HBM & the Memory Layer

Layer 4 of the AI Value Chain — where silicon meets bandwidth

Where Memory Sits in the Stack

LAYER 1 Raw Materials → LAYER 2 Equipment → LAYER 3 Foundries → **LAYER 4 Memory & HBM** →

What HBM Actually Is

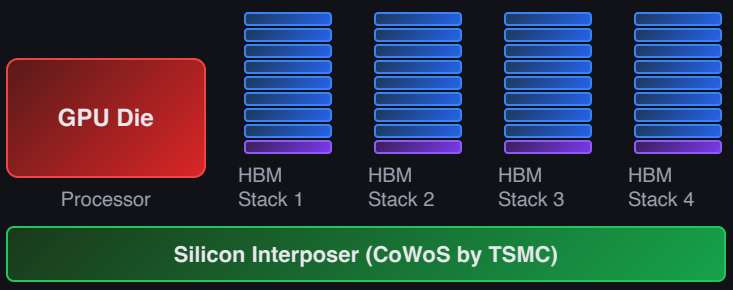
High Bandwidth Memory = DRAM dies stacked vertically like a tiny skyscraper, connected by thousands of tiny wires (TSVs) punched through each layer

- DRAM Die 12
- DRAM Die 11
- DRAM Die 10
- DRAM Die 9
- DRAM Die 8
- DRAM Die 7
- DRAM Die 6
- DRAM Die 5
- DRAM Die 4
- DRAM Die 3
- DRAM Die 2
- DRAM Die 1
- Logic Die (Controller)

↑ HBM3E: 12 DRAM dies stacked + 1 logic die

How HBM Connects to the GPU

GPU and HBM stacks sit side-by-side on a silicon interposer (CoWoS packaging by TSMC)



Nvidia H100 has 6 HBM stacks × 80GB = 80GB total · B200 has 8 stacks = 192GB

Connected by ~10,000+ TSVs (Through-Silicon Vias)

💡 Why HBM Exists — The Bandwidth Problem

AI models need to move **massive** amounts of data to/from the GPU every second. Regular DDR5 RAM sits far away on the motherboard and connects through narrow pipes. HBM sits *right next to* the GPU with thousands of parallel connections.

DDR5 (Regular RAM)	HBM3E
~50 GB/s	~4,800 GB/s
64-bit bus · Far from GPU	1024-bit bus × 8 stacks · Adjacent to GPU

HBM is ~100x faster than regular RAM. That's why AI chips need it.

⚖️ The HBM vs Consumer RAM Tradeoff

Same DRAM wafers, two very different products. Manufacturers must choose how to allocate limited fab capacity:

DDR5 (Consumer)	VS	HBM3E (AI)
PCs, phones, servers		GPU accelerators
~$3/GB		~$15/GB
		← 3-5x more profitable

↓ So what happens?

AI demand for HBM explodes 📈 → **Fabs shift wafers to HBM (higher margin)** → **Less DDR5 supply for consumers** →

PC/phone RAM prices rise 📈

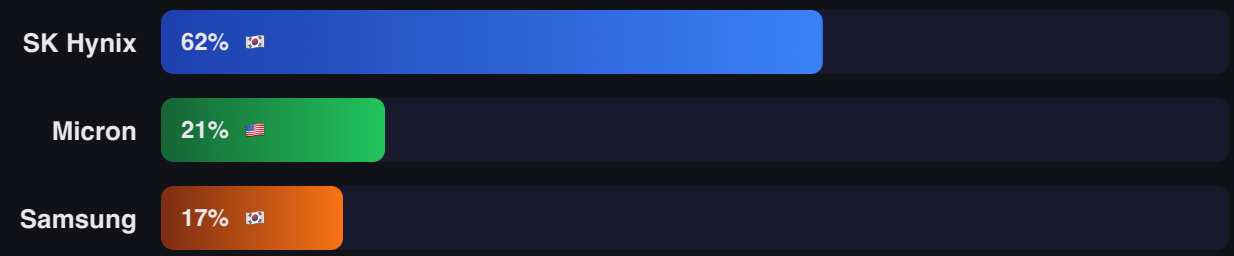
This is the "memory shortage" you're seeing in the news. It's not that we can't make RAM — it's that the economics of AI are pulling production away from consumers.

HBM Generations — The Evolution

GENERATION	BANDWIDTH	CAPACITY/STACK	LAYERS	USED IN
HBM2e	460 GB/s	16 GB	8 dies	A100
HBM3	819 GB/s	24 GB	8 dies	H100
HBM3E NOW	**1,200 GB/s**	**36 GB**	**12 dies**	**B200, MI300X**
HBM4 2026	1,800+ GB/s	48 GB	16 dies	Nvidia Rubin, AMD MI450

Each generation = more layers stacked, wider bus, more bandwidth. The trend: models keep getting bigger → need more memory → need faster memory.

Who Makes HBM — Market Share



Key insight: 79% of HBM production is in South Korea. Micron is the *only* US-based supplier — which is why it's getting $13.7B in CHIPS Act funding. All three producers are **sold out through end of 2026**.

Key Numbers to Remember

34×
More HBM per AI server vs previous generation

21.7 TB
HBM in Nvidia's GB300 rack (up from 640GB in DGX H100)

$200B
Micron's planned US investment in memory fabs

239%
Micron stock return in 2025

100×
HBM bandwidth vs regular DDR5 RAM

3-5×
HBM profit margin vs commodity DRAM

📌 Worth Knowing

TSVs (Through-Silicon Vias) — the tiny vertical wires punched through each DRAM die to connect the stacked layers. Manufacturing these without defects is extremely hard, which is why HBM yield rates are a competitive moat.

CoWoS bottleneck — TSMC's Chip-on-Wafer-on-Substrate packaging is needed to put HBM + GPU together on one interposer. CoWoS capacity was a major bottleneck in 2024-2025 and TSMC has been aggressively expanding it.

Why SK Hynix leads — they partnered with Nvidia early (2018) on HBM optimization. First to market with HBM3E. Samsung struggled with yield issues and only recently got Nvidia qualification.

Memory is no longer a commodity — historically DRAM was a cyclical boom-bust commodity. HBM has transformed it into a strategic, high-margin product. Micron's CEO: "Memory is now essential to AI's cognitive functions."

The heat problem — stacking 12 DRAM dies creates serious thermal challenges. HBM4 will need new thermal solutions, which is partly why it's being designed with a fundamentally different architecture.

Why not just use more regular RAM? — It's about bandwidth, not just capacity. AI models need to read billions of parameters every inference. The data bus on regular RAM is simply too narrow — like trying to fill a swimming pool through a garden hose vs a fire hydrant.