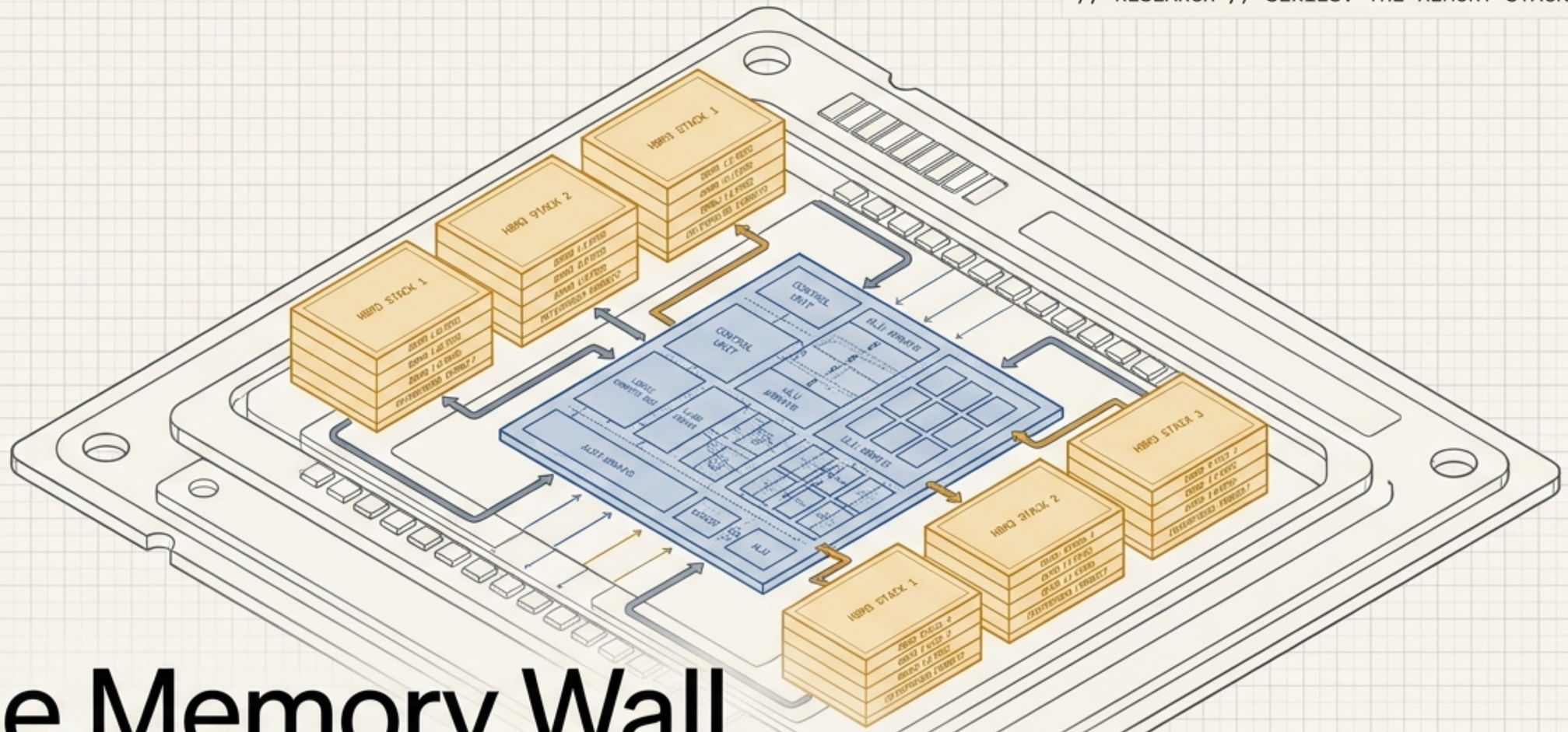
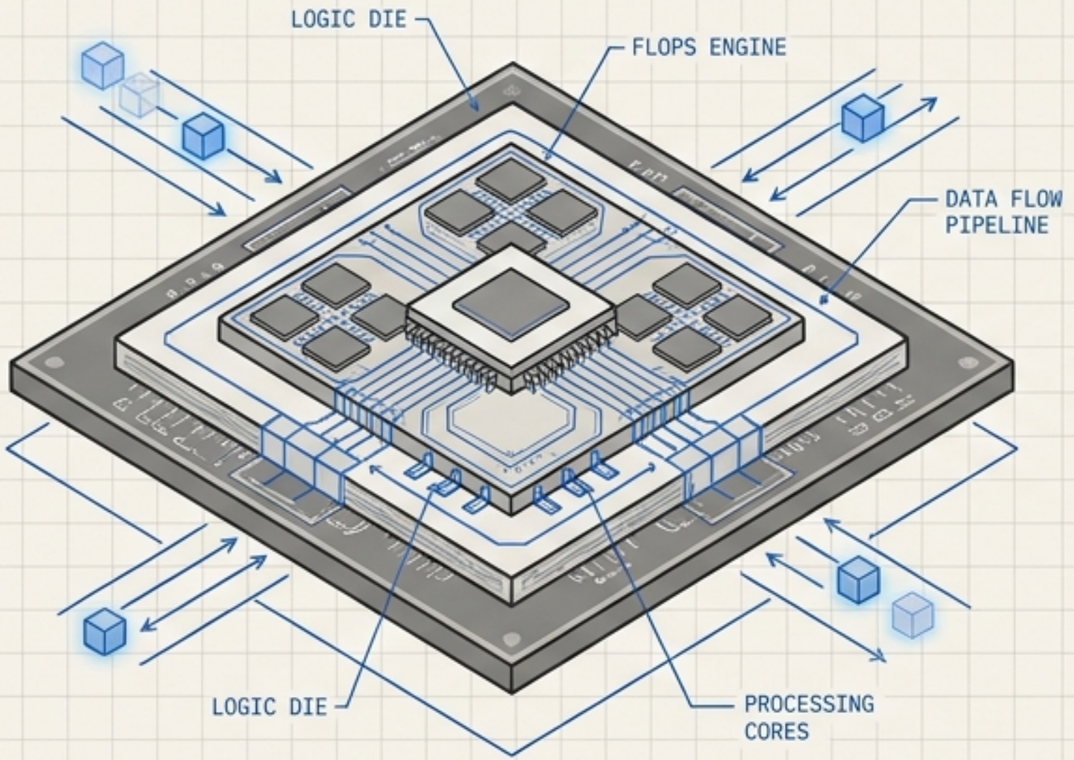


The Memory Wall.

Why the most critical constraint in AI hardware isn’t the logic die—it’s the memory stacked next to it.

The Hype: Compute

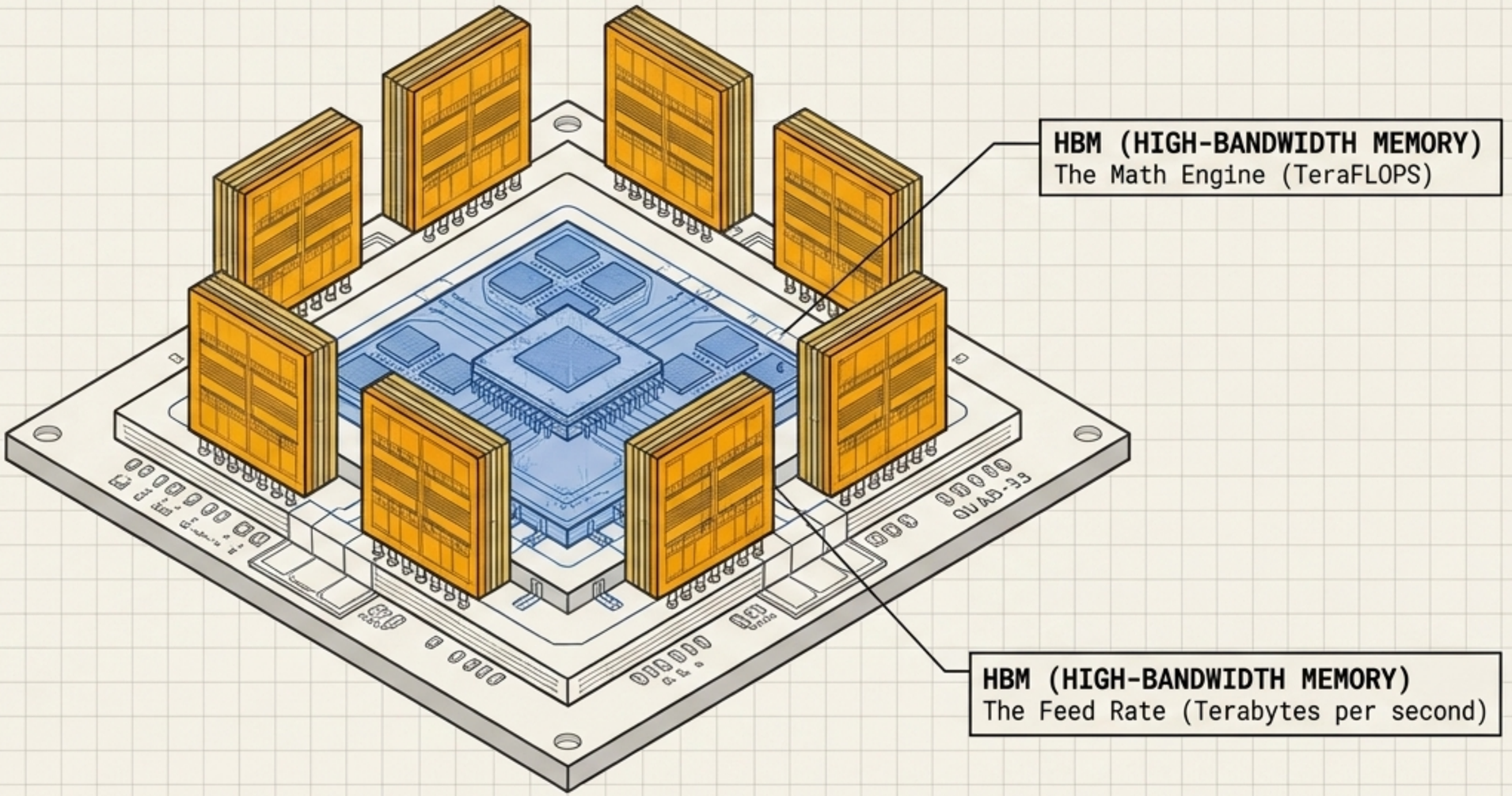


Most models assume the value and supply anxiety rest entirely on the logic die—the raw FLOPS.

The BOM: Memory

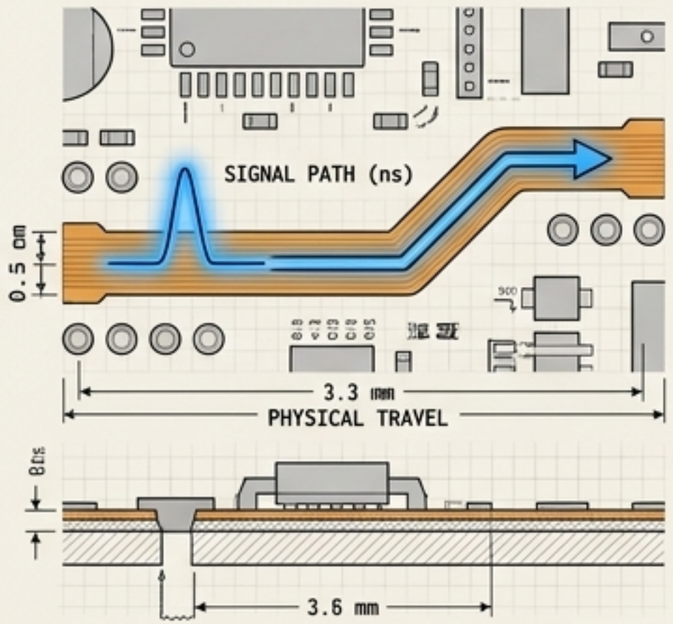
ITEM	QTY	PART NO.	DESCRIPTION	UNIT COST	TOTAL
1	1	SRV-CPU-X9	High-Perf Server CPU (Logic)	$2,500.00	$2,500.00
2	1	SRV-MB-ATX	Server Motherboard (ATX)	$450.00	$450.00
3	4	SRV-FAN-120	120mm Chassis Fan	$25.00	$100.00
4	1	SRV-PSU-1600	1600W Power Supply	$350.00	$350.00
5	8	SRV-HBM-64G	High-Bandwidth Memory (HBM) 64GB	$3,500.00	$28,000.00
6	1	SRV-CHS-4U	4U Rack Chassis	$800.00	$800.00
				TOTAL ESTIMATED BOM:	$32,200.00

When you finance hardware, you underwrite the residual value and supply chain risk. The bottleneck isn't fabbing logic; it's sourcing memory.



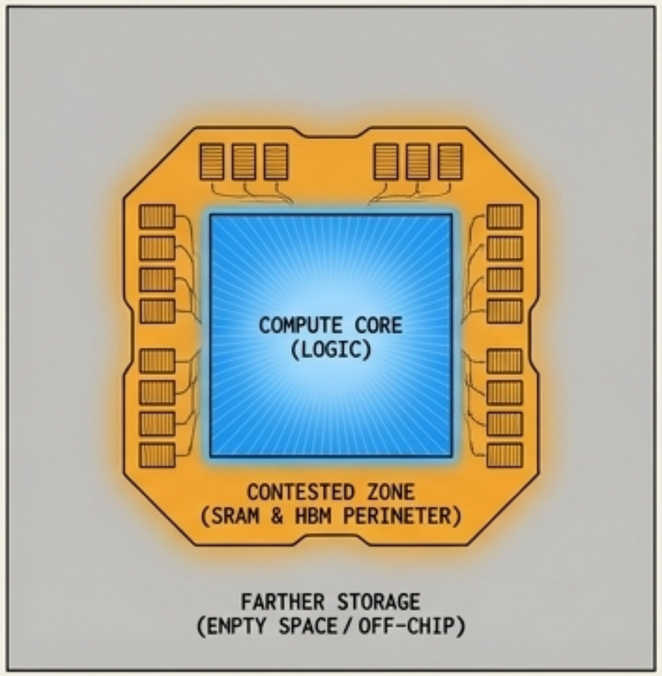
The logic die is what marketing slides count in teraFLOPS. The HBM ring is what actually gates how fast that logic can work. You can own all the FLOPS you want—if you can't feed them, they sit idle.

Distance is Latency



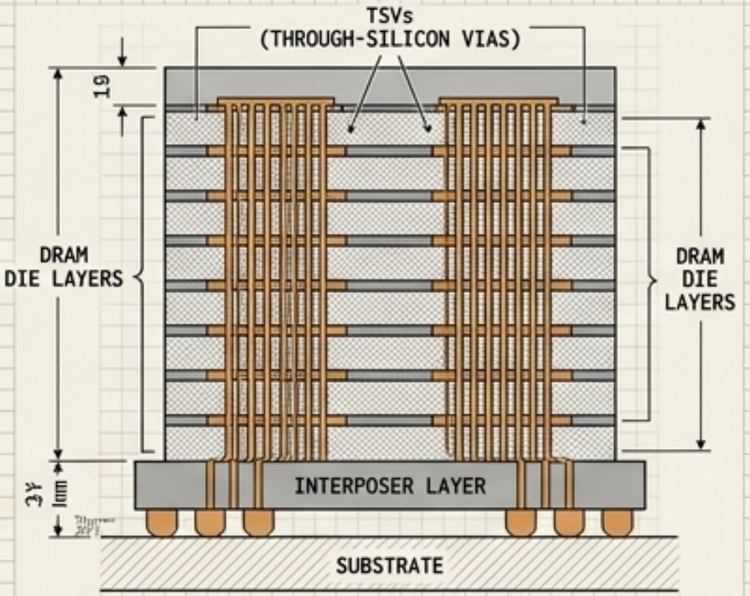
Electricity is fast, not infinite. Signals travel a few centimeters per nanosecond. To be fast, you must physically be close.

Close is Small

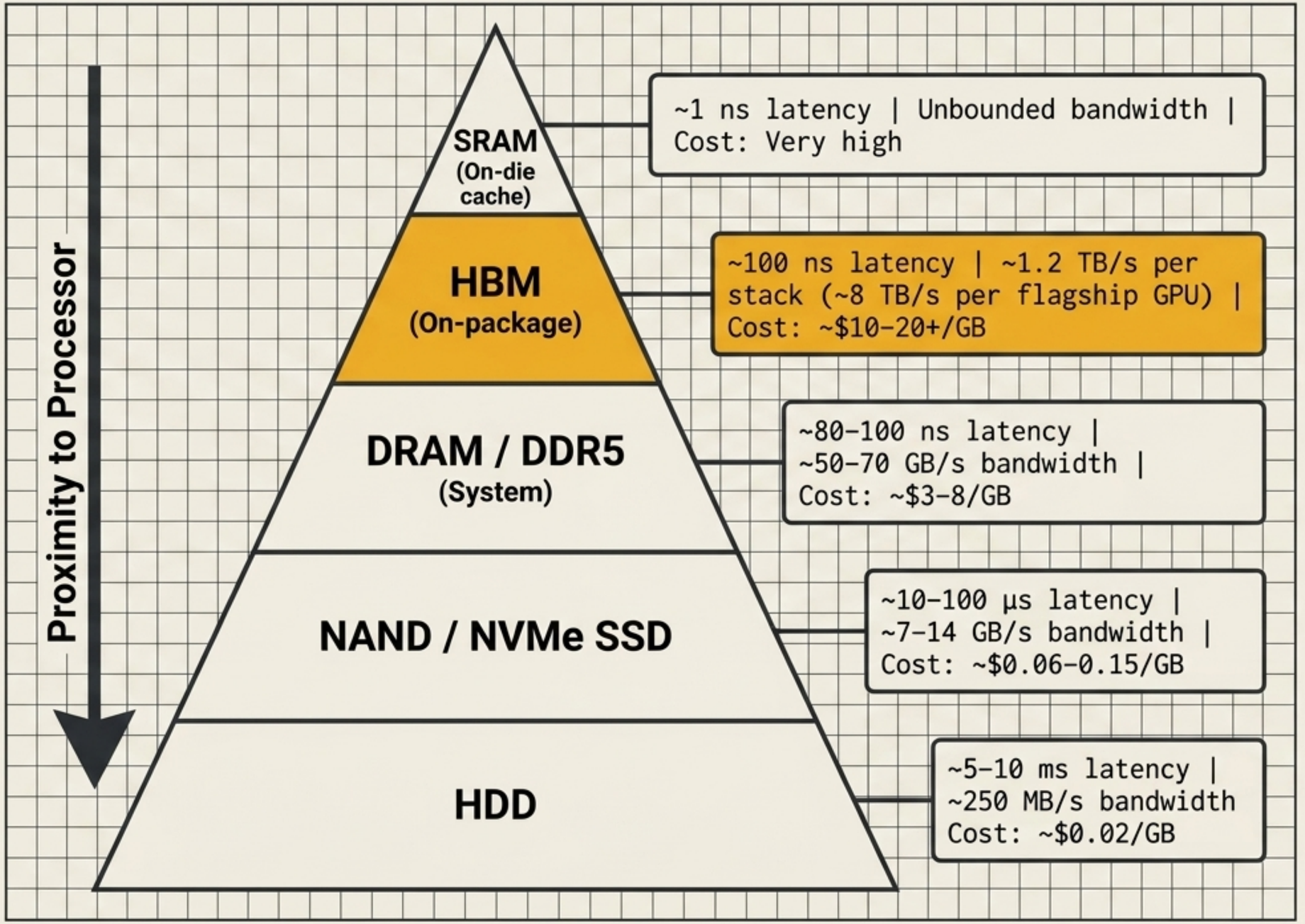


The perimeter around a die can only hold tens of megabytes of SRAM and a few HBM stacks (~a couple hundred GBs). Larger storage must live farther away.

Fast-and-Close is Expensive



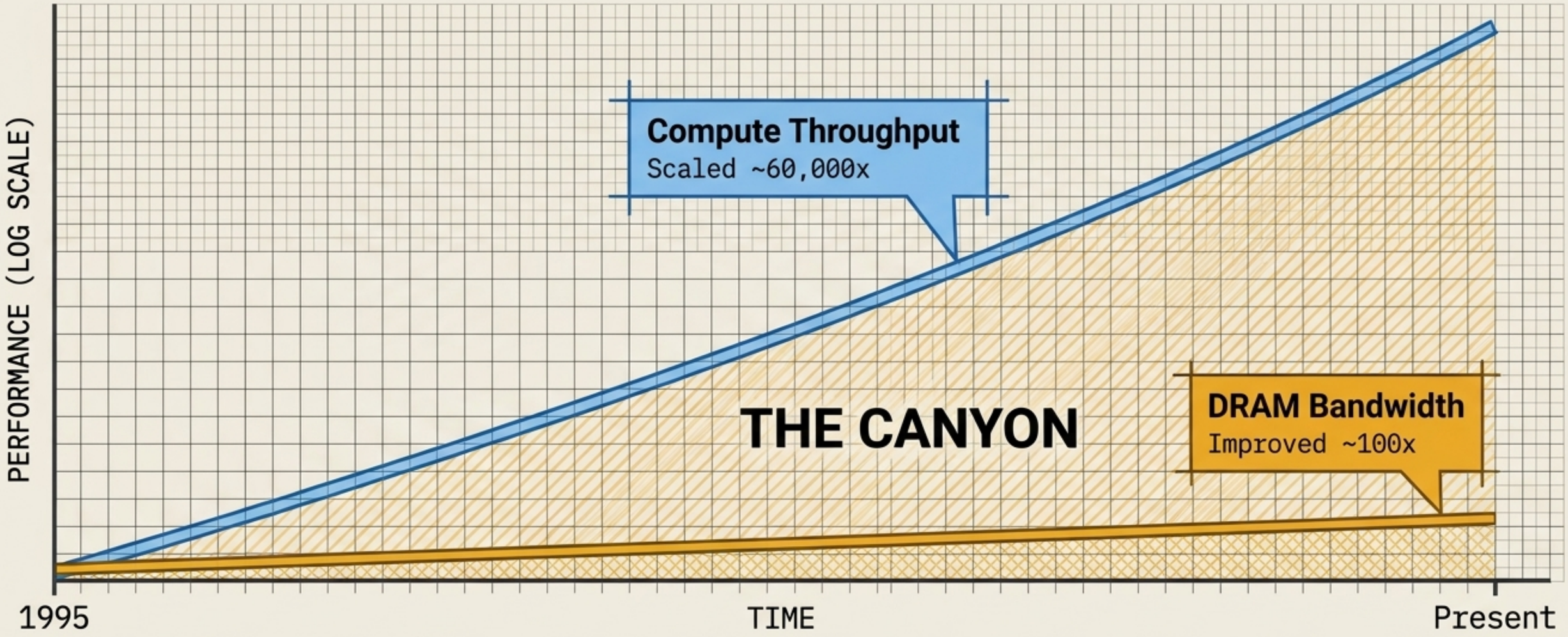
SRAM requires 6 transistors per bit. HBM demands brutal, low-yield 3D stacking processes. The speed has an exponential cost multiplier.



Read the extremes.

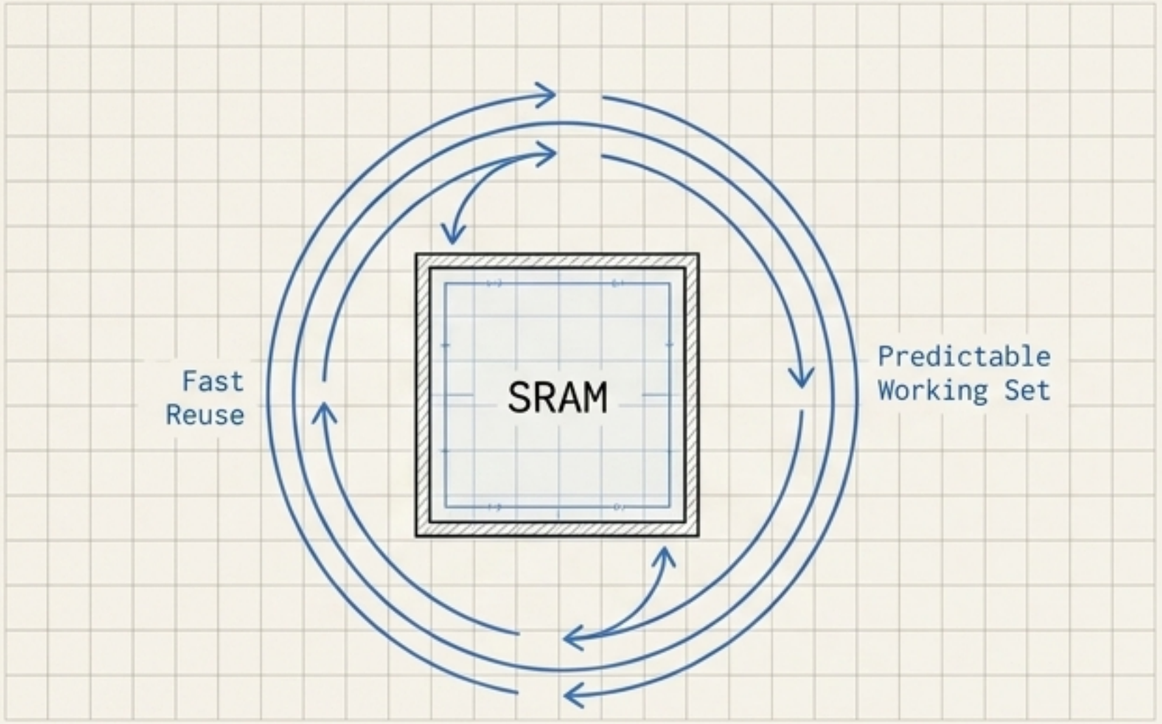
HBM moves data 1,000x faster than an SSD, but costs 100x more per gigabyte.

Every system design is a negotiation up and down this stack.



In 1995, architects warned of a “Memory Wall”—a point where faster processors would just spend all their time waiting for data. For decades, clever engineering (caches, prefetching) hid this canyon. Then came AI.

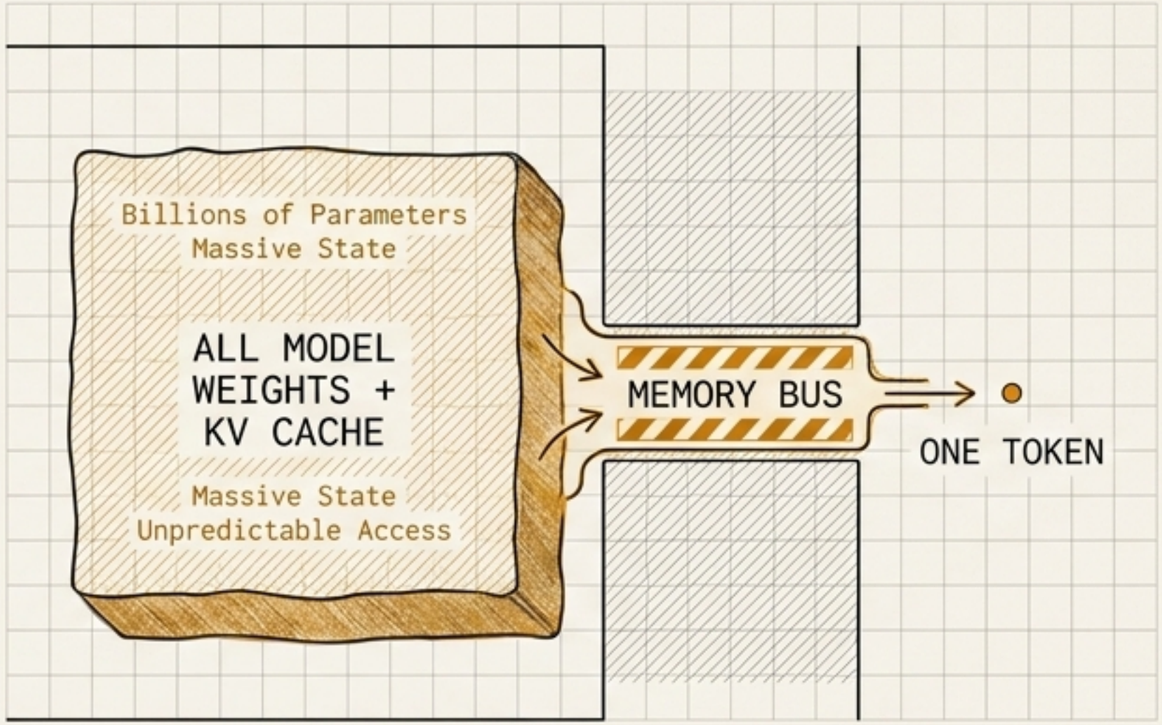
THE CACHE BET



Most software reuses a small working set of predictable data in tight loops.

You never feel the canyon because you never have to cross it.

THE AI REALITY

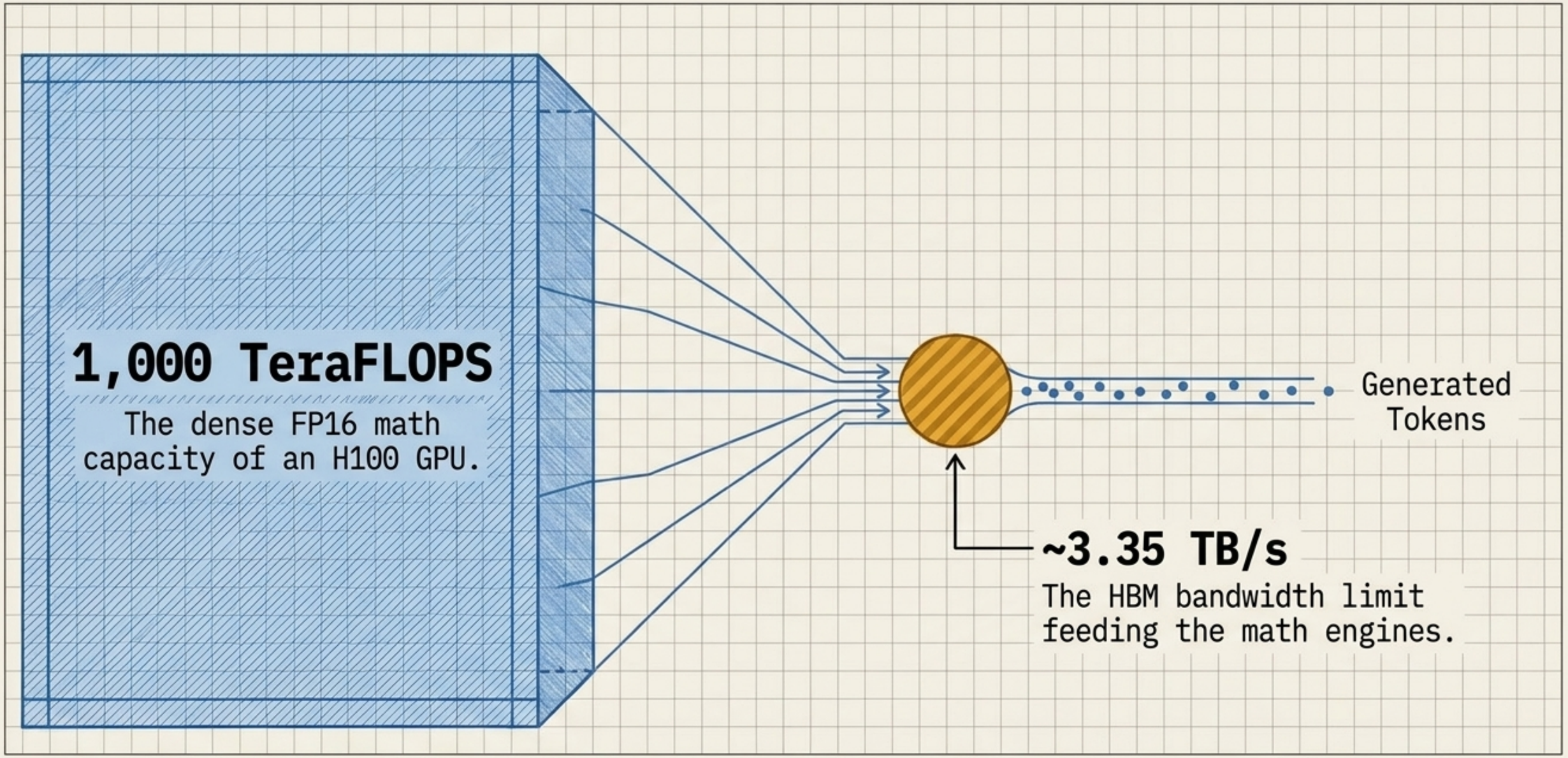


Token generation breaks the bet. The data you need next is all of it, every time.
No reuse.

No predicting your way around it.

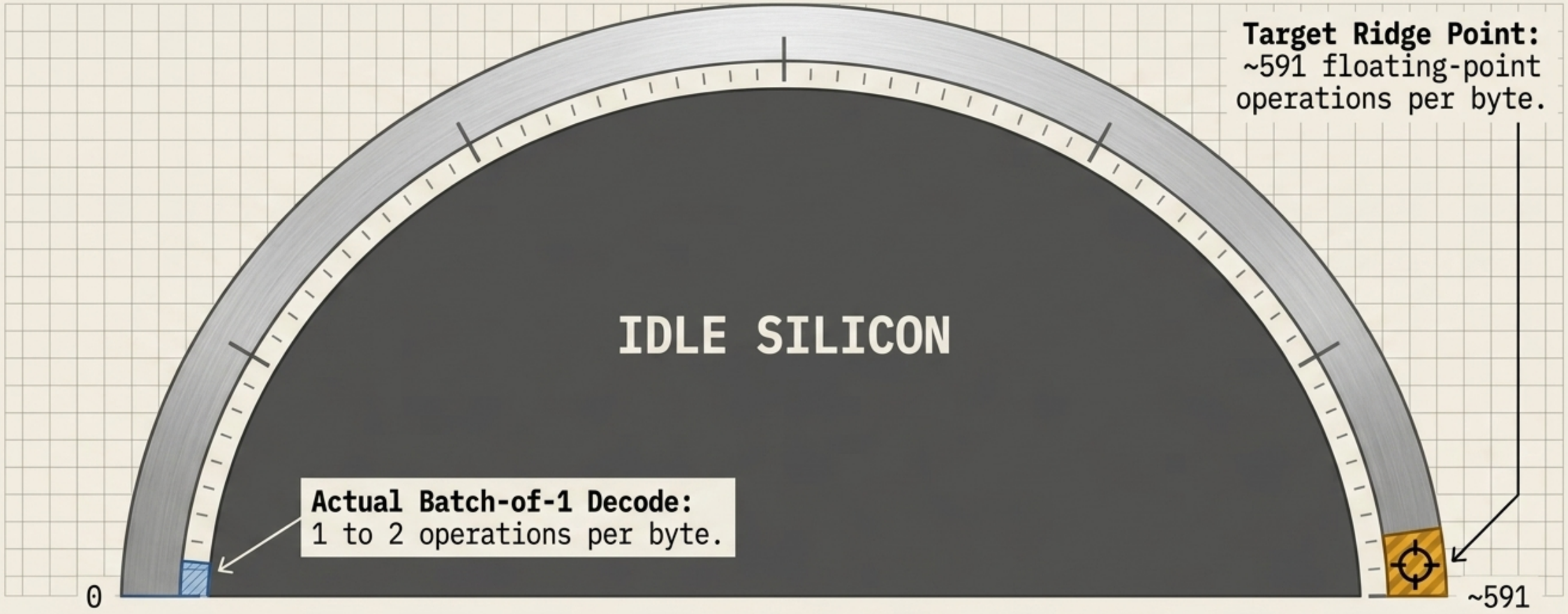
The Two Worlds of AI Compute

	Training & Prefill (The Benchmark)	**Decode / Token Generation** (The Reality)
Arithmetic Intensity	High (Lots of math per byte fetched).	Low (Tiny math per massive byte fetch).
Mechanism	Big dense matrix multiplies over large batches of data.	Dragging the entire weight set across the bus just to emit a single token.
Constraint	**Compute-Bound.** The FLOPS earn their keep. This is what marketing benchmarks measure.	**Memory-Bandwidth-Bound.** The bus chokes. This is where production AI spends overwhelmingly most of its life.



You bought a sports car and you're stuck behind the on-ramp's speed limit. Buying a faster engine doesn't help. Widening the on-ramp does.

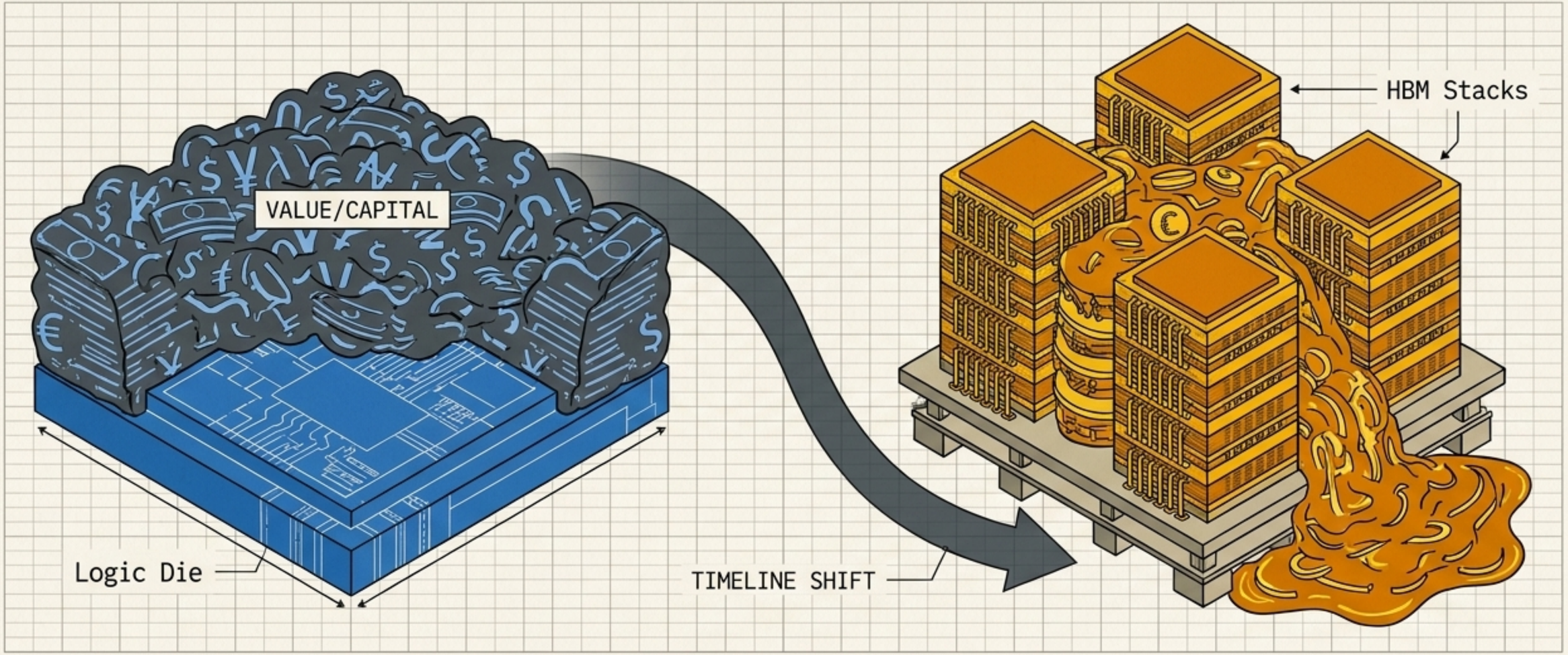
The Utilization Cliff: Why Your GPU Is Idle



During single-stream decode, a flagship GPU runs at just a few percent of its peak FLOPS. The expensive math units—the thing the chip is named and sold for—sit idle, starved, waiting on memory.

Past (Logic Era)

Present/Future (Memory Era)

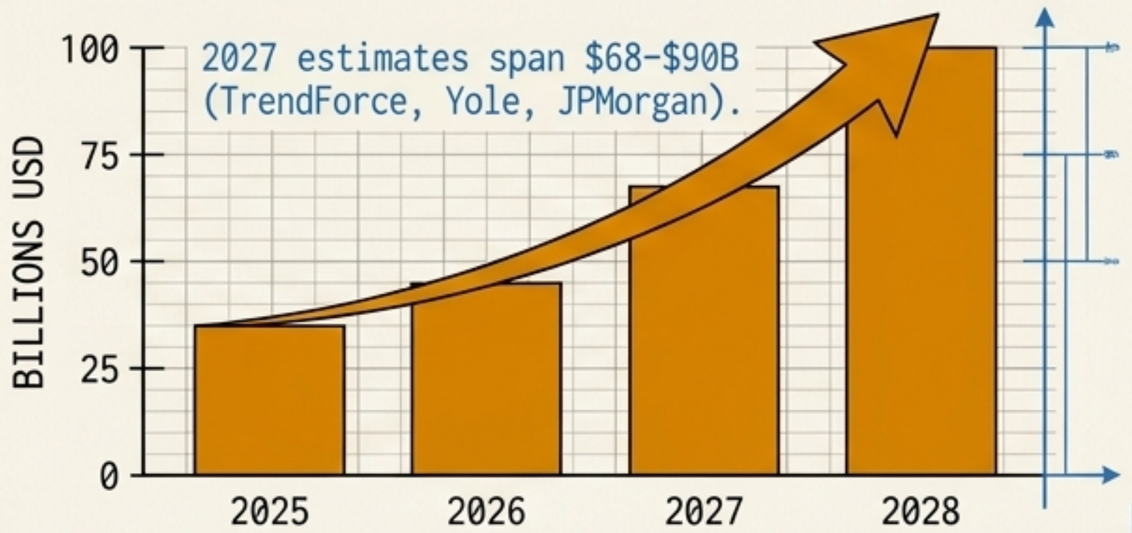


Follow the money, because the money already moved. When the constraint was logic, value pooled around the logic die. With the constraint shifted, value is violently reallocating to the memory supply chain.

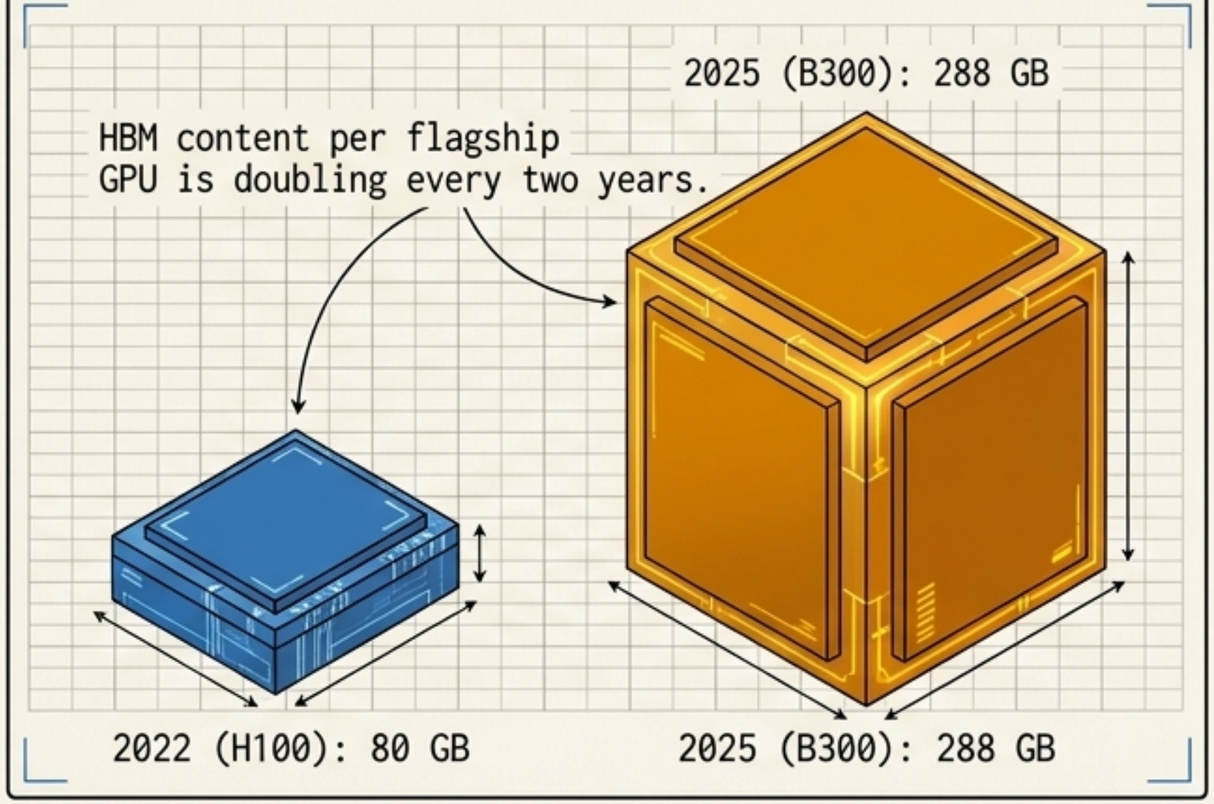
The HBM Financial Engine: Market & Physical Multipliers

MARKET EXPANSION

The HBM Market is projected to grow from **~$35 Billion** in 2025 to **~ $100 Billion** by 2028 (40% CAGR).



CONTENT PER CHIP MULTIPLIER



More memory per chip. More chips total. Higher prices per gigabyte.
Three multipliers stacked on top of each other, all driven by the Memory Wall.

~~The defining question of the last decade:~~
Who has the most **FLOPS**?

The defining question of the AI era: Vitocasvs Who can feed them?

Memory bandwidth and capacity are the real spec sheet now.