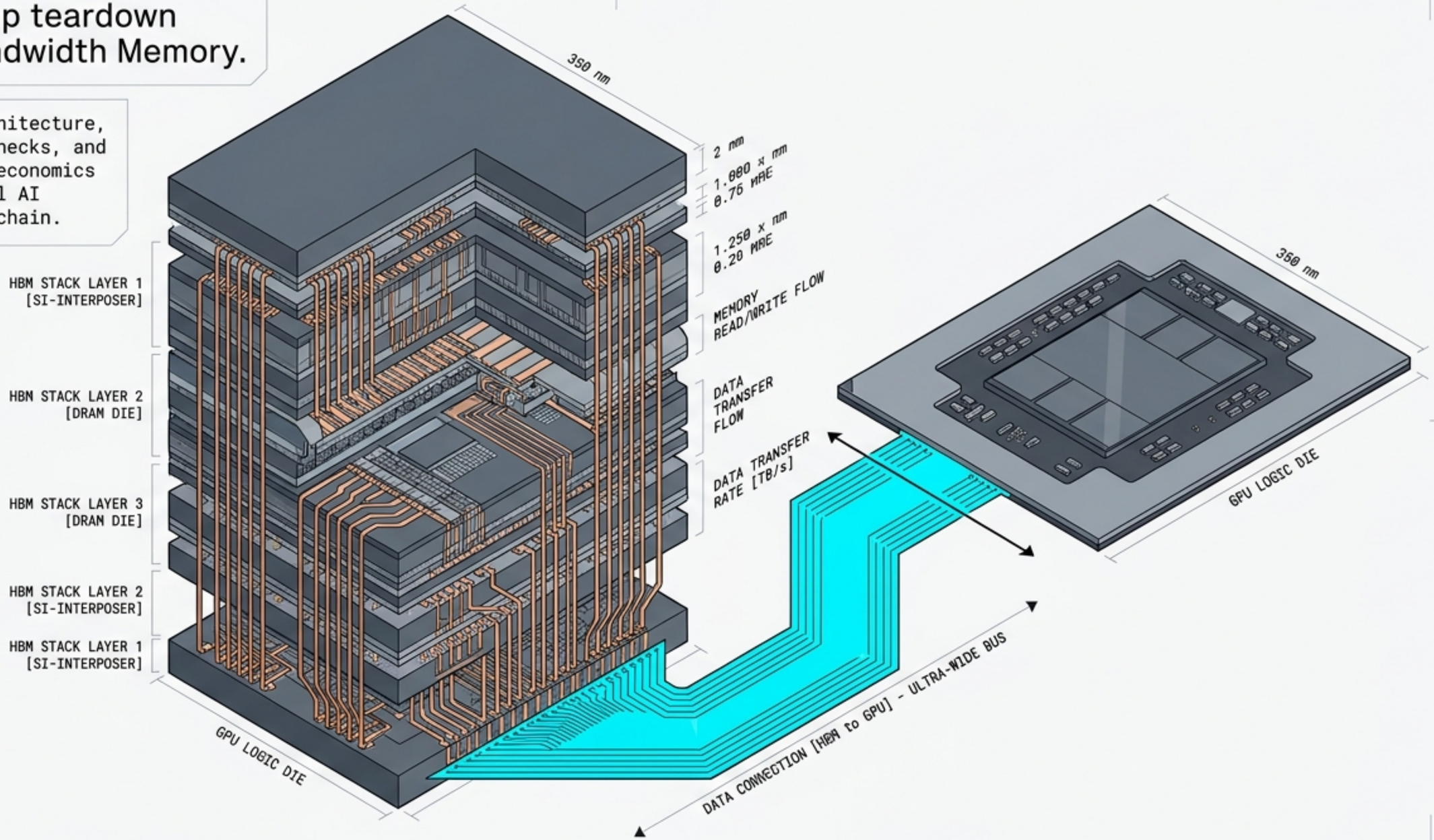
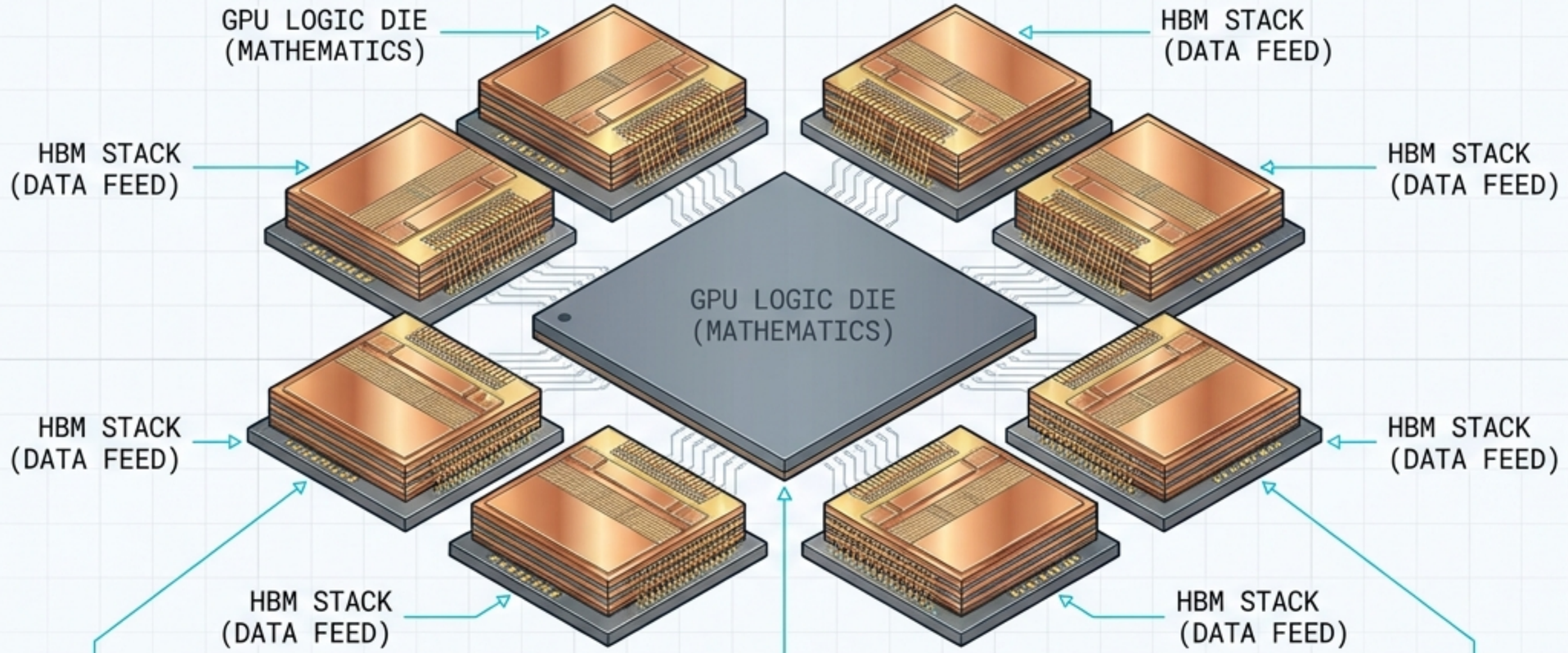
A physics-up teardown of High-Bandwidth Memory.

Decoding the architecture, packaging bottlenecks, and trillion-dollar economics gating the global AI hardware supply chain.





THE HARDWARE ILLUSION

Investors obsess over the GPU logic die, but it is not the most expensive component on a modern AI module.

```
ANALYSIS POINT:
VISUAL FOCUS != COST REALITY
```

THE BOM REALITY

HBM stacks account for approximately 40-50% of the total Bill of Materials on a flagship AI module.

```
DATA MATRIX:
[COMPONENT] | [BOM SHARE%] | [STATUS]
GPU LOGIC   | <20%         | COMMODITY
HBM MEMORY  | 40-50%       | CRITICAL
```

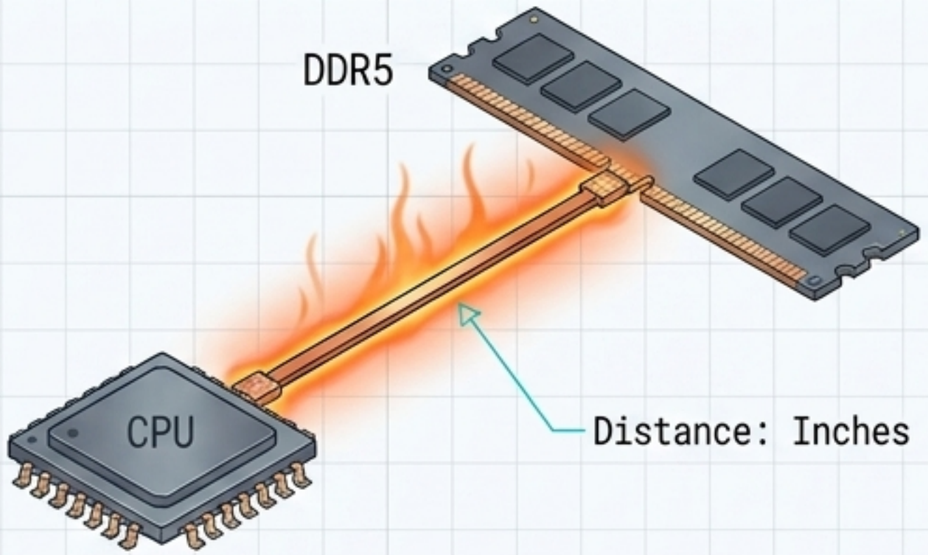
THE SUPPLY CONSTRAINT

HBM is the single most sold-out, heavily supply-constrained component in the datacenter ecosystem.

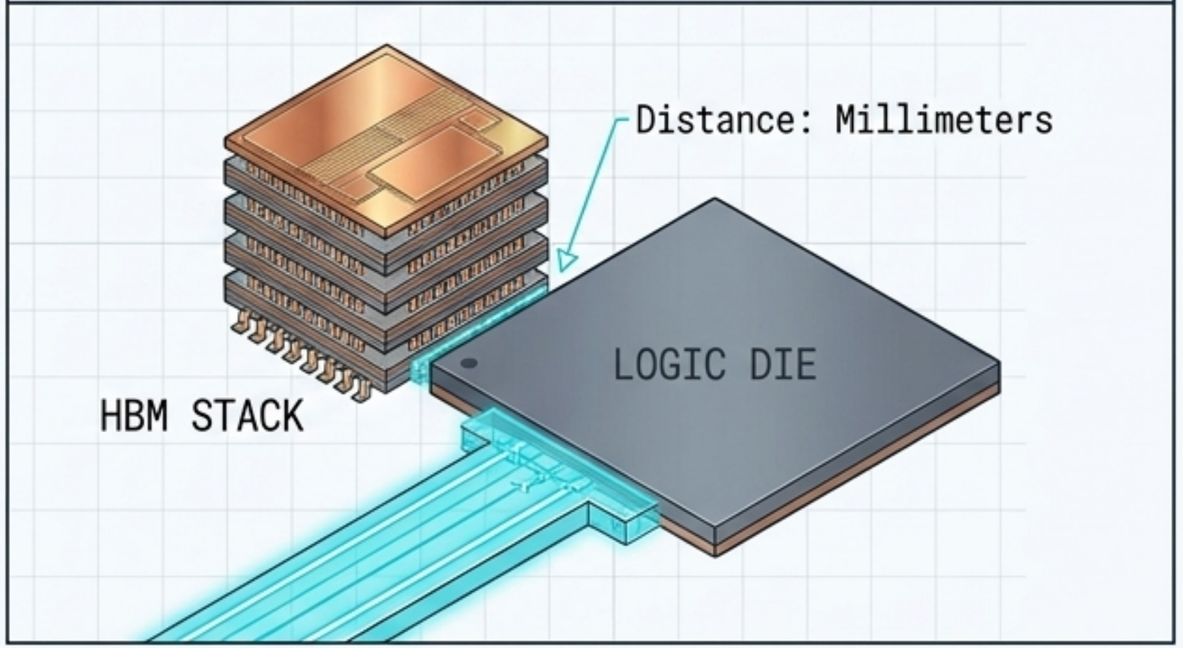
```
INDUSTRY NOTE:
MEMORY BANDWIDTH IS THE BOTTLENECK
```

The geometrical shift from “narrow and fast” to “wide and close”

Traditional DDR5 (Narrow & Fast)

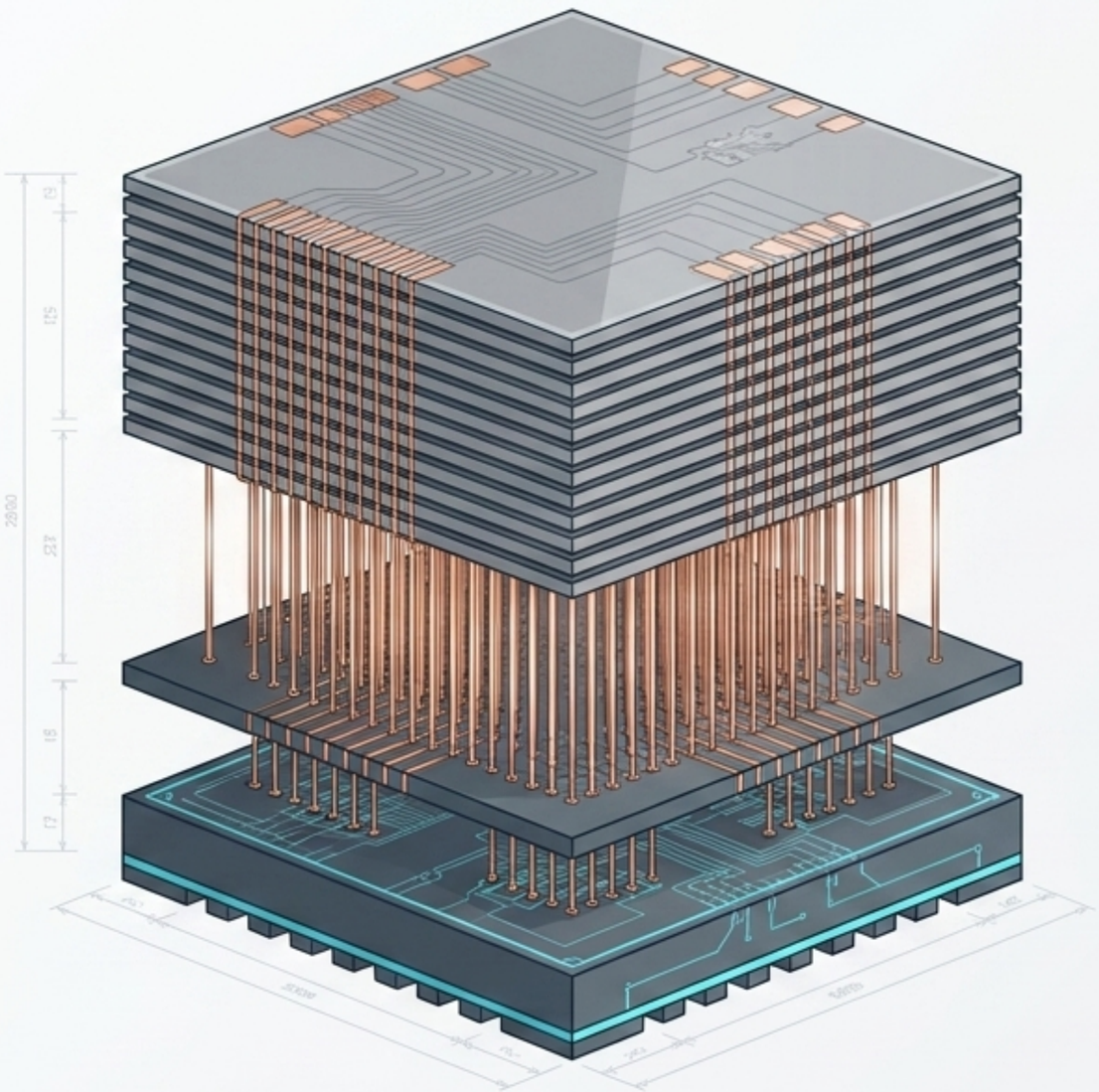


HBM (Wide & Close)



Metric	DDR5	HBM
Architecture	Flat spread	90-degree vertical stack
Interface Width	64-bit channel	1024-bit firehose
Distance	Inches	Millimeters
Power Efficiency	High energy burn across distance	Low picojoules-per-bit yielding massive bandwidth-per-watt

An anatomy of a 12-story memory tower



DRAM Dies

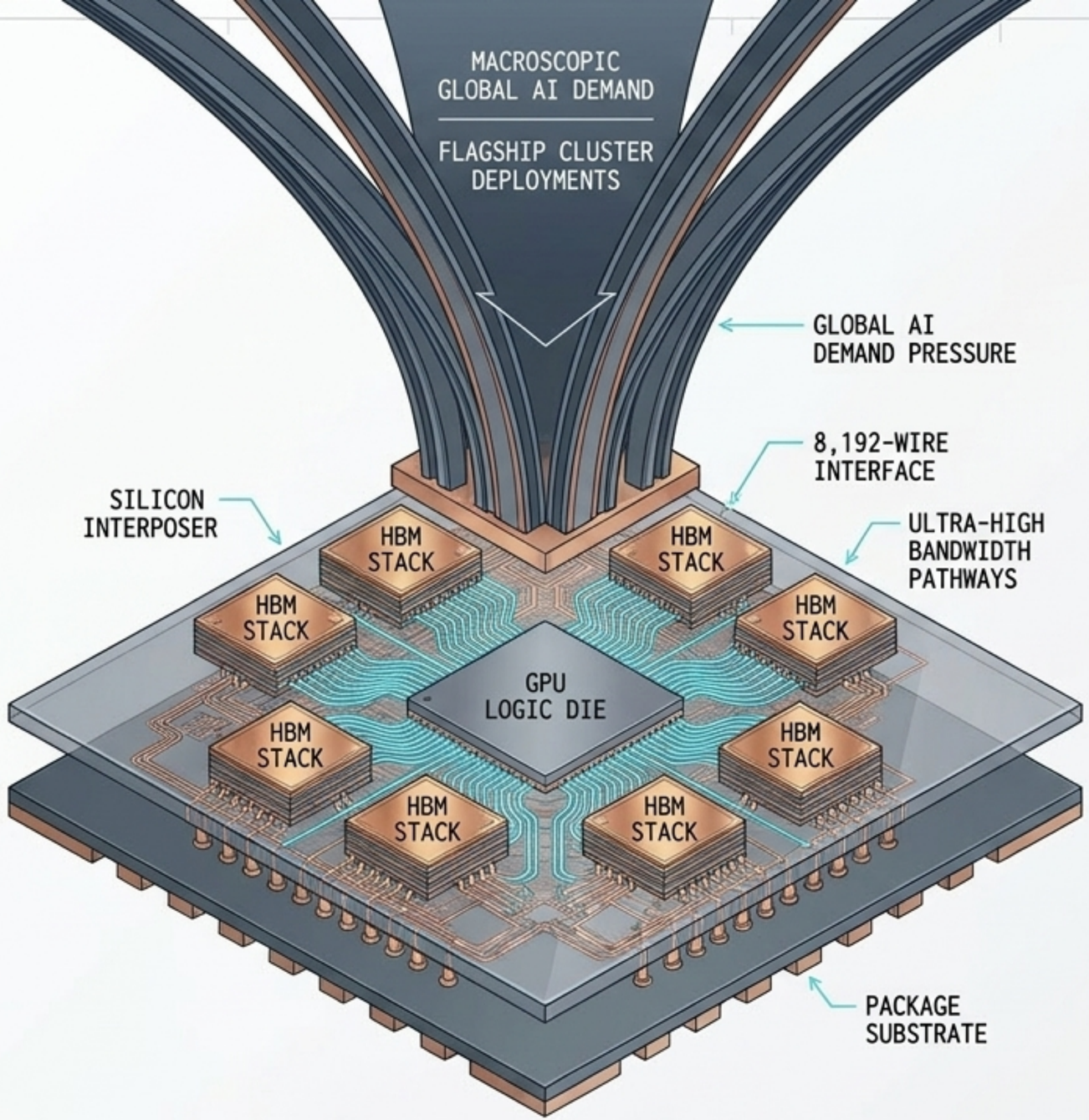
The top layers (8-Hi, 12-Hi, or 16-Hi configurations). The actual memory cells storing the bits.

Through–Silicon Vias (TSVs)

The physical spine. Thousands of microscopic copper-filled holes etched straight down through the silicon, carrying signal and power.

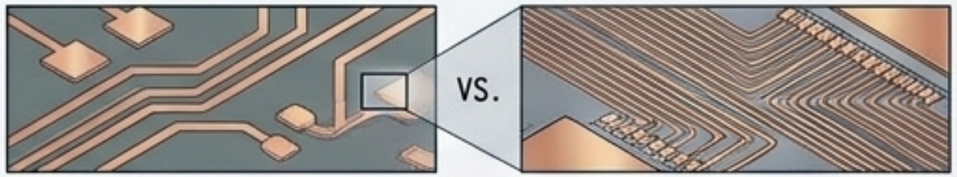
Base Logic Die

The foundation. Acts as the memory controller, handling test, repair, I/O, and presenting a clean interface to the GPU.



THE ASSEMBLY REALITY

You cannot simply bolt a GPU and HBM together on a standard circuit board. 8,192 wires (8 stacks × 1024 bits) require silicon-grade routing.



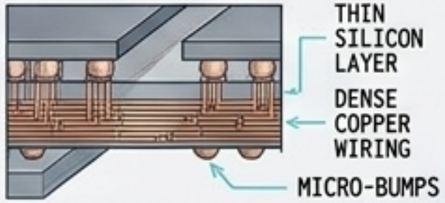
STANDARD PCB

SILICON-GRADE ROUTING (SUB-MICRON)

THE INTERPOSER

A thin slab of silicon etched with ultra-fine wiring that houses both the GPU and the HBM stacks.

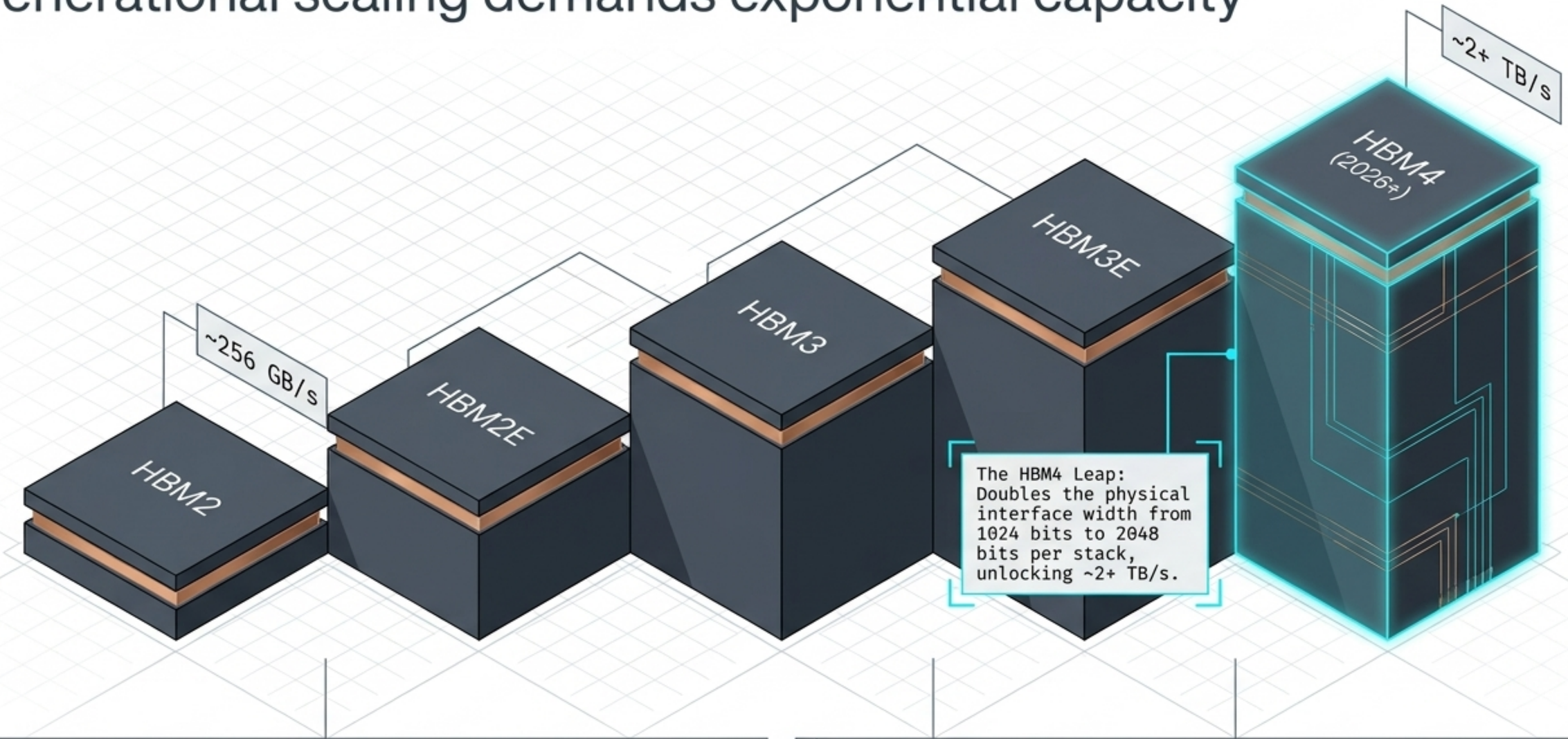
INTERPOSER CROSS-SECTION



THE CoWoS MOAT

TSMC's CoWoS process is required to marry the logic and memory. Despite doubling capacity annually, CoWoS remains the ultimate chokepoint gating all flagship AI cluster deployments.

Generational scaling demands exponential capacity



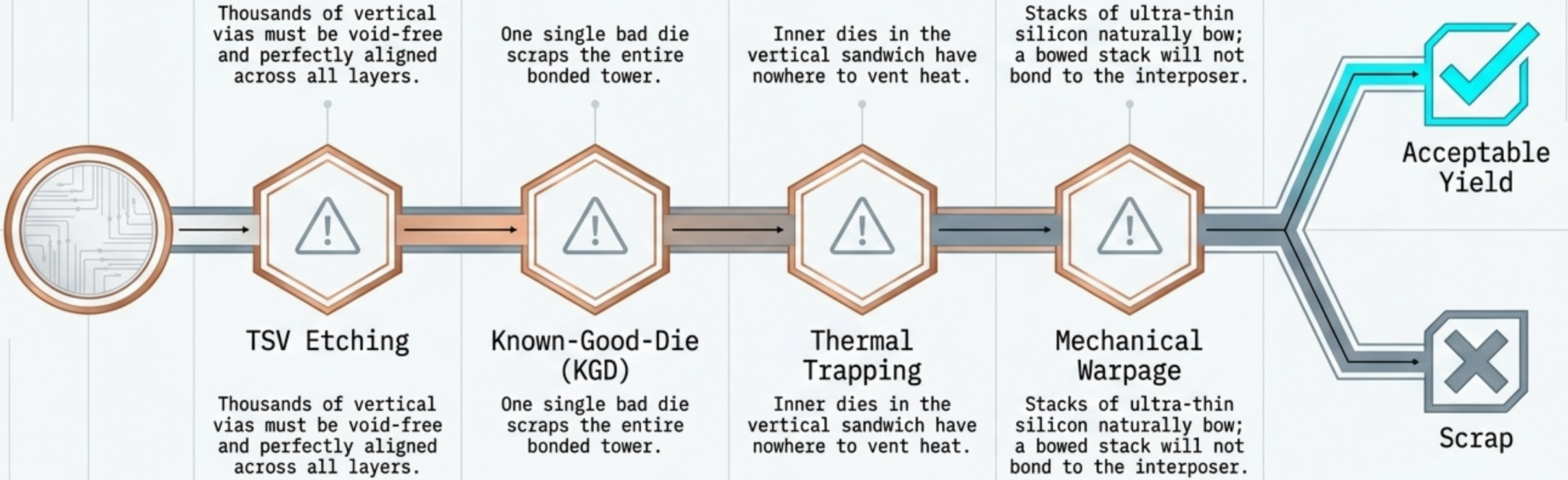
Bandwidth Velocity

HBM2 -> HBM3E: Bandwidth jumped ~5x from ~256 GB/s to ~1.2 TB/s per stack in just a few years.

Content Per GPU

HBM required per flagship GPU doubles roughly every two years (H100: 80GB -> H200: 141GB -> B200: ~192GB -> B300: ~288GB).

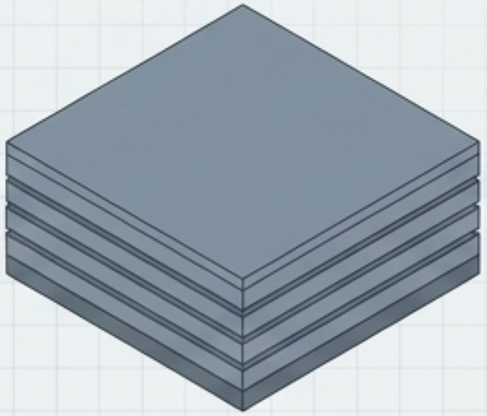
The spec is public, but compounding yield is the moat



Key Insight: Yield math compounds exponentially. Moving from 8-Hi to 12-Hi to 16-Hi is a fresh, structurally harder yield fight, trapping newcomers in years of catch-up.

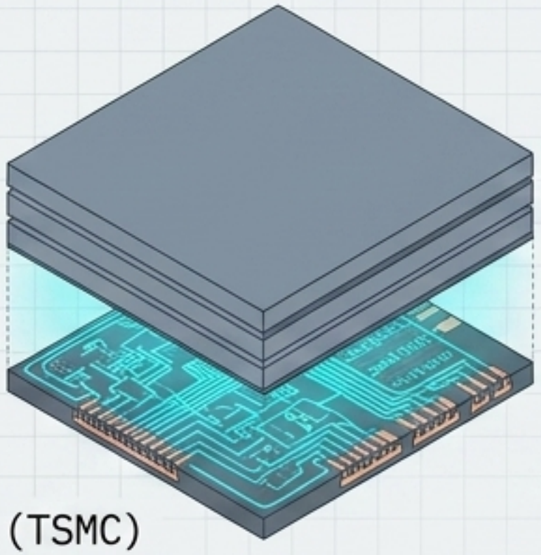
HBM4 transforms commodity memory into semi-custom logic

Before
(HBM3E)

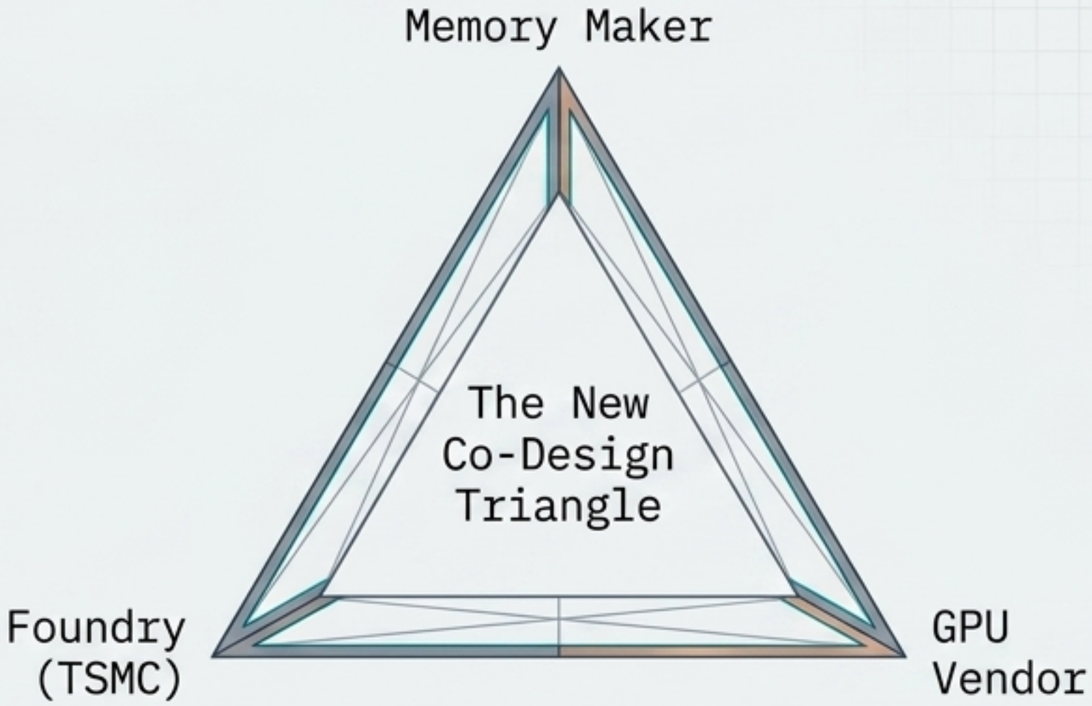


Fabricated by
Memory Maker

After
(HBM4)



Custom Logic
Fabricated by
Advanced Foundry (TSMC)



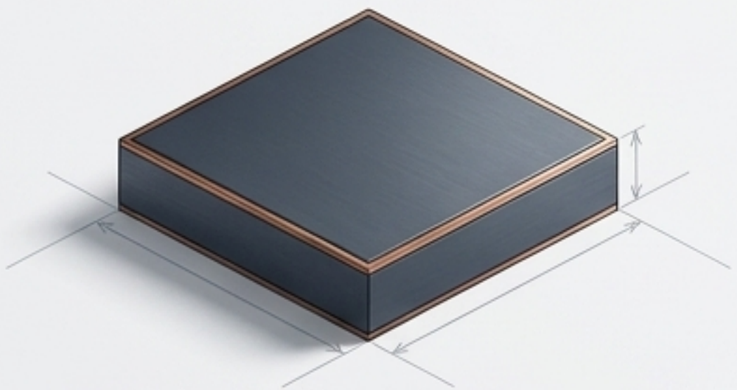
The Shift

Historically, the base die was a simple commodity fabricated by the DRAM maker.
HBM4 moves the base die to advanced foundry processes (TSMC).

The Impact

HBM ceases to be a spec-sheet part. It becomes a co-designed, customer-specific logic-plus-memory product, fundamentally reshuffling where margin is captured.

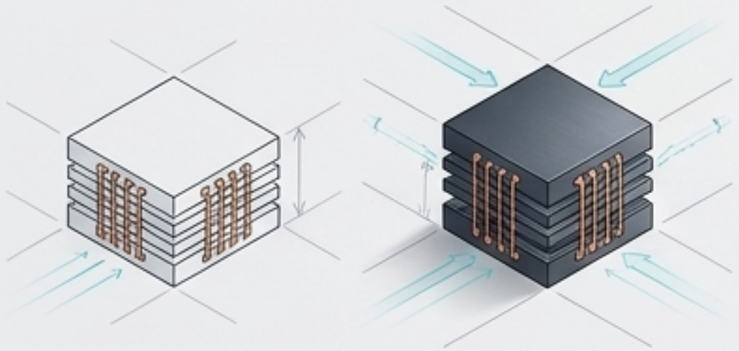
Milestone: All three crossed $1T market cap in May 2026.

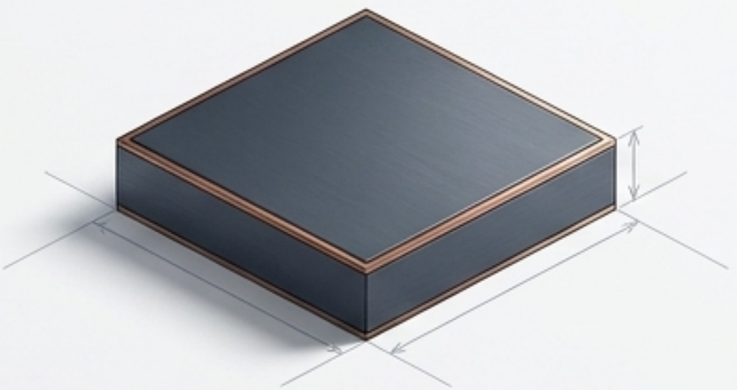


SK Hynix

The Leader

>50% market share. First to HBM3/3E. Nvidia's primary, critical partner.

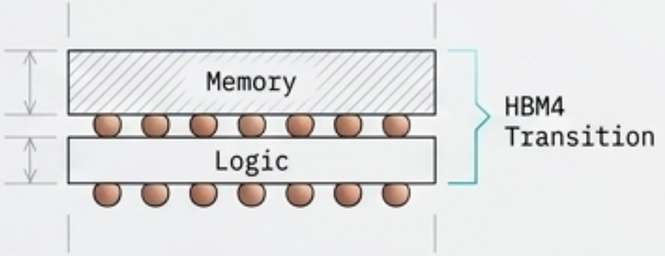


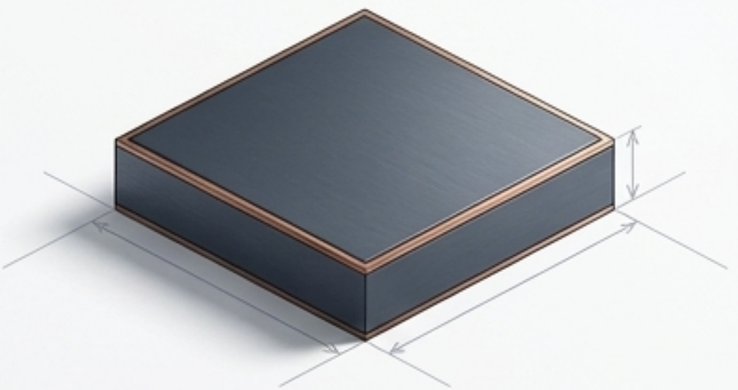


Samsung

The Integrated Giant

Owns both memory fabs and logic foundries. Leveraging in-house foundry capability for the HBM4 transition.



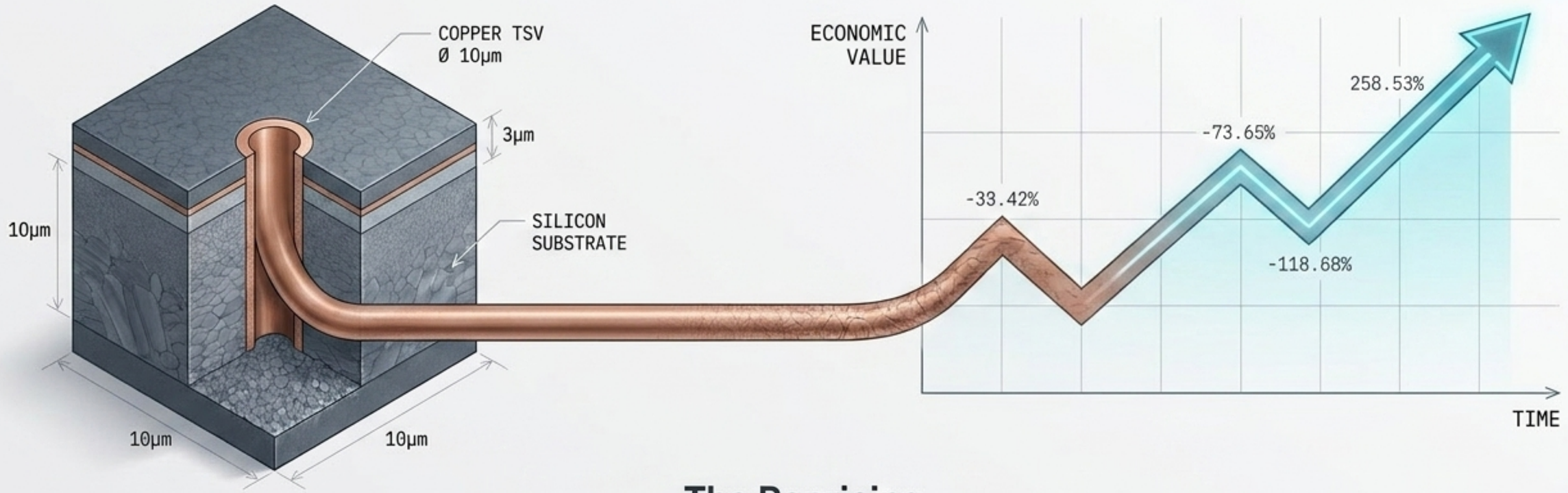


Micron

The US Champion

Ramping 12-Hi HBM3E, aggressively gaining share, backed by the US supply chain reshoring narrative.

Physics dictates economics



The Repricing

The physical reality of DRAM rotated 90 degrees and threaded with thousands of vias took the worst boom-bust business in tech and transformed it into an unclonable chokepoint.

The Insight

The hardest part of the AI boom was never the chip that does the math—it was the memory stacked beside it.