Raw Compute is an Illusion When the Engine is Starving

The Logic Die (GPU)

- ~1,000 TeraFLOPS of peak dense FP16 compute (H100).
- The highly marketed engine you pay for.
- *Incredibly fast, but currently starving.*

The Memory (HBM)

- ~3.35 TB/s to ~8 TB/s bandwidth.
- The physically constrained on-ramp feeding the engine.
- *The actual speed limit of artificial intelligence.*

The Core Constraint: When building an AI server, raw compute (FLOPS) is no longer the binding constraint. If you cannot feed the processor, the expensive math units sit idle. You cannot fix traffic by buying a faster car; you must widen the road.

A Thirty-Year Trajectory Finally Hits the Wall

Compute Scaling (The Runaway)

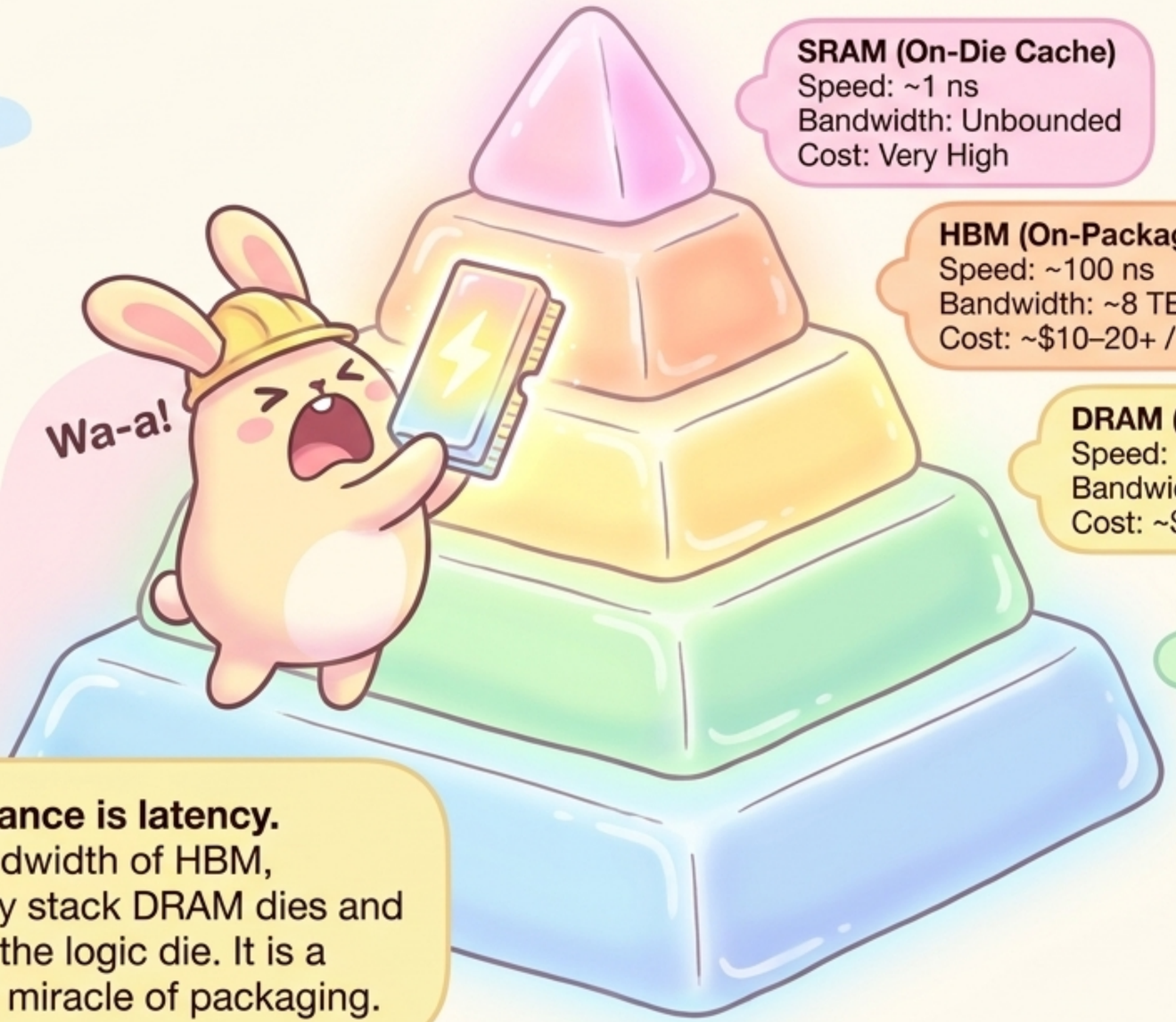
- Since 1995, raw processor throughput has scaled by roughly 60,000x.
- Engineers previously hid this gap using bigger caches, prefetching, and smarter prediction.

Memory Scaling (The Crawl)

- In the exact same timeframe, DRAM bandwidth improved only ~100x.
- Latency—the physical time required to fetch a single piece of data—has barely moved at all.

The gap between processor speed and memory speed didn't just widen—it became a canyon. In 1995, architects warned of the "Memory Wall." AI is the workload that finally crashed into it.

The Unforgiving Physics of the Memory Pyramid



SRAM (On-Die Cache)
Speed: ~1 ns
Bandwidth: Unbounded
Cost: Very High

HBM (On-Package)
Speed: ~100 ns
Bandwidth: ~8 TB/s
Cost: ~$10–20+ / GB

DRAM (System DDR5)
Speed: ~80-100 ns
Bandwidth: ~50-70 GB/s
Cost: ~$3–8 / GB

NAND (NVMe SSD)
Speed: ~10-100 µs
Bandwidth: ~7-14 GB/s
Cost: ~$0.06–0.15 / GB

HDD (Cold Storage)
Speed: ~5-10 ms
Bandwidth: ~250 MB/s
Cost: ~$0.02 / GB

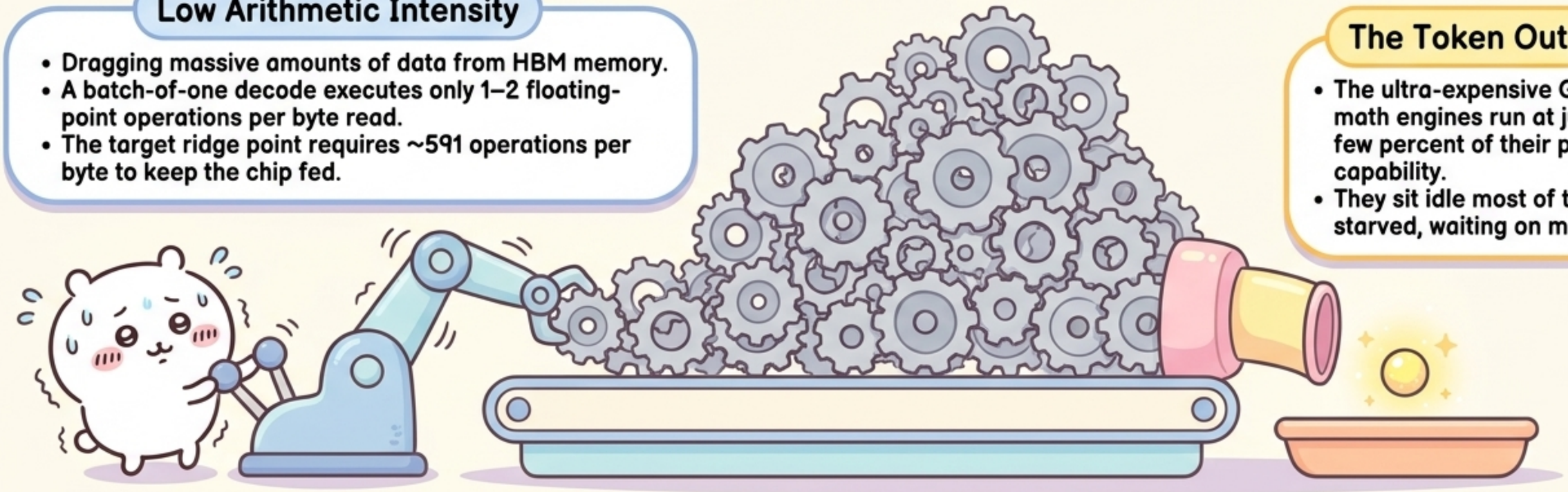
The Cost of Proximity: Distance is latency.
To achieve the immense bandwidth of HBM, manufacturers must vertically stack DRAM dies and wire them physically next to the logic die. It is a low-yield, brutally expensive miracle of packaging.

Token Generation is a Memory Problem, Not a Math Problem

The Decode Phase breaks the system: Generating an answer requires dragging the entire model weight set across the memory bus just to produce a single token.

Low Arithmetic Intensity

- Dragging massive amounts of data from HBM memory.
- A batch-of-one decode executes only 1–2 floating-point operations per byte read.
- The target ridge point requires ~591 operations per byte to keep the chip fed.



The Token Output

- The ultra-expensive GPU math engines run at just a few percent of their peak capability.
- They sit idle most of the time, starved, waiting on memory.

The Reality: Training uses dense math and loves FLOPS. But Decode—the production phase where AI serves billions of answers—is overwhelmingly memory-bandwidth-bound.

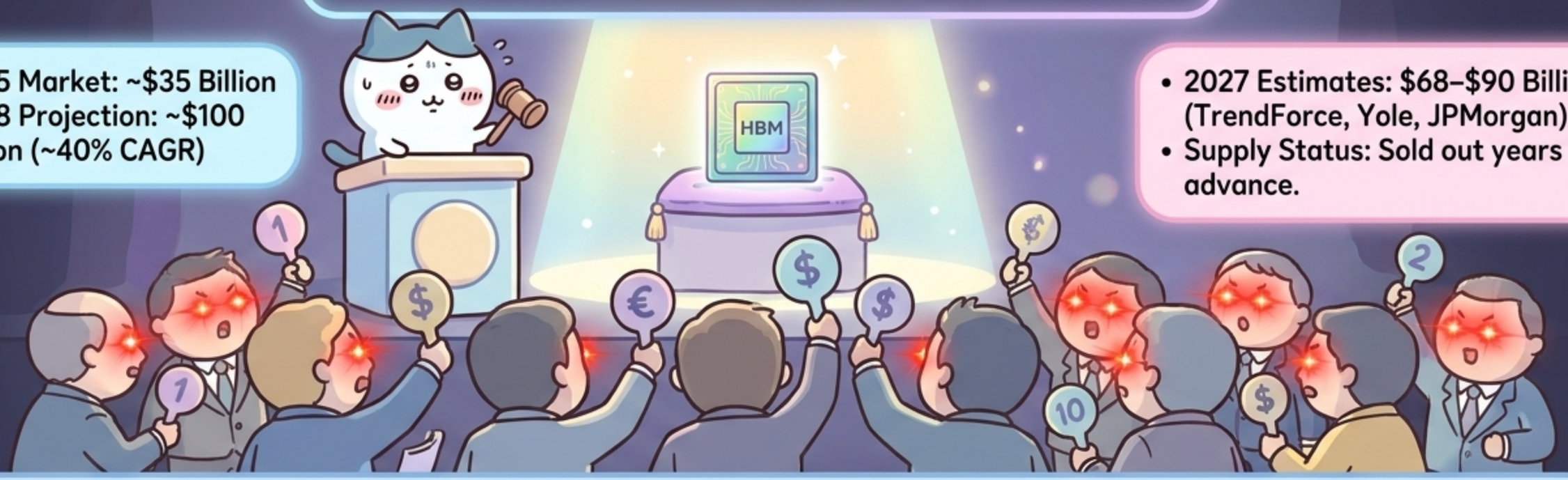
Follow the Money: The High-Bandwidth Supercycle

Multiplier Equation

More Memory per GPU (80GB in 2022 ➜ 288GB in 2025)
× More Total Chips Sold
× Higher Price per Gigabyte
= The Ultimate Hardware Prize

- 2025 Market: ~$35 Billion
- 2028 Projection: ~$100 Billion (~40% CAGR)

- 2027 Estimates: $68–$90 Billion (TrendForce, Yole, JPMorgan)
- Supply Status: Sold out years in advance.



The New Spec Sheet: For the past decade, the industry asked who had the most FLOPS. Today, the value pools around whoever supplies the memory. The only question that matters for AI infrastructure is: **Who can feed them?**