

Semiconductors 101, unit 25

How small can transistors get? The story of "nanometers"

The news talks about "2 nm" and "3 nm" every day. Once you know where the number comes from, why each shrink gets harder and why transistors keep changing shape, you can see where chips are heading next.

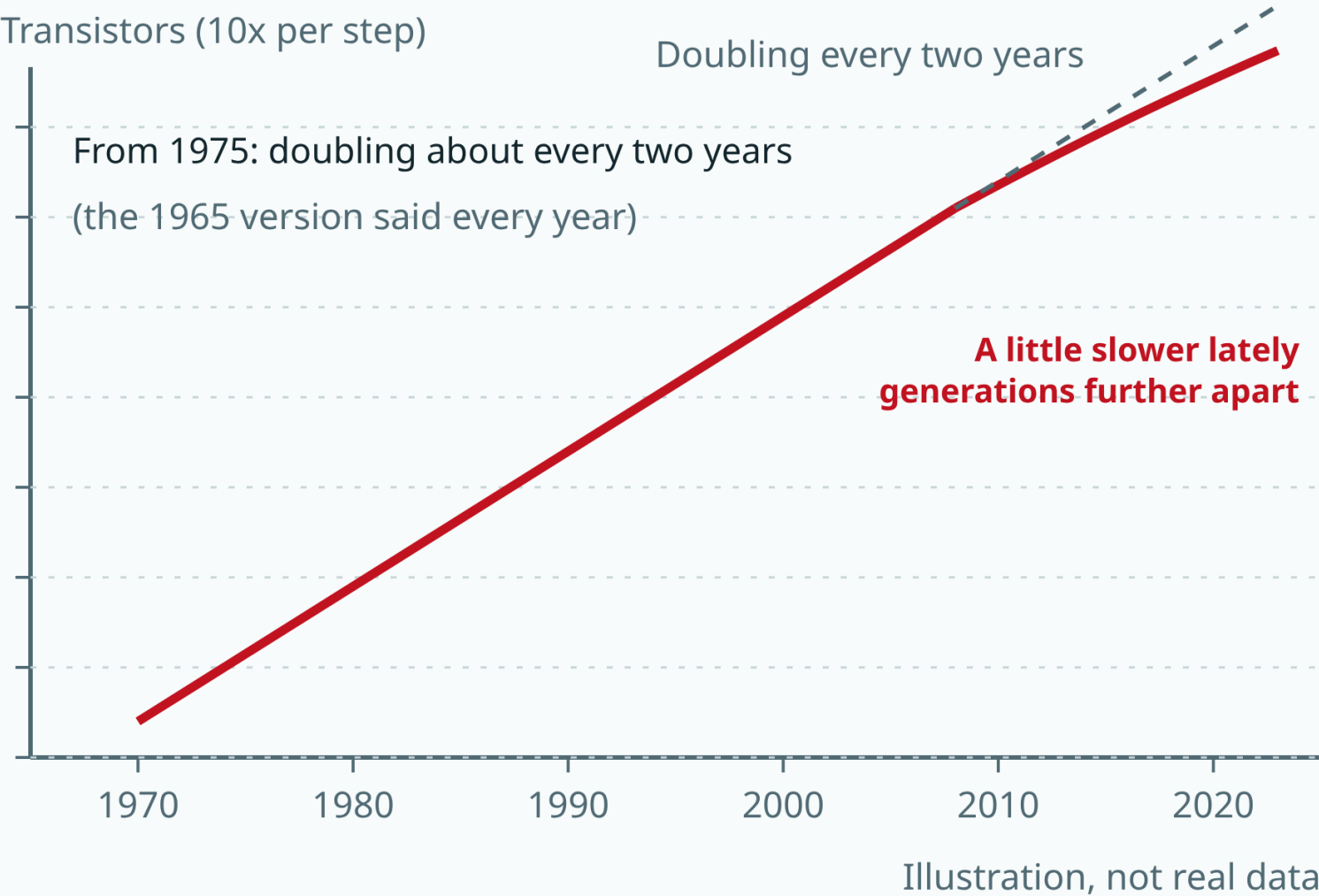
Semiconductors 101: a free course following the life of a chip

Free to use in class and for non-commercial teaching. Please credit fab24hr. Do not sell it or present it as your own work.

Goals

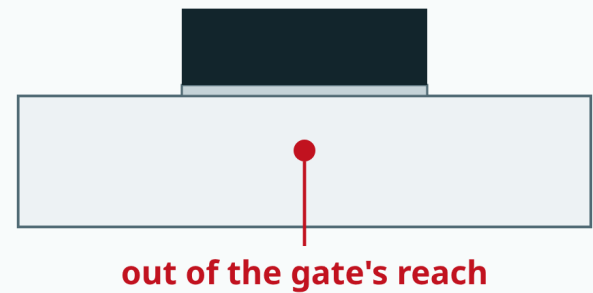
- 1 Know how small a nanometer is, and that a name like "2 nm" is the name of a process generation, not the length of anything on the chip.
- 2 Explain what Moore's law says, and why shrinking keeps getting harder: atoms, leakage and heat, slower wiring, and rising cost.
- 3 Explain why transistors went from flat to a standing fin (FinFET) and then to a gate wrapped around all four sides (GAA), and why power lines are starting to move to the back of the wafer.

Semiconductors 101, unit 25

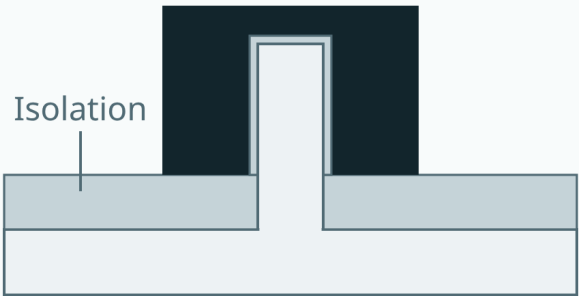


An illustrative line, not real data. The 1965 observation was a doubling about every year, revised in 1975 to about every two years; in recent years each generation has become harder and more expensive.

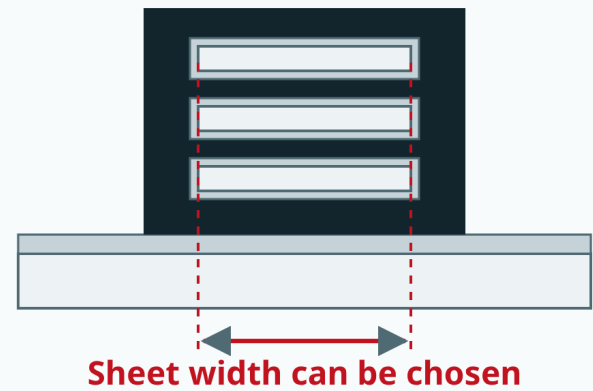
Flat: the gate covers only the top



FinFET: gripped on three sides



GAA: each sheet wrapped around



Next: two kinds of switch stacked



□ Silicon ■ Gate □ Insulator

Cut across the path of the current; a sketch, not to scale.

From flat, to standing up, to wrapped on all four sides: the more the gate surrounds the channel, the more tightly even a very small switch shuts. GAA sheets can be printed wider or narrower as needed, while fins can only be added one at a time. The last panel is still

research, only a sketch.

"So many nanometers": from a real size to the name of a generation

A nanometer is a billionth of a meter; a human hair is tens of thousands of nanometers thick. The whole way chips are made is called a process, and each generation has a name. Until roughly the 1990s, the name more or less matched a real size on the transistor, such as how narrow the finest line of that generation could be. Each generation aimed to fit twice as many switches in the same space: shrink the length and the width each to 0.7 times and the area is about half. So each name was the last one times 0.7: 90, 65, 45, 32, and on down to 3 and 2. Later, transistors could no longer get better just by shrinking everything in proportion, and new shapes and materials took over (more below). Each generation still held more than the last, but the name came to mean only which generation it is, and the same number does not always mean the same size from one maker to another, because there is no common standard. A public technology roadmap that the industry draws up together estimates that in the "3 nm" generation, transistors in a row are spaced in the forties of nanometers and the tightest metal lines in the twenties: far bigger than the name, and still more than a thousand times thinner than a hair.

Why does each shrink get harder?

Moore's law is an observation an engineer made in 1965: the number of components on a chip (later mostly transistors) was doubling about every year; in 1975 he changed it to about every two years. From around the 2010s, the pace slowed. First, sizes are close to atoms: a nanometer holds only a few silicon atoms, and an atom can't be cut in half. Second, a transistor that small can't shut off tightly, so a little current slips through even when it is off, which is called leakage; the voltage won't come down much further either, so the more tightly the switches are packed, the hotter the chip runs. Third, printing finer lines takes light of a shorter wavelength in the scanner, or splitting one layer into several exposures; the more steps, the more chances for something to go wrong, and the higher the cost. Fourth, the wiring: the thinner a wire, the higher its resistance, and the closer two wires are, the more they tug at each other and the slower the voltage changes. Making transistors smaller usually makes them faster, but making wires thinner makes them slower, so the more advanced the chip, the more of a signal's delay comes from the wiring.

Can't shut tightly? Change the shape: from flat, to standing, to wrapped all around

The part of a transistor that turns it on and off is called the gate. Transistors used to be flat: the channel the current flows through lay on the silicon surface, with the gate only on top. As they shrank, the entrance and exit moved closer and closer, the gate could no longer control the bottom of the channel, and even when it said "off", a little current slipped through, like pressing a garden hose from the top with one finger: water still leaks underneath. So the channel was stood up as a thin wall only a few nanometers wide (a fin), and the gate straddles it, gripping the top and both sides. This is the FinFET (fin field-effect transistor), which came into wide use around the early 2010s. Fins are hard to make any thinner, and as the gate gets shorter, three sides still can't quite control the very bottom, so the next change was gate-all-around (GAA): the channel becomes a few thin sheets stacked one above another, and the gate fills the gaps between them too, wrapping every sheet on all four sides, like wrapping your whole hand around the hose. Mass production only began in the 2020s, around the generations the news calls 3 nm and 2 nm. Further ahead, stacking the two kinds of switch that work as a pair (n-type and p-type) on top of each other, and making the channel from materials only a few atoms thick, are both still in research.

Power lines move to the back, and other roads

The wiring has become a bottleneck too. On the most advanced chips, a dozen or so layers of wiring sit above the transistors, and the power lines are crowded onto that side along with the wires that carry signals. The thinner a wire, the higher its resistance, so by the time power reaches a switch the voltage has already dropped a little, and the switch flips more slowly. To lose less, power lines have to be wide and many, which squeezes the signal wires into detours. So some of the newest generations have started to use backside power delivery (each maker does it a little differently): the transistors and signal wiring are built on the front as usual, the wafer is flipped and bonded to another wafer that holds it up, the original wafer is thinned from the back to around a micrometer or less, holes are dug to reach the transistors, and wide, short power lines are laid on the back. Less voltage is lost and the front is freed up for signals; the price is more steps, more cost, and heat that is harder to get out. Other roads are being taken at the same time: some memory chips stack their storage cells layer upon layer, advanced packaging puts several small chips together (see "Advanced packaging and new materials"), and the finest layers are printed with extreme ultraviolet light (EUV).

Common misconceptions

People say: "2 nm" means the transistors on the chip are only 2 nm across.

In fact: It is the name of a process generation, not the size of anything. By a public estimate, the tightest metal lines in the "3 nm" generation are still spaced 20-some nanometers apart, several times the number in the name. Some films on a chip really are only a few nanometers thick, but the name isn't measuring them either.

People say: Moore's law is dead, so chips will stop getting better.

In fact: People disagree, and it comes down to the definition. If it means "doubling every two years, with each transistor getting cheaper", that is hard to do now. If it only means "more in the same space each generation", it is still moving, just more slowly and at a higher cost. The industry has no agreement on this. And not every chip needs the newest generation anyway.

Play

**Play a shift as a litho process engineer:
look after every layer the scanner prints**

Scan with a phone to play from here

fab24hr.com/en/?job=lit&utm_source=slides



Think

- 1 A newer generation usually does the same work on a bit less power, or a bit faster on the same power, but it is also harder and more expensive to make. If you were making a chip that runs on a single button cell, would you go for the newest generation? What would you ask before deciding?
- 2 Process names stopped being real sizes long ago, yet each new one is still roughly the last one times 0.7 (with some half-steps in between). What is good about naming them this way, and what does it lead people to misunderstand? If it were up to you, how would you name them?

Read

What is Moore's law, and why do people say it's slowing down?

fab24hr.com/en/blog/moores-law/

What are FinFET and GAA? Why transistors went from flat, to standing up, to surrounded on all four sides

fab24hr.com/en/blog/finfet-and-gaa/

The news says "2 nm". What exactly is 2 nm?

fab24hr.com/en/blog/what-is-2nm/

Further

What is backside power delivery? Why chips are moving their power lines to the back of the wafer

Semiconductors 101: a free course following the life of a chip

<https://fab24hr.com/en/learn/scaling/>

Written from public textbooks, industry standards, university course notes and papers; the companies, people, machine numbers and figures in the game are fictional.

Free to use in class and for non-commercial teaching. Please credit fab24hr. Do not sell it or present it as your own work.

Generic illustration made with AI help from public sources; may contain mistakes; not any company's tool