

Semiconductors 101, unit 26

Advanced packaging and new materials: more than shrinking

Processors compute faster and faster, but memory can't deliver data fast enough to keep up, and this "memory wall" often leaves AI chips waiting.

Stacking memory right next to the processor, packaging chips together and sending data over longer distances as light all have to do with it; add new

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

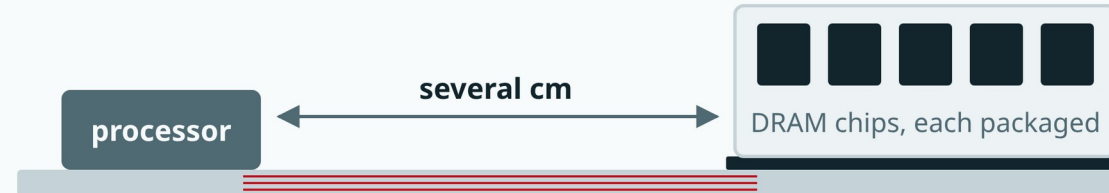
Free to use in class and for non-commercial teaching. Please credit fab24hr. Do not sell it or present it as your own work.

materials, and once you know these roads, the new terms in the news start to fit together.

Goals

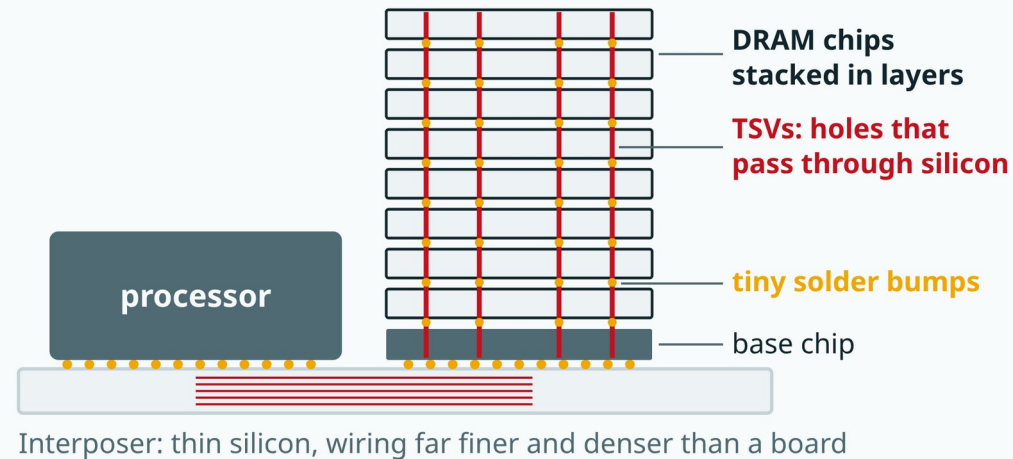
- 1 Explain what the memory wall is: why the processor so often stops to wait for data, why moving data costs more energy than computing, and why AI hits the wall hardest.
- 2 Explain what cache, HBM, chiplets, 2.5D and 3D each do, and know that these roads can be used together and that each has a price.
- 3 Tell which problem wide-bandgap materials and silicon photonics each solve, and know that neither is about replacing all silicon or all copper wiring.

Ordinary memory: plugged in, several cm from the processor



One memory module sends 64 zeros and ones at a time

HBM: stacked DRAM beside the processor on an interposer



One stack has 1,024 data lines, 16 times as many as a module
The newest generation doubles it again, to 2,048

HBM is still DRAM inside; what changes is how it is arranged and connected: stacked, connected through the silicon, and right next to the processor. With 16 times as many lines (twice that in the newest generation), much more data moves at once, and the shorter path uses less energy for each 0 or 1. Simplified: the numbers of layers, holes and lines are not drawn as they really are.

The memory wall: processors compute fast, but data can't arrive fast enough

The processor does the computing; memory holds the data to compute on. The two improve at different speeds: from the mid-1980s to the early 2000s, processors got roughly 50% or more faster every year, while the wait to read one piece of data from memory shrank by less than 10% a year. Memory chips care most about holding a lot and costing little; speed comes second. If data can't arrive fast enough, the processor can only stop and wait, however fast it computes, and this gap is called the memory wall. AI hits it hardest: a model's numbers are far too many to fit on the processor, so they sit in the memory outside. Every time a chatbot produces a little more text, a large share of them has to be read, and when it's working for just one person, each number brought in is used for about one multiplication and one addition before it's on to the next, so the computing circuits spend much of their time waiting. One study looked back over 20 years of servers: computing speed grew about 60,000 times, while the speed at which memory delivers data grew only about 100 times. Moving data also costs energy: every switch between 0 and 1 means charging or draining a whole wire, and reaching memory outside the chip is a trip of several centimeters. Researchers estimate that reading it once can take hundreds to thousands of times the energy of one simple addition.

Keep the data closer: cache and HBM

What the industry does is like running a kitchen: the cook is very fast, but the ingredients are in a warehouse far away, so keep what you use most within reach, move the warehouse closer, and use a bigger cart. Within reach means the cache, on the same chip as the processor: the numbers used most stay there, which is fast and saves energy. But cache is made of SRAM, which needs six transistors for each bit and takes up a lot of room, so a large AI model still won't fit. The wall people talked about early on was mostly "how long one wait takes"; today it is more often "how much can be sent every second". So several memory (DRAM) chips are stacked into one pile and placed next to the processor. That is HBM (high bandwidth memory): one stack has 1,024 data lines (the newest generation doubles that), while one memory module in an ordinary computer sends only 64 zeros and ones at a time. More lines send more at once, and the short trip takes less energy for each 0 or 1. The price: every layer has to be good and ground very thin, heat struggles to escape, and it is expensive and hard to make.

Packaging chips together: chiplets, 2.5D and 3D

What lets HBM sit right next to the processor is advanced packaging: a computing chip and its memory, or several computing chips, go into one package. So these roads are often used together: an AI chip made on the newest generation, with several HBM stacks beside it. There are two main ways to put them together: side by side on a thin slice with very fine wiring (an interposer, often made of silicon), called 2.5D; or stacked, connected by fine holes through the silicon filled with metal, called 3D. Compared with a few centimeters of wiring across a circuit board, the road is short and has many lanes, so data moves faster and uses less power. A big chip is also often split into a few small pieces, called chiplets: the bigger a chip, the more likely a defect on the wafer lands on it, and the lower the share that works (the yield); and the scanner prints one field at a time, while the largest chips are already about one field in size. Each piece is made and tested on its own and then put together: the computing piece on the newest generation, and pieces that would hardly get smaller on a newer one on older generations that are cheaper and well proven. The price: packaging gets harder and more expensive, and if one piece inside is bad, the whole set is wasted; stacked chips have to be ground very thin, and heat struggles to escape from the layers in the middle.

Two more roads: light for longer trips, and new materials

The faster data is sent, the shorter the distance copper can carry it, and the more energy it takes to restore the signal, so on longer trips, such as between racks of servers, electricity is turned into light and sent through optical fiber. Silicon photonics uses chipmaking methods to build thin paths for light on silicon, plus small parts that turn electricity into light and light back into electricity; co-packaged optics (CPO) then moves the electricity-to-light step into the same package as the chip that forwards the data, so electricity travels only a few millimeters to a few centimeters before it becomes light. New materials solve a different problem: silicon carbide and gallium nitride have a bandgap (the "step" an electron has to jump up before it can conduct) about three times as high as silicon's, which is why they are called wide-bandgap materials; they take higher temperatures and voltages. Used as switches that handle power, they waste less energy, and they go into electric cars and fast chargers; but the crystals are hard to grow, and processors and memory can't use the advantage, so most chips are still made of silicon. Some roads are mostly still in research or just getting started: building simple computing into the memory (the cheapest move is no move at all), or using materials such as gallium oxide and diamond, which the news sometimes calls "ultra-wide-bandgap" or "fourth-generation semiconductors".

Common misconceptions

People say: Wide-bandgap ("third-generation") semiconductors are a newer generation of chipmaking than "2 nm".

In fact: "Third generation", as Chinese-language news calls them, is about the material chips are made from, usually silicon carbide and gallium nitride, not a newer generation of silicon chipmaking. "2 nm" is about how many switches fit in the same space; silicon carbide and gallium nitride are about handling high voltage, large currents and heat. Power switches have to let large currents through, so they are big to begin with and don't need the finest lines. The two solve different problems; neither is more advanced than the other.

People say: The news says "replacing copper with light", so in the future, chips will compute with light inside.

In fact: The silicon photonics in the news is mostly about sending data, not computing with light; the computing inside chips is still done by transistors. "Replacing copper with light" usually means the links that keep getting faster and longer move to light; inside chips, and between chips sitting very close together, it is still mostly electricity. Light has its difficulties too: silicon is bad at giving off light, fibers must line up very precisely, and lasers are sensitive to heat.

Play

**Start at chapter 11, "The Ball Lifted":
walk the assembly line with Shu-Hui and
watch the chip go into one small
module along with resistors and
capacitors**

Scan with a phone to play from here



Think

- 1 Splitting a big chip into chiplets may improve yield and put the money where it matters most; but it needs extra room for connections, the packaging gets harder, and if one piece inside is bad, the whole set is wasted. If it were up to you, what kind of chip would be worth splitting? What kind would you keep in one piece?
- 2 To get past the memory wall, you can add more cache or put more HBM stacks beside the processor, and some researchers are working on computing inside the memory; every road has a price. If you were building a machine that only answers questions with AI, which road would you spend money on first? What would you want to know before deciding?

Read

What is the memory wall, and why can't data keep up with fast chips?

fab24hr.com/en/blog/memory-wall/

The news says "advanced packaging". What is it, and why do chips sit side by side or stack up?

fab24hr.com/en/blog/advanced-packaging/

The news says "HBM". What is it, and why does AI need special memory?

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

Further

The news says "wide-bandgap semiconductors". What are SiC and GaN?

The news says "silicon photonics". What is it, and why send data with light instead of electricity?

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

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

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