

Semiconductors 101, unit 27

Making sense of chip news

"Mature-node shortage", "AI chips", "a major breakthrough": with what you have learned so far, you can read these headlines and spot where they jump too fast.

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 how mature nodes differ from advanced nodes, why older generations still matter and cost less, and why the news so often swings between shortage and oversupply.
- 2 Explain why an AI chip packs in so many simple calculating units, and how chips for training and for inference differ.
- 3 Tell apart "made in research", "in mass production" and "widespread" in the news, and explain why a new technology often takes ten years or more to go from a paper to mass production.

Mature nodes: older generations made for years

A mature node is an older generation of chipmaking that has been in mass production for years; the newest generations are called advanced nodes. There is no single line for how many nanometers counts as mature. One common view puts it at 28 nm and older, and the line moves forward over time. Many chips don't need the most switches. Chargers and power chips have to handle high voltage and large currents, so their switches and insulation must be bigger and thicker; analog chips, which handle smoothly changing signals such as sound or temperature, gain little from shrinking. And older methods cost less: the tools in older fabs are mostly paid off, there are fewer steps, and the process is well understood, so the share of good chips (the yield) is high.

Why do shortage and oversupply take turns?

Because demand changes fast and capacity changes slowly. Adding a line in an existing fab means getting tools, installing them and tuning them until they are stable, which often takes several months or more; a new fab takes even longer, counted in years. And even when tools are free, a wafer takes weeks to months from wafer start (going into the line) to finish. A car that is missing one small, cheap chip cannot be finished. Car chips also have to survive cold, heat and constant vibration for well over a decade, and pass the car industry's public standards ("automotive grade" in the news) one test at a time; from the chip's own tests to testing the whole car often takes several months, sometimes more than a year, and changing the process or moving to another fab means qualifying it again whenever quality could be affected, so a shortage is hard to fill quickly. In a shortage everyone adds lines, and they all come on a few years later; if demand has slowed by then, there is too much capacity, fabs compete for orders, and the news says "oversupply".

AI chips: not the fastest at one sum, but doing many at once

Using an AI model is a lot like adding up a bill: a huge pile of numbers multiplied by other numbers, then all added together. Each step is simple, but there are a vast number of them, and most can be done at the same time. An ordinary processor (CPU) has a few to over a hundred powerful cores, so the work waits in a long queue; an AI chip puts thousands to tens of thousands of simple calculating units to work at once. It needs room for many units and uses a lot of power, so it often uses the newest process generation; and since a chip can't grow forever and data has to arrive fast enough, advanced packaging puts several chips and stacked memory together in one package (see "Advanced packaging and new materials"). Training teaches the model and often takes many large chips working together for days to months; inference uses the trained model to answer questions and cares more about saving power and responding quickly.

When the news says "breakthrough": how far along is it?

The news says things like "researchers build a 1 nm transistor" or "a major breakthrough in 2D materials", and it sounds as if your phone will have them in a year or two. That "1 nm" is not the same thing as a generation name like "2 nm". In fact, a new chip technology often takes ten years or more to go from its first paper to mass production. Making one in a lab only proves the idea works. Then many have to be made across a whole wafer, all the same; the new step has to fit in with hundreds of other steps; production tools and materials have to catch up; the chip goes through pilot runs and reliability tests on the R&D line, then to the production fab, where the yield still has to climb step by step. At first, a new technology is often used only in the chips that need it most and can afford it, and it spreads as more makers use it. The FinFET, the transistor stood up as a fin, appeared in a university paper in the late 1990s and went into mass production in the early 2010s. So when you read the news, first tell which it is: "made in research", "in mass production" or "widespread".

Common misconceptions

People say: The newest process is best, and every chip will move to it sooner or later.

In fact: Mature and advanced nodes split the work; the new does not replace the old. Chips that need the most switches use advanced nodes; chips that must handle high voltage, last long and cost little use mature ones. A phone has both: the processor that does the computing often uses the newest generation, while many of the chips around it that manage power, charging and sensing are made on mature nodes.

People say: An AI chip is just a faster CPU.

In fact: It is closer to the opposite. A CPU is like a few senior accountants who can handle any kind of bill, suited to work that keeps changing direction; an AI chip is like a whole auditorium of people, each with a calculator, all pressing keys on the same command. A dedicated AI accelerator goes further still: it does far more multiply-and-adds on less power, but it is not flexible enough for other work.

Play

Play a shift as the furnace R&D process engineer on the R&D line: the game's hearing-aid chip, Glowmere, isn't made on the newest generation, yet its furnace conditions still have to be settled on the R&D line

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

Scan with a phone to play from here



Think

- 1 Find a recent chip news story and retell it to someone at home in three plain sentences. Which words did you learn in this course? How far along is what it describes: made in research, in mass production, or already widespread? Which claims seem exaggerated or unclear?
- 2 What kind of chips do a car, a phone and an AI server each need? If you could add only one more production line, which would you build it for first? Why?

Read

The news says "mature nodes". Why do older chips still matter?

fab24hr.com/en/blog/mature-nodes/

The news says "AI chips". What are they, and how are they different from ordinary processors?

fab24hr.com/en/blog/ai-chips/

How long does a new chip technology take to go from a paper to mass production?

fab24hr.com/en/blog/lab-to-fab/

Further

The news says "automotive chips". What makes them different?

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

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

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