisible infrared), with IBM and MIT Lincoln Laboratory reporting similar devices nearly simultaneously, and Nick Holonyak - John Bardeen’s first doctoral student - at GE’s Syracuse lab made the first one that shone visible red:

the first practical LED [N8]. Holonyak predicted his “ultimate lamp” would go far. It did: LEDs became indicators, then displays, then general lighting, after the blue LED - the hardest color, won in the 1990s - completed white light. And the semiconductor laser went on to carry the internet: every fiber-optic cable on Earth is lit at one end by a descendant of Hall’s 1962 device. The transistor began by replacing the vacuum tube in the telephone network. Its optical cousins ended up carrying that network’s entire traffic, one photon at a time.

# PART IX - THE BUSINESS

# Chapter 32 - The cycle

For all its aura of exponential inevitability, the semiconductor business is one of the most violently cyclical in the world economy. The cycle is built into the physics of the factory: a fab takes years and tens of billions of dollars to build, and once built it must run full, because nearly all its cost is fixed. Demand, meanwhile, swings with consumer fashions, corporate budgets, and panic. When demand runs ahead, chips are allocated, prices spike, and everyone orders extra against shortage - which sows the glut. When demand falls behind, the fabs cannot slow down without bleeding cash, prices collapse, and the weakest players die. The industry has ridden this sine wave its entire life: the memory crashes of the mid-1980s that broke the American DRAM makers, the 2001 collapse after the dot-com bubble, the 2009 crash, the memory glut of 2019.

The classic illustration is the memory wars. In 1980 the United States still held more than half of the world semiconductor market and Japan about a quarter. Japanese memory chips, acclaimed by the American computer makers who bought them for quality, delivery and price, took the lead in the part that then drove the industry: Japan’s share of 64-kilobit DRAM passed America’s in 1981, and the Japanese share of the whole DRAM market reached about eighty percent by 1987 [I11]. Washington answered with dumping suits and, in 1986, the U.S.-Japan Semiconductor Agreement - the first great trade conflict of the transistor age [I11]. The American memory industry never won its lead back, and the next challenger rose in Korea: Samsung declared its entry into semiconductors in 1983, developed its own 64-kilobit DRAM within the year, and was mass-producing the 256-kilobit generation by 1985 [I12]. By the late 1990s the Koreans had overtaken the Japanese, completing memory’s migration from America to Japan to Korea [I12].

The most instructive modern lap began in 2020. Pandemic demand for computers and home electronics surged just as just-in-time supply chains canceled their chip orders; the result, from the second half of 2020 through 2022, was a global shortage severe enough to idle car factories on four continents - automakers, who had canceled first, discovered that a thirty-thousand-dollar vehicle cannot ship without a fifty-cent microcontroller [B2]. Scarcity pricing and double-ordering followed the classic script, and so did the hangover: by 2023 the industry was in a downturn, with market value falling from $598 billion in 2022 to about $544 billion [B2]. Then came the AI boom, and the sharpest upswing in the industry’s history: the market rebounded roughly nineteen percent in 2024 to about $627 billion, with memory up eighty-one percent in a single year, and forecasts - as of late 2024 - projected about $697 billion for 2025 and a trillion-dollar market around 2030 [B2]. The industry outran both marks at once: 2025 closed at $791.7 billion, up more than a quarter in a single year, and the industry’s trade association now projects roughly a trillion dollars for 2026 [M1].

But the averages now conceal a structural change, and it is the most important business fact of the present era. The profits of the boom are not shared. McKinsey's accounting of "economic profit" - profit above the cost of capital - tells the story: the whole industry generated about $38 billion in the 2000s, most of it at Intel alone; $450 billion in the 2010s; and $473 billion from 2020 to 2024, a sum concentrated in the AI-adjacent few - the leading foundry, the leading accelerator designer, the equipment makers [B2]. Between 2021 and 2024, the middle ninety percent of companies saw their average economic profit per company fall from $130 million to $17 million, below pre-pandemic levels [B2]. The transistor's market has become a barbell: a handful of firms at the frontier capture nearly all the value the device creates, while the long tail fights over the rest. The cycle still turns. It just turns around fewer and fewer axles.

# Chapter 33 - The foundry idea

The deepest restructuring of the semiconductor industry since the integrated circuit was not a device but a division of labor, and it came from a man the American industry had passed over. Morris Chang spent a quarter century at Texas Instruments, rising to run its worldwide semiconductor business, and left when the top job went elsewhere. In 1985 Taiwan's government - which had been methodically building a chip industry since the 1970s by licensing American technology and spinning companies out of its Industrial Technology Research Institute - recruited him to head ITRI and gave him, in effect, a blank check to design the island's semiconductor future [B1]. Chang's answer, founded in 1987, was the Taiwan Semiconductor Manufacturing Company, and its premise was a business model the industry had never tried: a pure-play foundry - a factory that manufactures chips designed by others, sells no products of its own, and therefore never competes with its customers [B1].

The founding was a triumph of statecraft as much as strategy. The startup capital was $220 million: about 48 percent from the government's National Development Fund, 27.6 percent from Philips - the only established chipmaker willing to sign on, after TI and Intel declined - which contributed technology and patents along with cash, and most of the rest from eight of Taiwan's wealthiest families, who were, by polite accounts, "asked" by the government to invest [B1]. The first CEO lasted a year; Chang took the chair himself and held it for three decades.

The foundry's significance took time to become visible. At first TSMC simply absorbed the overflow and the trailing-edge work of the integrated giants. But the model inverted the industry's economics. A fab's crushing fixed cost had always forced chip companies to be manufacturing companies first; now a designer could exist with no factory at all. Through the 1990s the "fabless" industry Chang had enabled - firms like Nvidia, Qualcomm, Broadcom, and hundreds of others that design chips and buy their manufacture - grew from nothing into the industry's creative center, and every fabless success made the foundry stronger [B1]. Meanwhile the integrated device manufacturers, carrying their fabs on their backs, found the cost of each new node harder to justify against TSMC's scale, and one by one exited the leading edge - most fatefully AMD, which sold its fabs in 2009, and Intel, whose struggles to keep its own factories at the frontier define the industry's current drama. By the 2010s TSMC was not merely the largest foundry but the most advanced manufacturer of logic chips on Earth. In 2025 it produced 12,682 distinct products for 534 customers, with an annual capacity of more than seventeen million twelve-inch-equivalent wafers [B1].

There is a coda this book cannot skip, because it is the hinge between business and power. The foundry model dispersed design across the world and concentrated

manufacture in one place: the most advanced chips on the planet are printed overwhelmingly on one island, by one company, in fabs clustered along Taiwan's western corridor. Chang's division of labor was an economic masterstroke that created, as a side effect, the single most consequential geographic concentration in the modern economy. What that concentration means - for Taiwan, for China, for the United States, and for everyone who depends on chips, which is everyone - is the subject of Part X.

# Chapter 34 - The cost of the next step

Moore’s law has a shadow twin, less famous and more binding, sometimes called Rock’s law after the venture capitalist Arthur Rock: the cost of a chip fabrication plant doubles roughly every four years [B5]. The first law says transistors get cheaper; the second says the factory that makes them does not. The history of the semiconductor business is the long arc of the second law closing on the first.

Gordon Moore himself documented the opening of the scissors. In a 1995 speech he recalled that when Intel built its first fab in 1968, a production tool - a diffusion furnace, an exposure machine - cost about twelve thousand dollars. By the mid-1990s a comparable tool cost twelve million, and it did not process wafers a thousand times faster to compensate. A fab that cost millions in 1968 cost over two billion by 1998, and Moore noted the capital cost curve was rising faster than industry revenue itself [B3]. He ended that talk by declining to extrapolate past the 0.18-micron generation: terra incognita, he called it, in the phrase of the old mapmakers [B3]. The man who made the industry’s most famous prediction refused to extend it past the point where economics, not physics, drew the map.

The territory beyond turned out to be chartable, just increasingly expensive. The bills for the frontier have already appeared in this book: a state-of-the-art fab at twenty to thirty billion dollars, EUV scanners at two hundred million and up, industry capital spending around 160 billion dollars a year [G4, G5, G6]. To those add the cost of design: by industry estimates, roughly 40 million dollars to design a chip on the mature 28-nanometer node, over 200 million at 7 nanometers, over 400 million at 5, and approaching 600 million at 3 [B4]. And beneath the accounting, one line in the ledger broke first: around the 28-nanometer generation, in the early 2010s, the cost per transistor stopped reliably falling. Lithography’s price rose exponentially below that point as each new node demanded multiple patterning passes; foundries began signaling that the next node might not be cheaper per transistor than the last [B4]. The economic engine of Moore’s law - cheaper transistors every two years - had coughed, though not stalled: density kept rising, and for the most advanced products the performance justified the price.

The consequences propagate everywhere in this book’s remaining chapters. Rising fab costs are why the integrated device makers quit the frontier one by one, why the foundry model won, and why leading-edge manufacture concentrated in a handful of firms and one island. Rising design costs are why the chip industry stratified into a few companies that can afford the frontier and a vast ecology living profitably on older nodes. And the cost of the next step is now the central question of industrial policy: when a single fab costs as much as a nuclear power plant, governments, not just companies, decide where it gets built - which is where the story goes next.

# PART X - POWER

# Chapter 35 - The island

No square mile on Earth now carries more economic and strategic weight than the science parks of Taiwan’s western corridor. The numbers, by now familiar in this book, remain hard to absorb: Taiwan produces over sixty percent of the world’s semiconductors and - by industry estimates cited through the early 2020s - more than ninety percent of the most advanced ones, nearly all of them from a single company, TSMC [B8]. Chips account for roughly fifteen percent of the island’s GDP [B8]. And the most advanced production was, until very recently, made only there. The previous chapters explained how this happened: Chang’s foundry model, the economics of Rock’s law, the one-by-one exit of the Western and Japanese giants. This chapter is about what it means.

It means, first, that the digital world runs through one earthquake zone on a contested island. Every advanced smartphone, every AI datacenter, every modern weapons system depends on chips whose leading-edge manufacture sits a hundred miles from the mainland that claims the island as its own. Taiwanese strategists call the industry the “silicon shield”: the idea that the island’s indispensability guarantees both American protection and Chinese caution, because neither superpower can afford the fabs destroyed or captured [B8]. The logic cuts both ways, and everyone knows it. The same concentration that deters also tempts, and the same indispensability that protects Taiwan also motivates every other power to reduce it.

It means, second, that the transistor - the device Bell Labs gave away in 1952 precisely so that it would spread - has reconcentrated into the tightest chokepoint of any foundational technology in history. Oil never had a Taiwan; neither did steel, electricity, or the internet’s software. A technology present in every device on Earth, manufactured at the frontier essentially in one place, by one firm, with machines from one Dutch company, is a structure with no precedent and no comfortable theory. The diversification now under way - TSMC fabs rising in Arizona, Kumamoto, and Dresden [B1, B7] - is real but measured in years and hundreds of billions of dollars, and even when complete it will leave the densest leading-edge capacity where it has always been.

And it means, finally, that chips have moved from the business pages to the war colleges. The remaining chapters of this part describe the two great policy responses of the 2020s - the American effort to rebuild domestic manufacture and to wall China out of the frontier, and the worldwide scramble to map and secure the supply web. Both are, at bottom, attempts to answer the question Taiwan’s existence poses: what happens to a civilization whose most essential artifact is made in one place it cannot control?

# Chapter 36 - The chip war

On October 7, 2022, the United States Department of Commerce issued a rule that marked the end of an era. Without new legislation, using export-control machinery built for arms, the Bureau of Industry and Security restricted the sale to China of advanced computing chips, the equipment to manufacture advanced semiconductors, and - in an extraordinary reach - the ability of “U.S. persons,” citizens and residents alike, to support advanced Chinese fabs at all [B6]. The stated purpose was military: to slow China’s acquisition of the chips that modern weapons, surveillance, and AI systems require [B6]. The practical effect was broader: the world’s largest chip market was partially walled off from the frontier, and the industry learned that its most important technology now moved at the pleasure of governments.

The controls tightened by ratchet. An October 2023 update closed workarounds - replacing chip-speed thresholds with a “performance density” rule designed to catch chips redesigned to slide just under the limit, extending licensing to arms-embargoed countries and to the foreign subsidiaries of Chinese-headquartered firms, and imposing on foundries new duties to police their own customers’ designs [B6]. The Dutch and Japanese governments followed with parallel restrictions on their equipment makers, turning ASML’s export licenses into geopolitical instruments [G8]. China answered with its own levers - restrictions on gallium and germanium exports, among others - and redoubled its subsidies for domestic production, with its national champion SMIC demonstrating surprising, costly progress behind the wall. Whether the controls will preserve a durable lead or merely build a self-sufficient rival is the open question on which billions now turn.

The other half of the response was subsidy. The CHIPS and Science Act, signed August 9, 2022, committed roughly $52 billion to rebuild American chipmaking: $39 billion in manufacturing incentives, $11 billion for research programs, $2 billion for a defense microelectronics network, plus a 25 percent investment tax credit [B7]. By late 2024 the awards were final: about $7.9 billion to Intel - later expanded, with the government taking an equity stake, to roughly $9 billion or more - $6.6 billion to TSMC’s Arizona complex, $4.7 billion to Samsung in Texas, $6.2 billion to Micron, and smaller sums across the supply chain [B7]. The results, as of 2026, are mixed and early: fabs are rising in Phoenix and elsewhere, TSMC’s second Arizona fab was delayed and then restored to schedule, and the deepest question - whether subsidized fabs outside the ecosystem that produced TSMC can match its cost, yield, and pace - remains unanswered. Europe, Japan, India, and others launched parallel programs of their own. After seventy years in which the industry’s geography was set by economics, the 2020s reasserted politics. The transistor was born in a government laboratory monopoly, spread by a government decree, and grew up in a global market. Its eighth decade is being negotiated, once again, between governments.

# Chapter 37 - The supply web

The transistor’s geography is usually told as a story about fabs. It is better understood as a web, and the web’s strands run through places no headline mentions. Follow a single advanced chip backward from the phone in your pocket and the journey quickly spans the world: designed in California on software from two American firms, built on intellectual property licensed from Britain, printed in Taiwan on wafers from Japan, with light from a Dutch machine powered by an American laser, through resists from Japanese chemistry, etched by American and Japanese tools, packed with memory from Korea, assembled and tested across East Asia. The 2021-2022 shortage taught this lesson with a ruler: when a February 2021 Texas ice storm cut power to the Austin fabs of Samsung, NXP, and Infineon for weeks, and a March 2021 fire gutted a Renesas automotive-chip line in Hitachinaka, car plants worldwide stalled - because a chip designed against one foundry’s process cannot simply be made elsewhere, and a three-month fabrication cycle turns any interruption into a quarter of lost output [B10].

The web’s deeper feature is its chokepoints - stages where one or two suppliers hold most of the world. Some are famous by now: ASML as the sole source of EUV scanners; TSMC with roughly three quarters of foundry revenue and nearly all leading-edge capacity; a three-company oligopoly - SK hynix, Samsung, and Micron - holding nearly all high-bandwidth memory [B9]. But the obscure ones are as stark. Japan supplies about ninety-five percent of EUV photoresist and roughly four fifths of all photoresist, dominates silicon wafers through Shin-Etsu and SUMCO’s combined half of the market, makes every EUV mask blank, and holds the grinders, testers, and specialty films - one company, Ajinomoto, better known for soup seasoning, supplies over ninety-five percent of the dielectric film inside chip packages [B9]. China refines about ninety-nine percent of the world’s gallium and ninety-one percent of its rare earths [B9]. The United States holds the design-software duopoly - Synopsys and Cadence - and eighty percent of the high-purity quartz from a single district in North Carolina [B9]. Even the gases were shown fragile: before 2022, Ukraine supplied most of the world’s neon for lithography lasers, and the invasion briefly threatened it [B9].

Two structural properties make the web hard to reroute. The first is qualification: every material, resist, and component is validated in a specific fab over seven to seventeen months, so substitution is slow even where a second source exists [B9]. The second is that the web’s strands were placed by decades of comparative advantage, and they are cheapest exactly where they are. The policy projects of the chip-war era - reshoring fabs, stockpiling materials, subsidizing second sources - are real and partially working, but they are rebuilding at the margin a structure that took fifty years to optimize. The transistor democratized electronics and then wove the world into a

fabric no single thread of which can be easily replaced. Understanding that fabric - its strengths, its fragilities, and its keepers - is now basic literacy for any state that intends to remain one.

# PART XI - LIMITS

# Chapter 38 - The physics wall

Every exponential meets its wall, and the transistor’s was always going to be the atom. A silicon atom sits about a quarter of a nanometer from its neighbor in the crystal lattice, so a “2-nanometer” transistor - whatever the marketing name means - has critical features only tens of atoms across. Below that there is no smaller to go: the device is made of atoms, and atoms are not subdividable parts. But the industry met its real wall earlier than the atom, and it was thermodynamic, not geometric.

The wall’s first brick was power, and Chapter 24 told the story: around 2005 Dennard scaling failed, voltage stopped falling, and chips hit a power ceiling that ended the free rise of clock speed. The reason voltage stopped falling deserves its own name, because it is the deepest limit the transistor has yet met: the Boltzmann tyranny [F2]. A transistor switches by raising and lowering an energy barrier, and at room temperature the thermal spread of the electrons sets a hard floor on how sharply that barrier can work - at best, about sixty millivolts of gate voltage per decade of current change. No material improvement removes this; it is thermodynamics, not engineering. Its consequence is that supply voltage cannot fall much below half a volt without the transistor leaking ruinously in its “off” state - and since power scales with the square of voltage, the industry’s main lever on energy is stuck [F2]. Everything in the modern era - FinFETs, gate-all-around, the whole menagerie of the next chapter - is, in large part, an attempt to improve the electrostatic grip on the channel within this constraint.

The second brick is leakage’s quantum cousin. As the insulating layers of the transistor shrank to a few atomic layers, electrons stopped respecting them: quantum tunneling lets a particle cross a barrier it classically cannot surmount, and an oxide thin enough leaks like a sieve. The high-k/metal-gate fix of 2007 - thicker insulators of higher permittivity - bought the industry a decade, but tunneling never sleeps; it merely retreats to the next layer of the stack [F1]. The third brick is statistical. When a transistor’s doped regions contain countable dozens of impurity atoms, the exact number varies from device to device, and variation is the enemy of circuits that must switch reliably by the billion. At the frontier, transistors are no longer identical parts but populations with distributions - and design has become the management of distributions.

And beyond these lies the accounting of physics itself. Erasing one bit irretrievably dissipates a minimum energy, kT ln 2 - Landauer showed in 1961 that this is not an engineering target but a law, the thermodynamic price of forgetting [F2]. Today’s transistors operate thousands of times above that floor, so the law does not yet bind; but it defines the end of the road as surely as the atom does. The honest summary of the physics wall is this: the transistor did not stop improving; it stopped improving for

free. The gains now come at rising cost, in money and in ingenuity, from geometry, materials, and architecture - which is exactly what the next chapter describes.

# Chapter 39 - After planar

For five decades the transistor was a flat device - a channel in the silicon surface, a gate on top - and scaling meant shrinking the flat drawing. The planar era ended not with a new physics but with a new posture: the transistor stood up.

The fin came first, and it came from Berkeley. In 1996, on a flight to Japan, Chenming Hu sketched the specifications of a transistor whose channel would rise out of the wafer as a thin vertical silicon fin, wrapped by the gate on three sides instead of overlaid from one; by 1999 his group with Jeffrey Bokor and Tsu-Jae King Liu had built it and named it the FinFET, answering a DARPA challenge to design for a scale ten times beyond anything then in production [F1]. Berkeley declined to patent it - one of the great gifts in this book's history of gifts - and by 2002 the industry's roadmap had chosen the fin as its future [F1]. Intel brought it to high-volume manufacturing in 2011 at the 22-nanometer node, calling the Tri-Gate transistor the most radical shift in fifty years, with the numbers to argue it: a third faster at low voltage, or half the power at the same speed [F1]. The industry followed, and the FinFET carried the frontier from 22 down through 3 nanometers.

The fin's successor wraps the channel completely. The gate-all-around transistor - the channel as a stack of suspended silicon nanosheets, gate metal on all four sides - maximizes the electrostatic grip the Boltzmann tyranny makes so precious [F1, F2]. Samsung shipped it first, beginning 3-nanometer GAA production in June 2022; Intel's version, RibbonFET, paired with a second revolution - backside power delivery, moving the power rails to the wafer's underside so the front side carries only signals - was slated for its 20A node ramping in 2024, with 18A following [F1]. The plans landed, unevenly. TSMC's 2-nanometer nanosheet process entered high-volume manufacturing in the fourth quarter of 2025 and supplied 8 percent of the company's wafer revenue by the third quarter of 2026 [S2]. Intel's 18A - RibbonFET with backside power - reached volume production for Intel's own products by late 2026, with management reporting yields at what it called industry-standard levels on its October 2026 earnings call; no external foundry customer had yet been publicly named [S3]. Samsung's 2-nanometer GAA generation shipped in its Exynos 2600, powering some Galaxy S26 phones in select markets; its yields, reported around 60 percent, come only from the trade press, since Samsung discloses none [S4]. As of late 2026 the frontier runs on nanosheets, and the roadmap beyond them - stacked complementary transistors, two-dimensional channel materials - is a matter of public research, not yet production.

The third dimension has a second meaning, equally consequential: when the transistor would not shrink fast enough, the industry began to stack and to split. Chiplets break a giant die into smaller ones - cheaper to yield, mixable across process nodes - joined by

advanced packaging into one system; high-bandwidth memory stacks DRAM dies vertically beside the processor; TSMC’s CoWoS packaging, assembled entirely in Taiwan, has become as much a chokepoint as lithography itself [B9]. Moore’s law, in other words, has migrated upward, out of the transistor and into the package. The modern chip is no longer a chip: it is a city of dies, stacked and stitched, whose skyline keeps rising even as the ground plan approaches the atom.

# Chapter 40 - Different switches

Suppose the transistor as we know it truly does reach its end - the channel a handful of atoms, the oxide a handful of angstroms, the Boltzmann tyranny uncheated. What then? The research literature answers with a menagerie of candidate switches, and surveying them is the fitting close of this part, because it reveals what the transistor really is: not a device but an idea - a controllable bottleneck in the flow of a physical quantity - that has already outlived three of its own embodiments.

The nearest candidates keep the electron but change the mechanism. Tunnel field-effect transistors let electrons quantum-tunnel through the barrier rather than climb it, promising to break the sixty-millivolt floor and switch at voltages ordinary transistors cannot reach; ferroelectric FETs store state in the switch's own polarization; two-dimensional materials - sheets of molybdenum or carbon one atom thick - promise channels without dangling bonds at scales silicon cannot reach [F2]. All remain laboratory devices in 2026, and all have been "five years away" for fifteen years, which is itself information: displacing an incumbent that improves while you chase it is the hardest trick in technology.

The stranger candidates abandon charge as the signal entirely. Spintronics computes with the electron's magnetic moment - already victorious in memory, where magnetic tunnel junctions store bits as magnetization - and proposes collective switches that flip a material's order parameter, magnetic or ferroelectric, at energies potentially fifty times below what a scaled transistor can reach [F2]. Photonics moves the signal from electrons to light, already dominant in carrying data between chips and now pressing into the interconnects within them. Neuromorphic architectures change the question rather than the device, distributing memory and computation together as brains do instead of shuttling data between separate stores. And quantum computing replaces the bit with the qubit - not a better transistor but a different physics of information altogether, powerful for a narrow class of problems and irrelevant to most of what computers do [F2]. None of these is the next transistor, and it matters to say why. The transistor's successors will not arrive as one device the way the transistor did, because the transistor's own success has become the barrier: the installed base - the fabs, the design tools, the software, the trained world - is the deepest moat any technology has ever had. Whatever comes next will grow first in the transistor's shadow, as a complement at its edges, long before it replaces the thing itself. The menagerie's lesson is the book's lesson in miniature: revolutions in this field are real, but they arrive slowly, then all at once, and they are always built, never inevitable.

# PART XII - THE CONNECTED SYSTEM

# Chapter 41 - Energy and environment

The transistor’s environmental ledger has entries on both sides, and honesty about it is a condition of taking the device seriously. The credit side is real: semiconductors are the great dematerializers. A meeting that does not require a flight, a map that does not require printing, a library that weighs nothing, a power grid balanced by power electronics, an engine controlled to its most efficient stroke - across the economy, computation substitutes information for energy and matter, and the IGBT alone, as Chapter 31 noted, plausibly saves more energy than the industry consumes. The debit side is growing, concrete, and measurable, and it comes in two parts: making the chips and running them.

Running them is now visible at the scale of nations. The International Energy Agency estimates the world’s data centres consumed about 415 terawatt-hours of electricity in 2024 - roughly 1.5 percent of global supply - and projects, in its base case, more than a thousand terawatt-hours by 2030, approximately Japan’s entire electricity consumption, driven above all by AI [E1]. The emissions attached depend on the grid: the IEA projects data-centre electricity emissions peaking near 320 million tonnes of CO2 around 2030 - still under one percent of the global total, but among the fastest-growing sources [E1]. The transistor, invented to replace power-hungry tubes precisely because it sipped milliwatts, now aggregates its sips into national electricity budgets. The lesson is not that the device became wasteful - each computation costs astonishingly less energy than in 1947 - but that efficiency at the component level is always spent at the system level. Jevons’ paradox, the nineteenth-century economist’s observation that efficient steam engines increased total coal use, has found its definitive case.

Making them is the quieter half of the ledger. Fabs are among the most resource-intensive factories ever built: ultrapure water by the millions of liters daily, electricity in industrial quantities - TSMC alone consumed 6.4 percent of Taiwan’s power as early as 2021 [E1] - and a chemical inventory that includes persistent fluorinated gases with global-warming potencies thousands of times CO2’s. The industry abates most of these at the stack and has made genuine progress per wafer, but absolute footprints grow with the boom, and the geographic concentration that Chapter 35 described is also an environmental concentration: the water and power demands of the world’s most advanced chips fall disproportionately on one drought-prone island. The transistor’s eighth decade thus presents the industry with the same shape of problem it solved in 1947 - replace a profligate technology with a frugal one - but this time the incumbent is itself, and the replacement must come from within its own physics and economics. Whether the next seventy years of progress can be paid for in watts the planet can spare is not a footnote to this book’s subject. It is, increasingly, the subject.

# Chapter 42 - The connected system

Step back from the device, the factory, and the market, and what remains is the hardest thing to see precisely because it is everywhere: the transistor is no longer a component inside a system. It is the substrate on which the systems run. This chapter takes one pass across the connected whole, to record what the previous forty-one chapters add up to.

Consider what a single ordinary day now routes through the device. The alarm is a chip; so is the phone it rings on, the network that feeds it, and the grid that charges it. Breakfast arrives through refrigerated logistics directed by computers; the car or train that carries you to work is a network of processors on wheels - an electric one carries more than twice the semiconductor content of its combustion ancestor, and the silicon content of a vehicle roughly doubled in four years around the early 2020s [B2]. The office is a terminal of a planetary computation fabric; the hospital's imaging machines, the pharmacy's supply chain, the bank's ledger, the farmer's weather forecast, the satellite overhead - each is a colony of transistors, and the colonies are now connected into a single metabolism. Humanity passed, decades ago, an invisible milestone: the number of transistors manufactured per year exceeded the number of grains of wheat harvested, of letters written, of most countable things our species makes. They are the most numerous artifact in history, and the cheapest: a dollar buys tens of millions of them.

The system-level consequences have been this book's recurring theme, and they can be stated plainly here at the end. The transistor created an industry with an unusual property - its product improves exponentially while its price collapses - and that property pulled the world's information into digital form, which made the industry indispensable, which funded the next doubling. Around that flywheel grew the structures this book has described: the venture-capital model born at Fairchild, the foundry division of labor, the equipment oligopolies, the Asian manufacturing heartland, the research consortia, the roadmaps, and finally the industrial policies of nations. Each ring of the system fed the next, and the whole now has the property its inventors never planned: no part of it can be removed without the rest failing. The internet without chips is nothing; chips without the global supply web are nothing; the web without the equipment makers is nothing; the equipment makers without the frontier fabs are nothing. The transistor's ultimate product is interdependence itself.

That is the vantage from which the controversies of the present - the chip war, the subsidies, the chokepoints, the energy demands - look less like anomalies than like the predictable late stage of a long process. A technology that reorganizes civilization will eventually be governed as civilization's infrastructure. The transistor spent its first

sixty years as a product. It is spending its next decades becoming something closer to territory: contested, mapped, fortified, and shared.

# Chapter 43 - The long arc

This book began in a laboratory in New Jersey in the last days of 1947, with a sliver of germanium, two strips of gold foil, and a team trying to build a better amplifier for the telephone network. It ends here, in a world where the descendants of that device are manufactured by the quintillion, hold up the economy, focus the anxieties of superpowers, and press against the limits of thermodynamics and the atom. The long arc between those points has a shape, and the shape is the book’s argument.

The arc bends through four transformations. First, the physics: a quantum-mechanical effect in an unpromising material, understood deeply because a great industrial laboratory was free to understand things deeply, became a controllable device. Second, the manufacture: the device learned to be printed - on silicon, by the thousand, then the billion - and printing turned a laboratory curiosity into the cheapest artifact ever made. Third, the economics: cheap switching organized itself into an exponential, and the exponential organized an industry around a self-fulfilling schedule that ran for sixty years. Fourth, the politics: when the exponential concentrated its manufacture into a few firms and one island, the device escaped the market that made it and became an object of statecraft. Physics begat factories, factories begat abundance, abundance begat dependence, dependence begat power. That sequence, more than any single invention, is the transistor’s story.

The arc also carries warnings the record supports. Genius mattered - Bardeen’s theory, Shockley’s junction, Hoerni’s planar insight, Noyce’s and Kilby’s integrations, Dennard’s arithmetic - but every genius in this book stood on an institution, and the institutions were built by policy: the Bell monopoly’s research tax, the consent decree that gave the invention away, the military’s patient early purchases, the Taiwanese state’s blank check. The lone-inventor myth is not only false but misleading, because it hides the lesson the true story teaches: that foundational technology grows from patient, public-minded investment in understanding, years before the market can see the return. The nations now spending tens of billions to rebuild chip capacity are relearning the second half of that lesson. The first half - that the Bell Labs of the story was a place where people were paid to understand germanium, not to ship product - is the harder half, and the more often forgotten.

What comes next, the book has been careful not to prophecy. The roadmap still has chapters in it: nanosheets and stacked transistors, backside power, packaging that turns chips into cities, and beyond them the strange menagerie of tunneling, spin, and light. The transistor will be displaced someday, the way it displaced the tube, or it will metastasize into its successors so gradually that no one will mark the day. Either way, the pattern the device established - that understanding matter deeply enough lets you print civilization’s substrate, and that whoever prints it holds power no charter granted

- is now permanent. The long arc of the transistor is the arc of the modern world learning what it is made of. It bends, as it always has, through the material, toward whoever understands it best.

# Chapter 44 - A transistor glossary

Plain-language definitions of the terms this book uses, gathered here for reference.

**Band gap.** The energy difference between a material’s filled and empty electron states; the property that makes a semiconductor a semiconductor. Silicon’s is about 1.1 electron-volts, germanium’s 0.67.

**Bipolar transistor (BJT).** The transistor family descended from the 1947 invention, in which both electrons and holes carry current. Superb for amplifying; largely displaced in logic by the MOSFET.

**Boltzmann tyranny.** The thermodynamic floor on a conventional transistor’s switching sharpness - about 60 millivolts of gate voltage per decade of current at room temperature - which limits how low supply voltage can go.

**Channel.** The region of a transistor whose conductivity the gate controls; the bottleneck the whole device exists to regulate.

**Chip (die, integrated circuit).** A piece of semiconductor bearing many interconnected devices fabricated together; “chip” properly names the physical die, “integrated circuit” the circuit on it.

**Chiplet.** A partial system on its own die, combined with others through advanced packaging into a single functioning whole.

**Cleanroom.** The particle-controlled environment in which wafers are fabricated; modern fabs measure allowable dust in counts far below a surgical theater’s.

**CMOS.** Complementary MOS: circuits pairing n-type and p-type transistors so that essentially no current flows except during switching. The basis of all modern low-power logic.

**Dennard scaling.** The 1974 observation that shrinking a transistor’s dimensions and voltage together keeps power density constant - the license that made Moore’s law affordable until it expired around 2005.

**DRAM.** Dynamic random-access memory: one transistor and one capacitor per bit, dense and cheap, but leaky - every cell must be refreshed thousands of times a second.

**Doping.** Deliberately introducing impurity atoms (parts per billion to parts per million) to give silicon its useful n-type or p-type character.

**EUV lithography.** Printing with 13.5-nanometer-wavelength light generated by vaporizing tin droplets with lasers; required for the most advanced chips, and supplied only by ASML.

**Fab (foundry).** A semiconductor fabrication plant. A “foundry” is a fab that manufactures other companies’ designs; an “IDM” (integrated device manufacturer) designs and makes its own.

**FinFET / gate-all-around (GAA).** The three-dimensional transistor generations: the channel raised as a fin gripped on three sides (FinFET, from 2011), then as stacked nanosheets gripped on all four (GAA, from 2022).

**Flash memory.** Nonvolatile storage that traps charge on a floating gate; NOR and NAND are its two architectures.

**IGBT.** Insulated-gate bipolar transistor: a MOS-gated power switch combining easy control with heavy conduction; the standard valve of high-power electronics.

**Landauer limit.** The thermodynamic minimum energy, kT ln 2, irretrievably dissipated when one bit is erased.

**Mask.** The patterned plate through whose image a circuit layer is printed onto the wafer; a modern chip requires dozens.

**Moore’s law.** Gordon Moore’s 1965 observation, revised 1975, that the economical number of transistors per chip doubles about every two years - an economic regularity, not a law of nature.

**MOSFET.** Metal-oxide-semiconductor field-effect transistor: the insulated-gate device whose current is controlled by an electric field across a thin oxide. The most manufactured artifact in human history.

**Node.** A named generation of manufacturing technology (180nm, 7nm, 3nm...). Since the mid-1990s the number is a label, not a measured dimension.

**Photolithography.** The photographic process - coat, expose, develop, etch - by which circuit patterns are printed on the wafer; the productive heart of the fab.

**Planar process.** Jean Hoerni’s 1959 method of building transistors beneath a protective oxide layer on the flat wafer surface; the foundation of all modern manufacture.

**Semiconductor.** A material - above all silicon - whose conductivity lies between conductor and insulator and can be commanded by doping, fields, and light.

**Transistor count / yield.** The two numbers that rule the business: how many devices a chip carries, and what fraction of a wafer’s chips work.

**Wafer.** The polished single-crystal disc - 300 millimeters across in modern fabs - on which chips are printed by the hundreds at once.