Semiconductor Technology from A to Z

Advanced Packaging / 3D Integration

Limits of Classical Integration

Why “More Moore” Is Reaching Its Limits

For decades, the semiconductor industry followed Moore's law: the number of transistors per chip doubled roughly every 18–24 months, primarily through shrinking feature sizes. At feature sizes in the single-digit nanometer range, however, this scaling runs into physical and economic limits. Leakage currents rise due to short-channel effects, lithography cost per exposure step is exploding (EUV systems can cost well over 200 million euros), and yield drops disproportionately for large monolithic chips, since the probability of a defect grows with chip area.

At the same time, demand for computing power – driven not least by AI accelerators – calls for ever more transistors per system. The industry's answer increasingly lies not in smaller transistors but in a smarter arrangement and connection of multiple chips: advanced packaging.

“More than Moore” as an Alternative

Instead of integrating every function of a system onto a single monolithic die, individual functional blocks – processor cores, memory, I/O, analog circuitry – are implemented on separate chips, often manufactured in different technologies, and then tightly coupled into an overall system. This is known as heterogeneous integration. This strategy allows each building block to be manufactured in the technology best suited and most cost-effective for it (e.g. logic at 3 nm, memory in an older, cheaper node), instead of accepting compromises for a single manufacturing process.

The challenge thus shifts from the transistor to the interconnect technology between chips – from the question “how small can a transistor become?” to the question “how densely and how fast can two dies be connected to each other?”

Context: Classical Wire Bonding and Flip-Chip

The techniques covered in the assembly chapters – wire bonding and flip-chip – each connect a single die to a package or substrate. The electrical connections sit at the edge or on the chip surface and are limited in density: wire bonds typically achieve a pitch (spacing between adjacent connections) of 40–50 µm, while flip-chip bumps fall in the range of 100–150 µm.

Advanced packaging techniques such as TSV, interposers, and hybrid bonding

(covered in the following chapters) achieve interconnect densities that are orders of magnitude higher – down into the sub-micrometer range. Only this makes it possible to couple several dies so tightly that they behave electrically almost like a single monolithic chip.

Through-Silicon Via (TSV)

The Principle

A through-silicon via is a vertical, electrically conductive connection that runs completely through the silicon substrate of a die. Unlike classical wiring layers, which sit exclusively on the chip surface, the TSV opens up the third dimension: signals and supply voltages can be routed from the front side of a die directly to the back side – and thus to the next die in a stack – without the detour via bond wires at the chip edge.

Fabrication Process

The process broadly breaks down into three approaches, which differ in when the via is created relative to the rest of the fabrication flow:

- Via-first: the TSV is etched into the bare wafer before transistor fabrication. This approach allows high-temperature-stable fill materials such as polycrystalline silicon, but is difficult to reconcile with modern CMOS processes.
- Via-middle: the TSV is created after transistor fabrication (FEOL) but before metallization (BEOL). This is the industrially dominant approach today, as it leaves the established transistor process untouched.
- Via-last: the TSV is etched only after the chip is fully fabricated, sometimes even from the wafer back side. The advantage is complete decoupling from front-side processing.

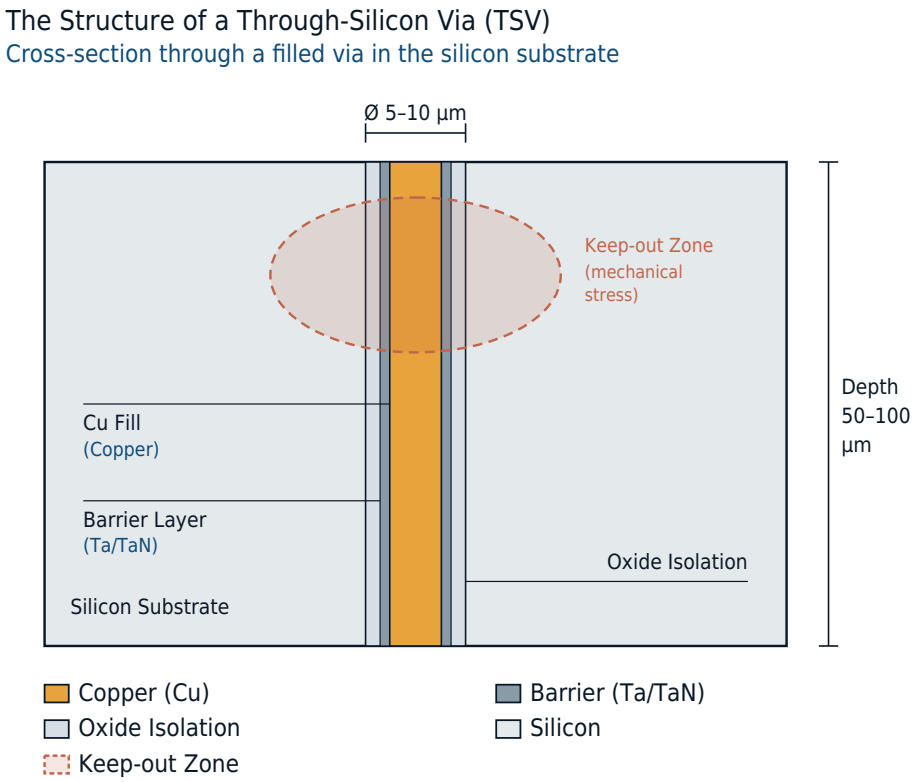
The via itself is etched using deep reactive ion etching (Bosch process, see the dry etching chapter) and reaches aspect ratios of 10:1 or higher – typical diameters are 5–10 µm, at depths of 50–100 µm.

Isolation and Cu Fill

Since the surrounding silicon is itself semiconducting, the via wall must first be electrically isolated with a thin oxide layer (usually via PECVD, see the deposition chapter). This is followed by a barrier layer (typically Ta/TaN) that

prevents copper from diffusing into the silicon – the same issue encountered with copper metallization (see the copper technology chapter). Only after that is the via electrochemically filled with copper.

Structure of a TSV in cross-section



Challenges of the Aspect Ratio

The higher the aspect ratio, the more difficult it becomes to achieve complete, void-free copper filling – similar to the gap-fill problem in CVD processes, only on a much larger scale. In addition, because the thermal expansion coefficient of copper differs strongly from that of silicon, the filled-in copper creates mechanical stress in the surrounding material, which can degrade transistor performance in the immediate vicinity of the TSV (the “keep-out zone”).

2.5D Integration: The Interposer

The Basic Principle

In 2.5D integration, several fully processed dies are not stacked directly on top of one another but placed side by side on a shared carrier substrate – the interposer. The interposer itself contains no active transistors; it serves purely as a high-density wiring layer that connects the individual dies to one another and to the package substrate beneath. The name “2.5D” reflects that an additional vertical layer is added, while the actual chip integration still happens laterally (side by side).

Silicon vs. Organic Interposer

- Silicon interposer: manufactured from a silicon wafer, penetrated by TSVs for through-connections and several metallization layers (RDL, see below) on top. Silicon allows very fine wiring structures (sub-micrometer pitch), but is comparatively expensive to produce and limited in area (typically to the size of a lithography reticle).
- Organic interposer: based on printed-circuit-board-like materials, significantly cheaper and producible in larger areas, but only achieves coarser wiring densities.

Redistribution Layer (RDL)

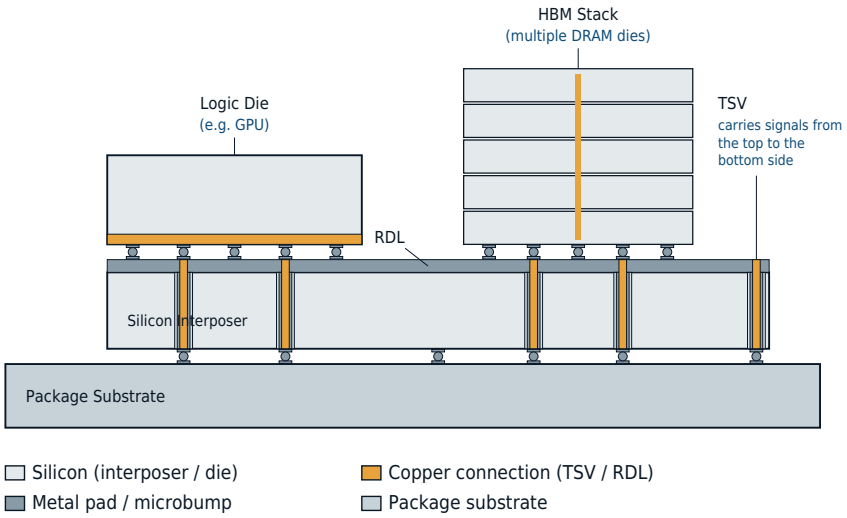
The redistribution layer is an additional metallization layer on the interposer (or directly on a die) that “redistributes” a chip's comparatively coarse contact pads onto the much finer connection structures of the interposer – hence the name. RDL structures are manufactured with lithography and copper technologies similar to the BEOL wiring of a regular chip, but with coarser feature widths and thicker traces, since current-carrying capacity matters more here than maximum density.

Example: HBM on an Interposer

The most prominent application example is the pairing of a logic die (e.g. a GPU) with several high-bandwidth-memory stacks on a shared silicon interposer. The extremely short, dense connection paths across the interposer enable memory bandwidths that classical PCB wiring could not achieve (covered in more depth in chapter 6).

Interposer with logic die and HBM stack

2.5D Integration: Interposer with Logic Die and HBM Stack
Cross-section of a shared silicon interposer



3D Stacking

Die-to-Die vs. Wafer-to-Wafer Bonding

While 2.5D integration arranges dies side by side, true 3D stacking places several dies directly on top of one another and connects them vertically via TSVs. Two fundamental manufacturing approaches are available:

- Die-to-die (D2D): individual dies that have already been singulated and tested are stacked on top of each other. Advantage: before stacking, each die can be individually verified as functional (“known good die”), which significantly improves the overall yield of the stack.
- Wafer-to-wafer (W2W): two complete, unsingulated wafers are bonded together as a whole and only singulated afterward. Advantage: significantly higher throughput and placement accuracy; disadvantage: a defective die on either wafer causes the entire stacked device at that position to fail, since no pre-selection is possible.

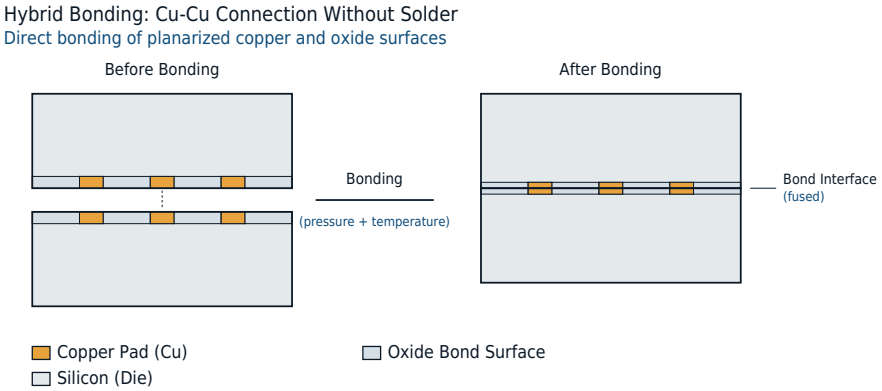
A hybrid approach, die-to-wafer (D2W), combines the advantages of both methods: tested individual dies are stacked onto a complete wafer.

Hybrid Bonding: Direct Cu-Cu

The densest form of 3D interconnect currently available is hybrid bonding. Here,

both surfaces to be joined have copper areas (for the electrical connection) and oxide areas (for mechanical bonding and isolation) planarized with high precision at the same time – using chemical-mechanical polishing, see the corresponding chapter – and are then joined directly, without any additional solder or bump material. Under pressure and moderate temperature, the oxide surfaces first fuse chemically; the copper then expands thermally and forms a direct metallic connection. This makes it possible to achieve interconnect pitches below one micrometer – an order of magnitude unreachable with classical solder or bump processes.

Hybrid bonding before and after the bonding process



Thermal and Mechanical Challenges

Stacked dies generate heat during operation that must be conducted away through several active layers – a die in the middle of a stack has a significantly worse thermal path to the heat sink than a single die sitting flat on a substrate. In addition, mechanical stress caused by differing thermal expansion coefficients of the stacked materials increases with every additional layer. Both factors limit, in practice, how many active, high-power logic dies can reasonably be stacked on top of each other, while passive or low-power memory layers (as in HBM) are far less critical.

Chiplets & Heterogeneous Integration

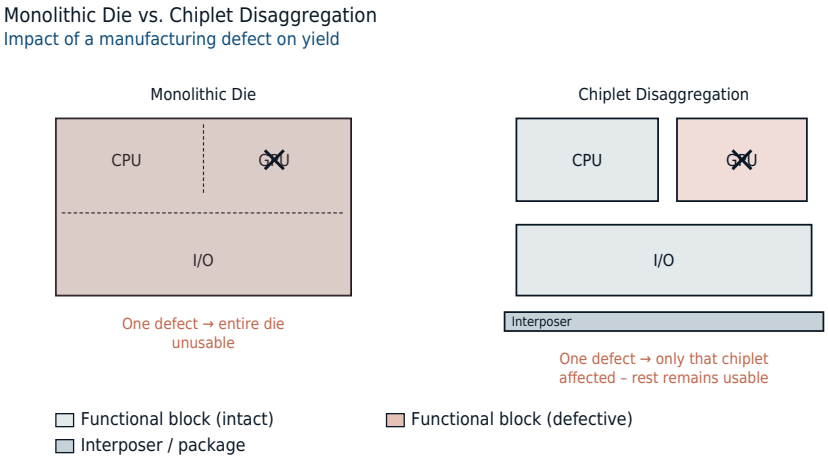
Disaggregation of SoCs

A classical system-on-chip (SoC) integrates every functional block – CPU cores, GPU, I/O controllers, memory interfaces – onto a single monolithic die. The chiplet approach deliberately breaks this SoC apart into several smaller, functionally distinct dies (“disaggregation”), which are then reassembled into a functionally coherent system using the advanced packaging techniques described in the previous chapters (interposer, hybrid bonding).

Advantages: Yield and Cost

The central advantage lies in manufacturing yield: since a chip's defect probability grows disproportionately with its area, the yield of one large monolithic die drops drastically faster than that of several smaller dies with the same total area. A defective small chiplet is discarded while the rest of the system can still be used – with a monolithic chip, the same defect causes total failure. In addition, as mentioned in chapter 1, each chiplet can be produced in whichever process node is economically and technically optimal for it, instead of committing the entire chip to the most expensive process it needs.

Impact of a defect: monolithic die vs. chiplet disaggregation



Disadvantages: Interconnect Overhead

These advantages come with a downside: the connection between chiplets – even via the densest advanced packaging techniques – has higher latency and higher energy consumption per transmitted bit than on-chip wiring within a

monolithic die. System architects must therefore carefully weigh which functional blocks are coupled tightly enough to benefit from monolithic integration, and which are suited to being split into separate chiplets.

Standardization: UCIe

For chiplets from different manufacturers to be combined into a shared system, standardized physical and protocol-level interfaces are needed. The industry standard Universal Chiplet Interconnect Express (UCIe) defines both the electrical layer (signal levels, pitch, bump arrangement) and the transfer protocol, with the goal of enabling a vendor-independent “chiplet marketplace” analogous to today's standard IC procurement.

Applications & Outlook

HBM as a Flagship Example

High Bandwidth Memory is the most consistent commercial implementation of the technologies covered in this section: several DRAM dies (see the DRAM cell chapter) are connected via TSV and wafer-to-wafer or die-to-wafer bonding into a vertical stack of eight, twelve, or more layers, and then coupled to a logic die via a silicon interposer. The resulting memory bandwidth far exceeds that of classical DRAM modules wired planarly on a circuit board, while also achieving significantly lower energy consumption per transmitted bit.

AI Accelerators

Modern AI training accelerators are the main driver behind the packaging technologies described here: they typically combine several logic chiplets with surrounding HBM stacks on a shared interposer, maximizing both the memory bandwidth critical to AI workloads and compute density, without depending on a single, uneconomically large monolithic die.

Foundry Comparison

Different foundries have established their own brand names for their advanced packaging platforms, each combining and marketing the fundamental principles described in this section – interposer, TSV, hybrid bonding – in different ways. These platforms differ, among other things, in the choice of silicon versus

organic interposer, the supported TSV aspect ratio, and the maximum number of dies that can be combined.

Outlook

The trend clearly points toward ever tighter, ever more finely resolved 3D integration – from TSV-based stacks toward hybrid bonding with steadily shrinking pitch. At the same time, heat dissipation and cost remain the limiting factors, which is why advanced packaging today functions less as an outright replacement for classical transistor scaling and more as a necessary complement to it.