

Semiconductors 101, unit 2

Transistors: tens of billions of switches

A chip can compute and remember because it holds a great many tiny switches controlled by electricity; a phone chip has more than ten billion. This unit goes from a single switch to processors and memory.

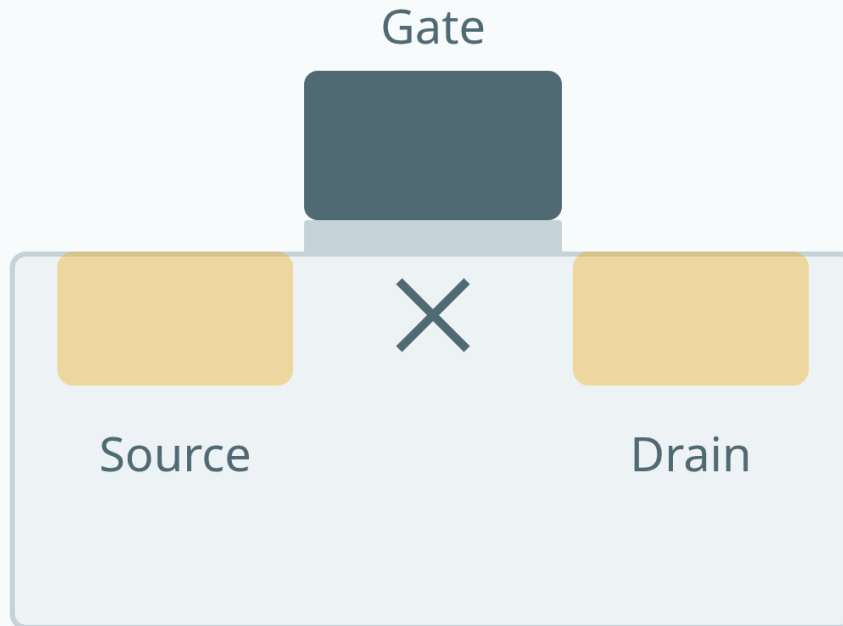
Semiconductors 101: a free course following the life of a chip

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

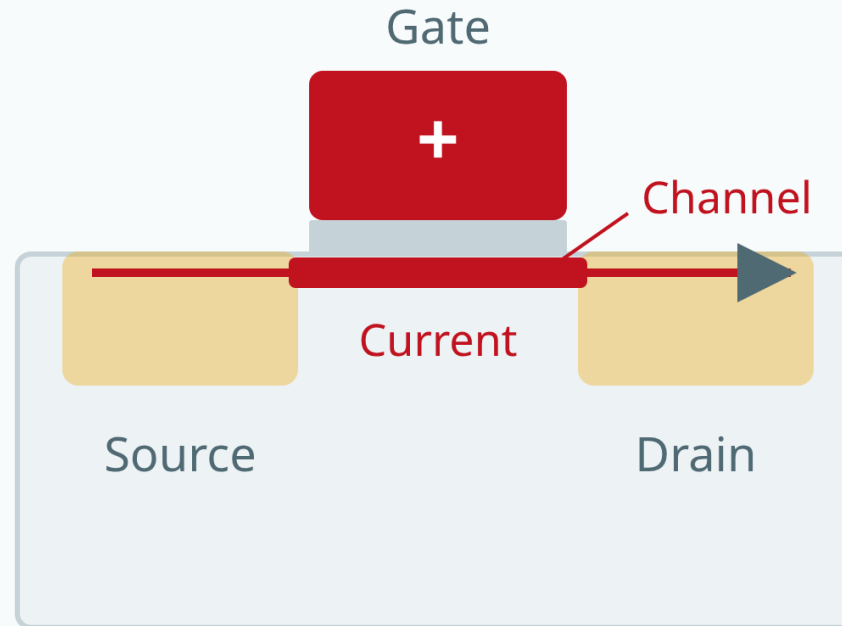
Goals

- 1 Explain what a transistor does using the idea of a switch, why it is normally off, and that a 0 or a 1 is really the voltage on a wire.
- 2 Explain why two kinds of switch are paired (CMOS), how switches in series and in parallel become logic gates, and how gates make a circuit that adds.
- 3 Describe what making transistors smaller gains, and why a chip uses power even when it is doing nothing.

Off: no current



On: current flows



Under the gate is a very thin insulating layer

A simplified cross-section. A very thin insulating layer separates the gate from the channel; the gate's voltage decides whether the channel opens.

A transistor is a gate run by electricity

Think of a transistor as a tiny waterway. One end is the entrance, called the source; the other end is the exit, called the drain; and over the channel between them sits a gate, with a very thin insulating layer in between so it doesn't touch the channel. In a common kind, the entrance and the exit are n-type silicon and the silicon under the middle is the opposite type, p-type (the doping unit explains this), so normally no current gets through and the switch is off. Put a voltage on the gate and it pulls electrons to the silicon surface, laying a temporary conducting path, so the switch is on; take the voltage away and the path disappears. A switch is controlled by electricity and passes on electricity, so one switch's output can drive the next switch.

Two kinds of switch in pairs: CMOS

There are two kinds of switch: one turns on only with a voltage on its gate (NMOS), the other the reverse, on only when its gate voltage is low (PMOS). The simplest circuit, the inverter, puts a PMOS between the output and the power supply and an NMOS between the output and ground, with their gates joined as the input. Input 1: the NMOS turns on and the output is connected to ground, a 0. Input 0: the PMOS turns on and the output is connected to the supply, a 1. You often hear "on is 1, off is 0"; that isn't accurate. In logic circuits such as processors, a 0 or a 1 is the voltage on a wire, and whether the switches are on or off decides whether that wire is connected to the supply or to ground. When nothing is switching, one of the pair is always off, so almost no current flows from supply to ground and little power is used. This pairing is called CMOS.

Switches in series and in parallel compute

Two switches in series make a path only if both are on; in parallel, if either one is. Using these two arrangements to set when the output is connected to ground and when to the supply gives a logic gate. For example, two in series on the ground side and two in parallel on the supply side make the output 0 only when both inputs are 1: that's NAND. Add an inverter after it and you get AND, "1 only if both are 1". Numbers in a chip are written in binary, where each digit is only 0 or 1. To add two one-digit numbers, the carry is AND and this digit is XOR, "1 only if the inputs differ"; add in the carry from the digit to the right as well, put a row of these side by side, and long numbers add up. Build on that the circuits that compare numbers and choose what comes next, add control circuits that direct them step by step through each instruction, and you have the core of a processor. Memory holds the program and the data: DRAM is fast but forgets when the power goes off; flash memory is much slower but remembers.

Smaller is faster and saves power, but leaks

Make the switches smaller and more fit in the same space; each flip needs less charge filled in and let go, and it switches faster. That's how a phone chip came to hold more than ten billion transistors. A nanometer is a billionth of a meter, or a millionth of a millimeter, and a hair is tens of thousands of nanometers thick; on the newest chips, transistors sit just a few tens of nanometers apart, more than a thousand times finer than a hair. But too small and they don't shut tightly: even when the gate says "off", a little current slips through. That's leakage. One switch leaks very little, but ten billion add up, so a chip uses power even while doing nothing. That's why transistors changed shape, from flat, to standing up, to having the gate wrap the channel on all four sides, so the gate grips it more firmly.

Common misconceptions

People say: The more RAM a phone has, the more photos it can store.

In fact: The RAM on a spec sheet is DRAM, which holds the programs and data in use right now and is wiped when the power goes off. Photos and apps are stored in flash memory, the "128 GB" or "256 GB" of storage. The two kinds of memory do different jobs, and the storage is often ten times the RAM or more.

People say: The "nanometers" in the news are the real size of the transistors.

In fact: The nanometer number says which generation of shrinking a chip belongs to. Names kept going down by roughly 0.7 each generation, but real dimensions did not shrink at that rate, and later transistors changed shape too, so the name no longer matches any real size.

Play

Start at chapter 10, "Standby Current": a chip that does nothing yet draws too much power

Scan with a phone to play from here

fab24hr.com/en/?story=10&utm_source=slides



Think

- 1 A phone left alone still slowly loses charge. Where does the energy go? Try to explain it with what you read in this unit.
- 2 One kind of memory forgets when the power is cut; another remembers. Why does a phone need both? Can you think of a similar division of labor in everyday life?

Read

What is a transistor, and why does a phone chip hold billions of them?

fab24hr.com/en/blog/what-is-transistor/

What is CMOS, and why do chip switches come in pairs?

fab24hr.com/en/blog/what-is-cmos/

What are logic gates, and how do switches become a chip that computes?

fab24hr.com/en/blog/logic-gates/

Further

How is memory different from a processor, and DRAM from flash?

Why do chips use power? Leakage and standby current, explained

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

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

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