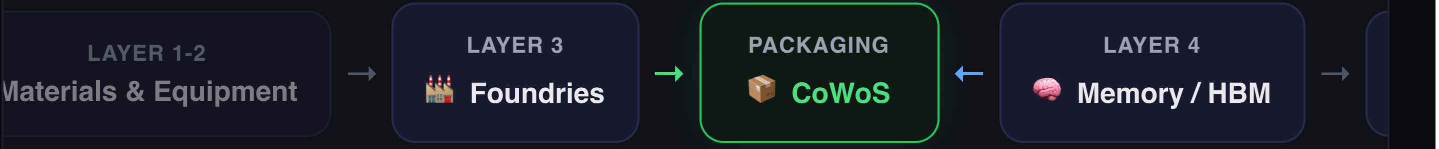
CoWoS & Advanced Packaging

The hidden bottleneck between making chips and shipping them — where GPU dies meet HBM

🗺️ Where CoWoS Sits in the AI Value Chain

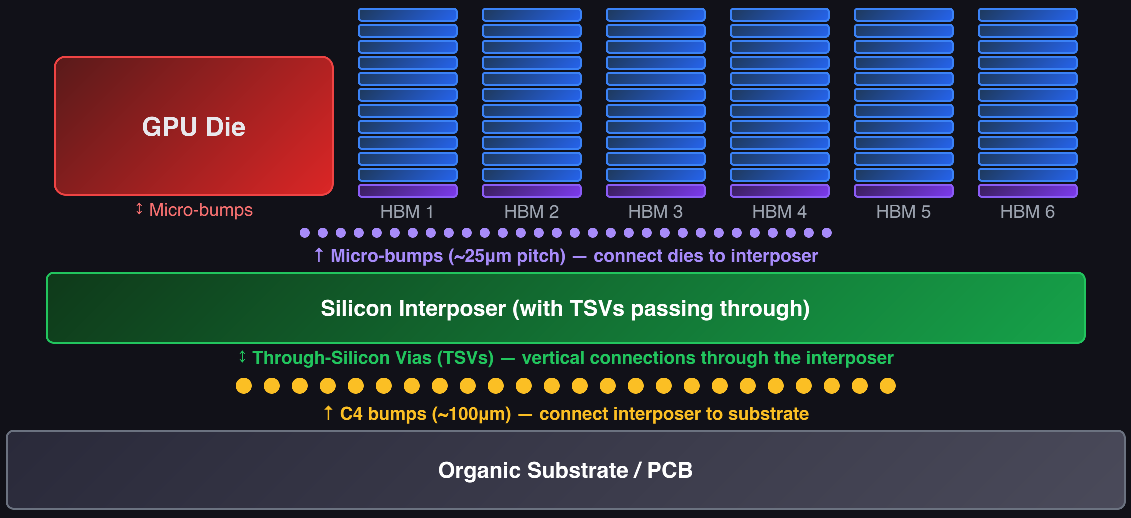
CoWoS is the **packaging step** — it takes a finished GPU die (from the foundry) and finished HBM stacks (from memory makers) and combines them into one working chip package. Without it, you just have separate pieces that can't talk to each other.



↑ CoWoS merges the GPU die (Layer 3) + HBM stacks (Layer 4) into one package → Layer 5

🔬 What CoWoS Actually Is — Cross-Section View

Chip-**o**n-**W**afer-**o**n-**S**ubstrate. The GPU die and HBM stacks are placed on a silicon interposer, which acts as a "bridge" connecting them with ultra-dense wiring. The interposer then sits on an organic substrate (PCB).

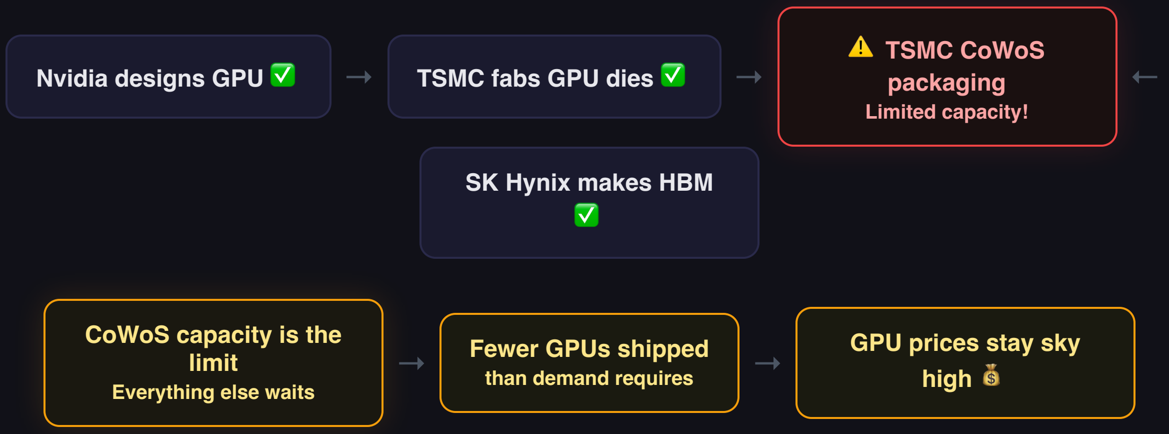

GPU Die
↕ Micro-bumps
HBM 1
HBM 2
HBM 3
HBM 4
HBM 5
HBM 6
↑ Micro-bumps (~25µm pitch) — connect dies to interposer
Silicon Interposer (with TSVs passing through)
↕ Through-Silicon Vias (TSVs) — vertical connections through the interposer
↑ C4 bumps (~100µm) — connect interposer to substrate
Organic Substrate / PCB

🟥 GPU Die 🟦 HBM DRAM Dies 🟪 HBM Logic Die 🟩 Silicon Interposer ⬛ Substrate/PCB ● Micro-bumps

● C4 Bumps

🚧 Why CoWoS Is the Bottleneck

TSMC is essentially the **only company on Earth** that can do advanced CoWoS packaging at the scale needed for AI chips. No matter how many GPU dies Nvidia designs or how many HBM stacks SK Hynix makes — they all have to pass through TSMC's CoWoS lines. It's a literal bottleneck.

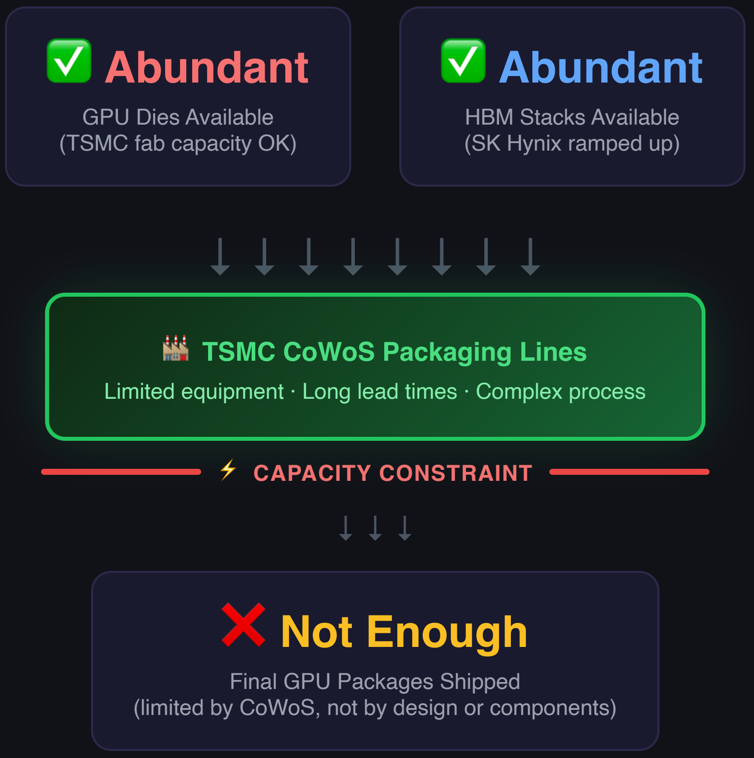


Nvidia designs GPU ✅ → TSMC fabs GPU dies ✅ → ⚠️ TSMC CoWoS packaging
Limited capacity! ←

SK Hynix makes HBM ✅

CoWoS capacity is the limit
Everything else waits → Fewer GPUs shipped
than demand requires → GPU prices stay sky high 💰

The Bottleneck Funnel



✅ Abundant
GPU Dies Available
(TSMC fab capacity OK)

✅ Abundant
HBM Stacks Available
(SK Hynix ramped up)

🏭 TSMC CoWoS Packaging Lines
Limited equipment · Long lead times · Complex process

⚡ CAPACITY CONSTRAINT

❌ Not Enough
Final GPU Packages Shipped
(limited by CoWoS, not by design or components)

📈 CoWoS Evolution — Getting Bigger and Bigger

As AI chips grow larger and need more HBM, the interposer must grow too. Each generation pushes the boundaries of what's physically possible in packaging.

CoWoS-S

2012 → present (original)

Si interposer

Silicon interposer, single reticle size (~~800mm^2). Used for H100 and earlier GPUs. Limited by reticle — the interposer can only be as big as what the lithography machine can expose in one shot.

Used in: Nvidia H100, A100
Interposer: ~1,700mm^2
HBM stacks: Up to 6

CoWoS-L NOW

2024 → present (larger)

Large Si + LSI chips

Uses local silicon interconnect (LSI) chiplets to stitch together a **larger-than-reticle** interposer. Enables much bigger packages with more HBM and even multi-die GPUs.

Used in: Nvidia B200, B300
Interposer: ~3,300mm^2+
HBM stacks: 8 or more
Key change: 2× the area of CoWoS-S

CoWoS-R FUTURE

~2026+ (next-gen)

RDL interposer

Replaces the expensive silicon interposer with an **organic RDL (Redistribution Layer)**. Much cheaper and easier to scale, but slightly lower interconnect density. Likely for cost-optimized products.

Used in: TBD (rumored mid-range AI chips)
Interposer: Organic RDL
Cost: Significantly cheaper
Tradeoff: Lower density vs silicon

VARIANT	INTERPOSER TYPE	MAX AREA	COST	DENSITY	STATUS
CoWoS-S	Silicon (single reticle)	~1,700mm^2	High	Highest	Mature
CoWoS-L	**Silicon + LSI chiplets**	**~3,300mm^2+**	**Very High**	**High**	**Ramping**
CoWoS-R	Organic RDL	Flexible	Lower	Medium	Development

🔢 Key Numbers to Remember

2×
TSMC CoWoS capacity roughly doubled in 2025 — still not enough to meet demand

$1-2K
Estimated cost CoWoS adds to each GPU package (on top of die + HBM costs)

~2×
B200 interposer is ~2× the area of H100's — uses more CoWoS capacity per chip

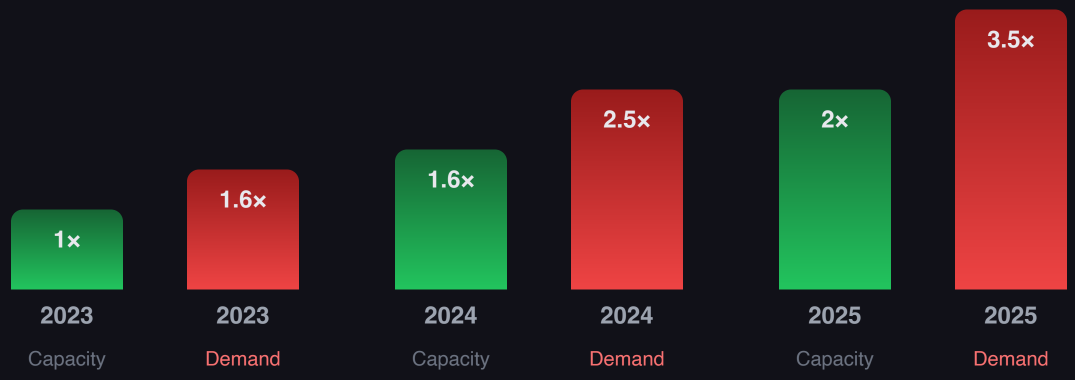
$B+
TSMC has invested billions in new CoWoS production lines in Taiwan & Japan

90%+
TSMC's share of advanced 2.5D/CoWoS packaging for AI chips

18-24mo
Lead time to build a new CoWoS production line from scratch

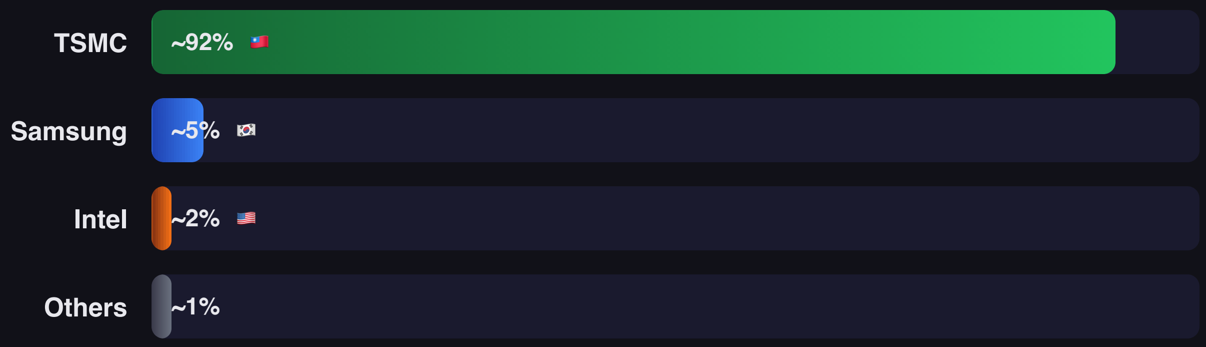
📊 CoWoS Capacity vs Demand (Illustrative)

Capacity has grown aggressively, but demand keeps outpacing it — especially as each new GPU generation uses a bigger interposer.



🔴 **Demand consistently outpaces capacity — the gap is GROWING because each B200 uses ~2× the CoWoS area of an H100**

🏢 Competitive Landscape — Who Else Can Package?



TSMC — CoWoS

2.5D Silicon Interposer

The gold standard. Highest density, best yield, most mature. Used by Nvidia, AMD, Google, and Amazon for their AI accelerators. Basically a monopoly on advanced AI chip packaging.

Samsung — FOPLP

Fan-Out Panel Level Packaging

Uses large rectangular panels instead of round wafers — potentially much cheaper at scale. But still early stage, lower density, and limited customer adoption. **Not yet proven for AI chips.**

Intel — EMIB

Embedded Multi-die Interconnect Bridge

Uses small silicon bridges embedded in the substrate (instead of a full interposer). Used in Intel's own Ponte Vecchio and Gaudi chips. **Not available as a foundry service at TSMC-scale.**

Bottom line: if you want to build a state-of-the-art AI GPU today, you're going through TSMC's CoWoS. There is no alternative at scale.

📌 Worth Knowing

Why can't they just build more CoWoS lines faster? — CoWoS packaging requires extremely specialized equipment (e.g., thermocompression bonders, advanced lithography for interposers). The supply chain for this equipment is itself constrained, and each new line takes 18-24 months to build, qualify, and ramp. You can't throw money at it and get instant capacity.

The interposer size problem — as chips get bigger (B200 → B300 → Rubin), the silicon interposer must grow too. But larger interposers are exponentially harder to manufacture with good yield. A single defect across that huge silicon area can kill the whole package. This is why CoWoS-L uses "stitched" chiplets — it's an engineering workaround to the reticle limit.

How this affects GPU pricing — the CoWoS bottleneck is a key reason Nvidia can charge $30,000-$40,000 per GPU. Even if die costs fell, the limited CoWoS capacity creates artificial scarcity. Cloud providers are willing to pay any price because the revenue from AI services far exceeds the GPU cost.

Connection to the HBM layer — CoWoS is literally the glue between the GPU die and HBM stacks. No CoWoS = you can't physically attach HBM to the GPU. Even if SK Hynix doubles HBM production, those stacks are useless without CoWoS to integrate them. This makes CoWoS the critical dependency for both the foundry and memory layers.

TSMC's strategic leverage — TSMC doesn't just make the GPU dies (foundry) — they also do the packaging (CoWoS). This gives them incredible leverage over the entire AI chip supply chain. Nvidia, AMD, Google, and Amazon are all dependent on a single company in Taiwan for both fabrication AND packaging.

The cost breakdown of an AI GPU — roughly: GPU die (~$2,000-3,000) + HBM stacks (~$2,000-4,000) + CoWoS packaging (~$1,000-2,000) + testing & other (~$500-1,000) = $6,000-10,000 manufacturing cost on a chip that sells for $30,000-40,000. Packaging is a significant chunk of the total.