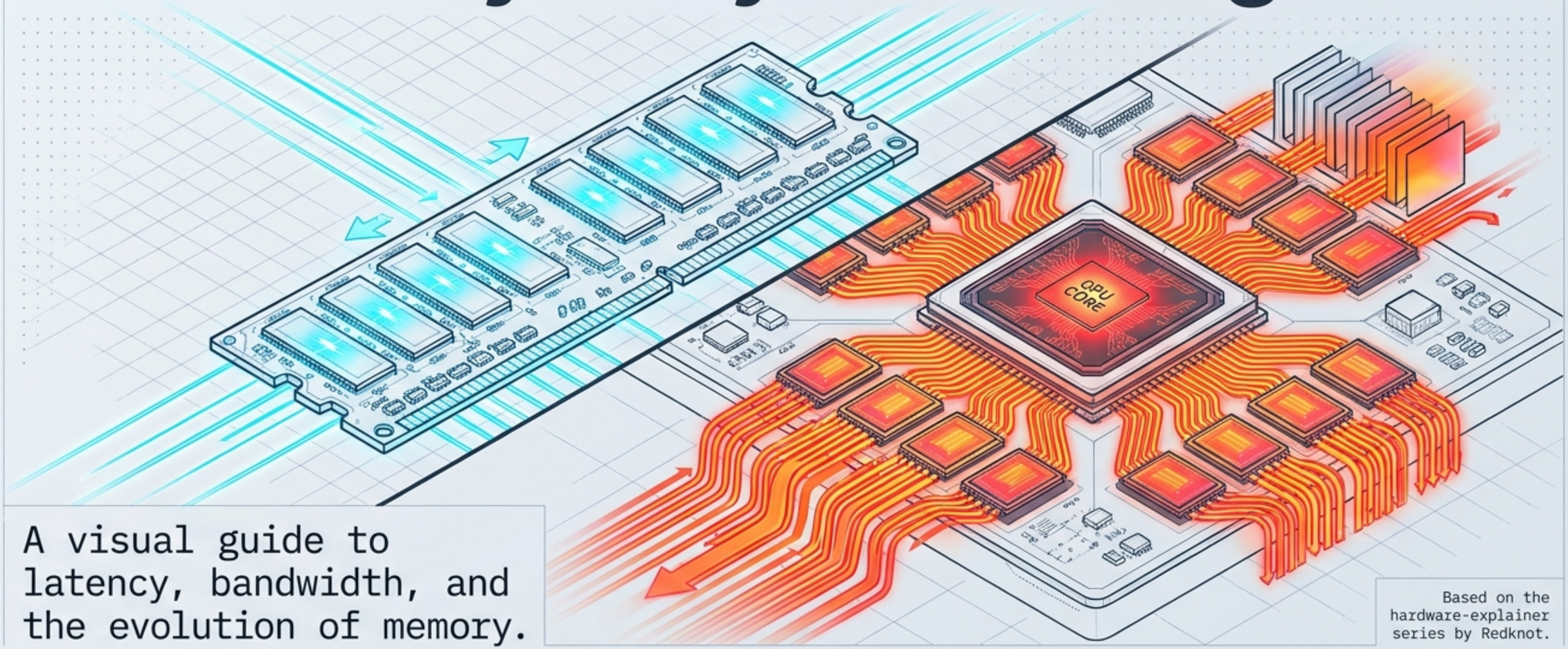
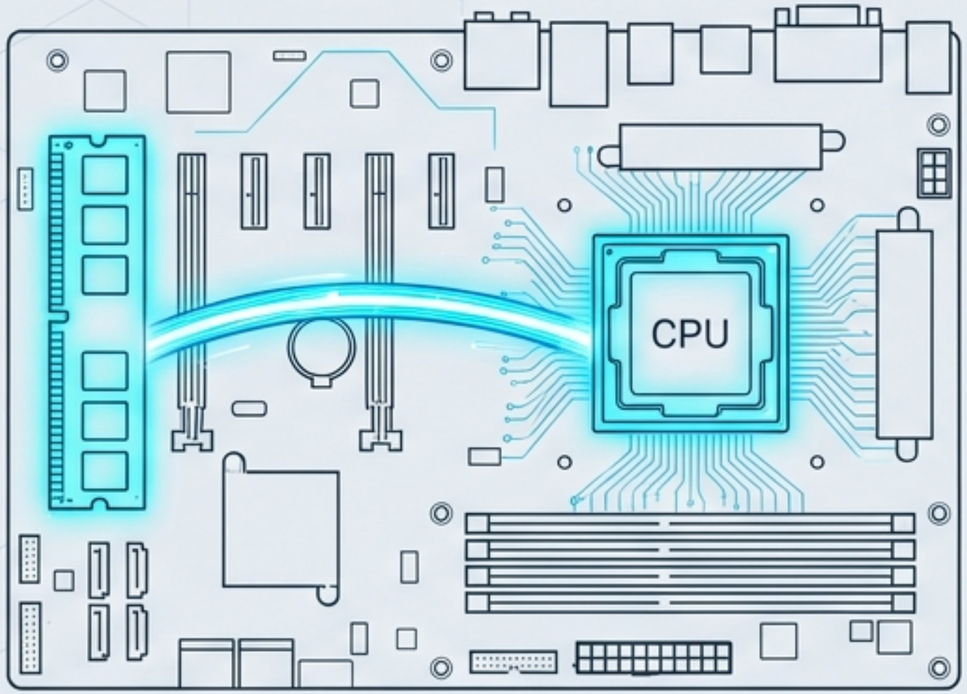
RAM and VRAM Both Store Data. Here's Why They're Nothing Alike.



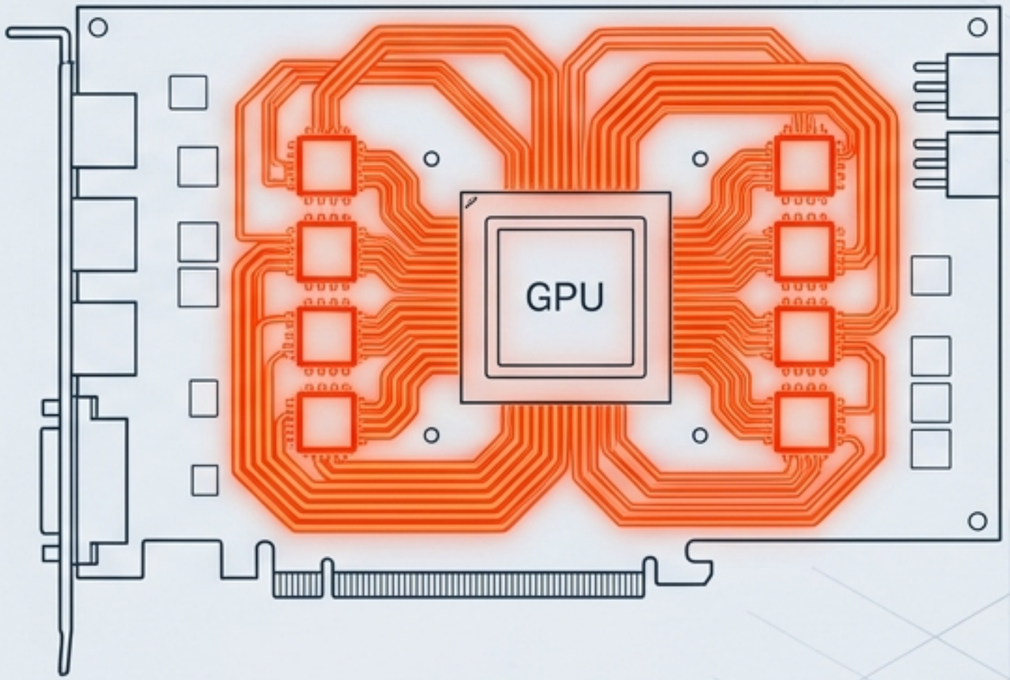
A visual guide to latency, bandwidth, and the evolution of memory.

Based on the hardware-explainer series by Redknot.

The jobs sound identical, but the spec sheets look like they are from different planets.



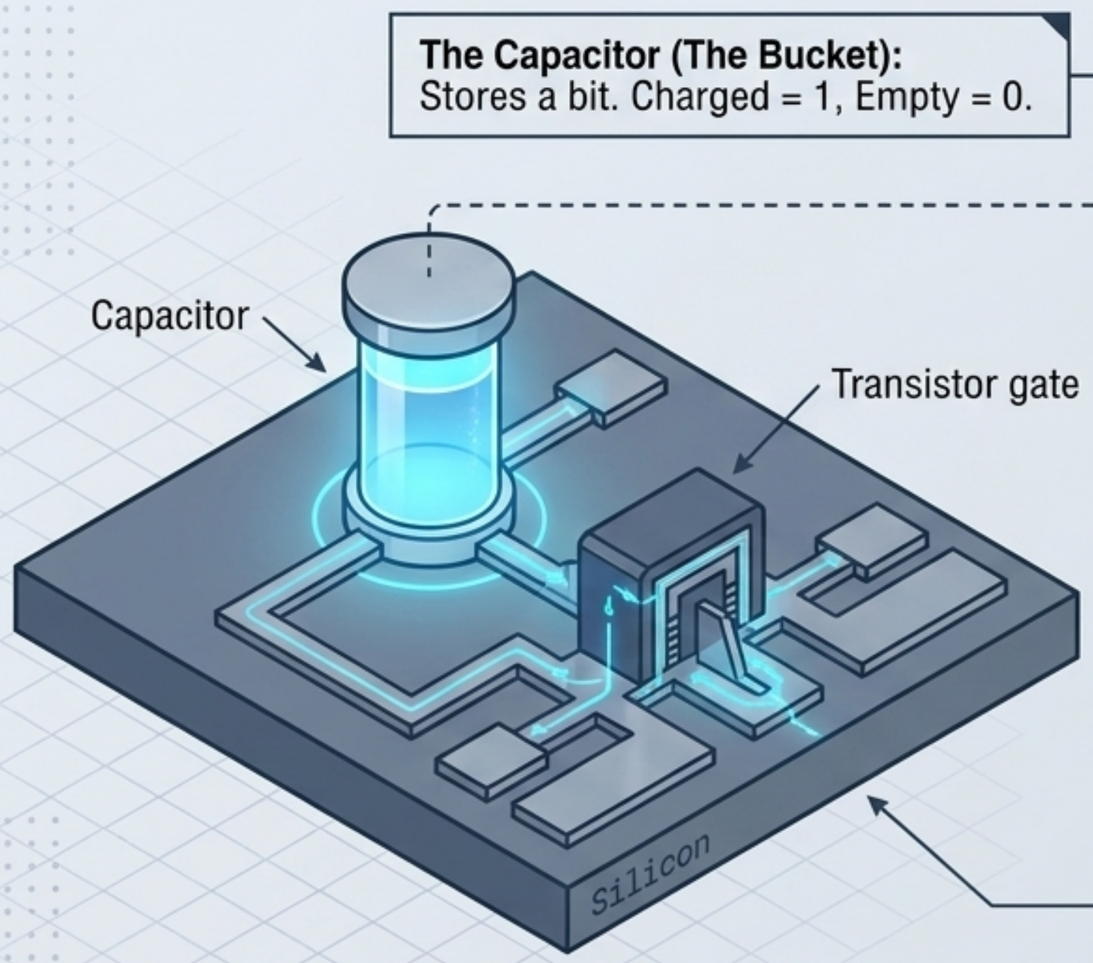
Main Memory: Buffering data for the Central Processing Unit.



Video Memory: Buffering data for the Graphics Processing Unit.

At the absolute foundation, all dynamic memory is just a leaky bucket and a valve.

The Capacitor (The Bucket):
Stores a bit. Charged = 1, Empty = 0.

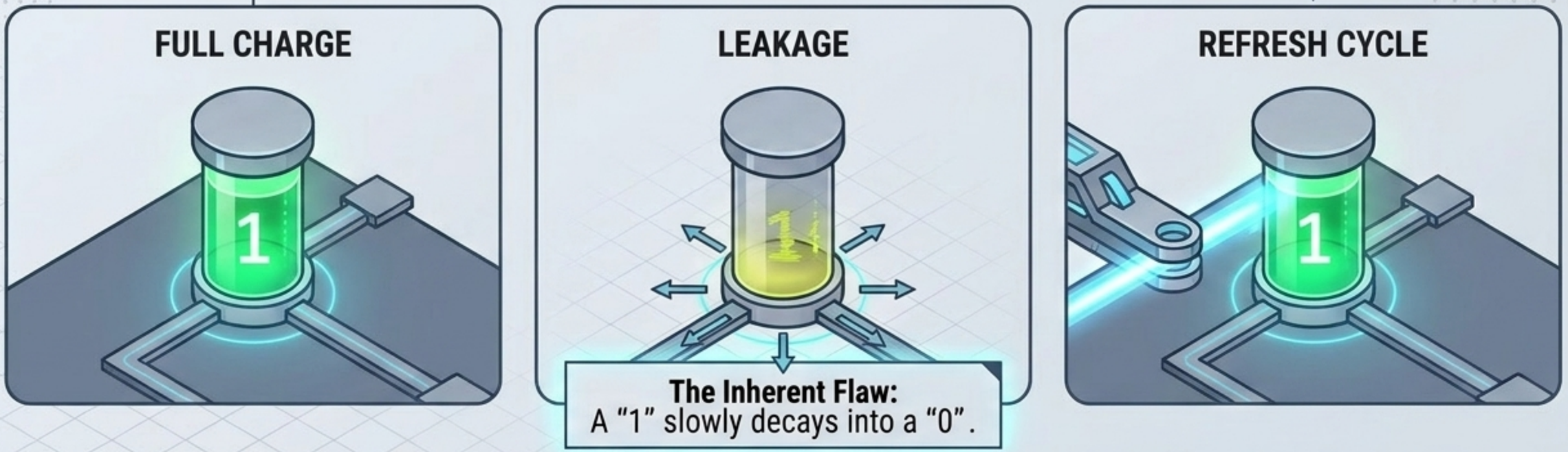


The Transistor (The Valve):
Acts as a gate. When turned on, the capacitor talks to the outside circuit; when off, it sits quietly.

INSIGHT NOTE: If wired straight into a circuit without the transistor, the capacitor would leak charge everywhere and disturb its neighbors.

The constant need to recharge memory cells makes DRAM inherently slower than the processor.

The Refresh Penalty



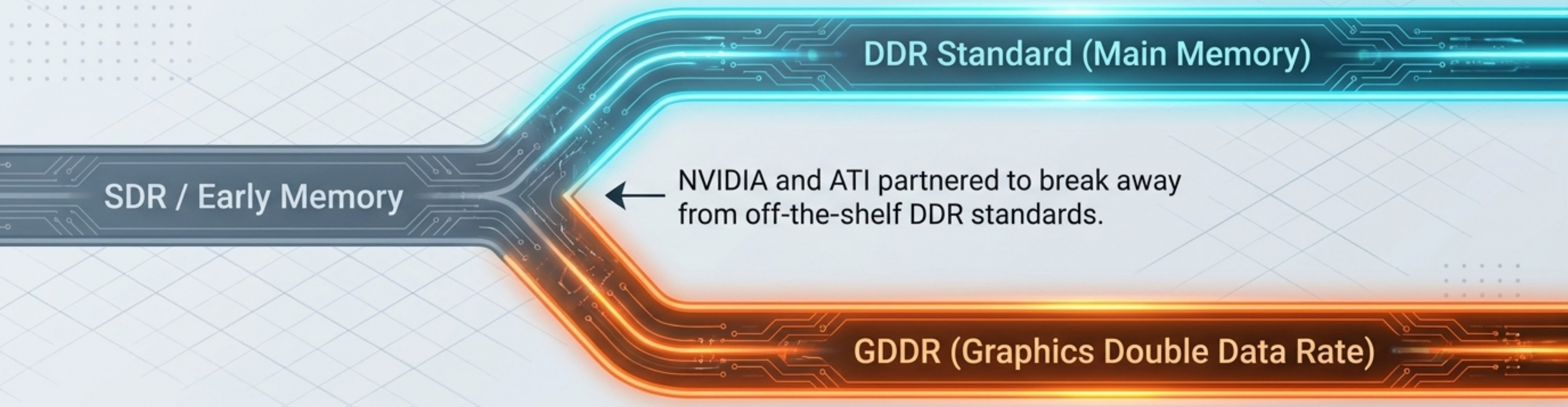
Bit stored as ‘1'.

Voltage leaks, value decays.

Data re-written and recharged.

Because of this leak, DRAM must read every cell and write it back on a strict schedule. This constant refreshing, layered on top of fussy read/write operations, ensures memory will always lag behind the CPU.

When graphics processors got hungry, standard memory couldn’t feed them.

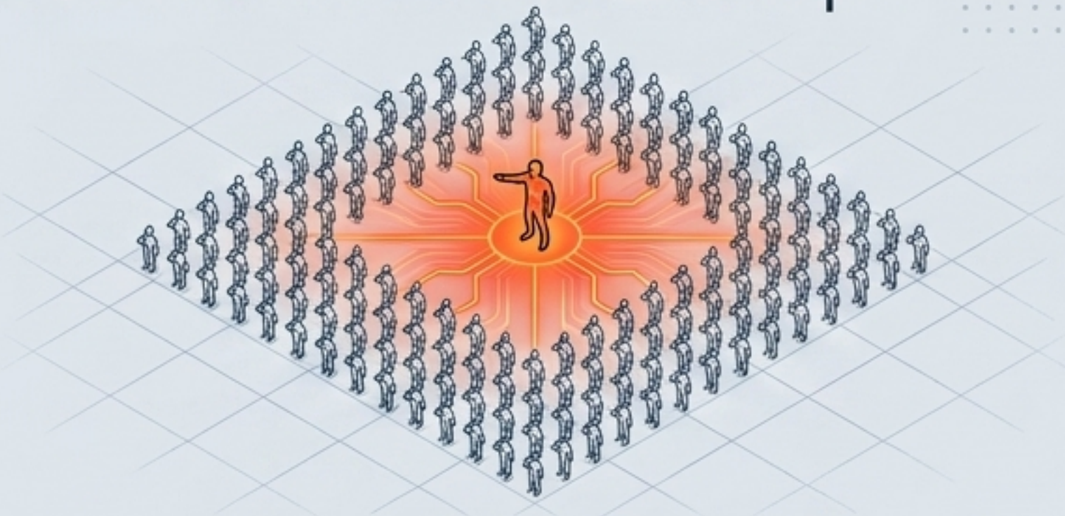


The split wasn’t about the memory itself. It was about the radically different temperaments of the masters they served: the CPU vs. the GPU.

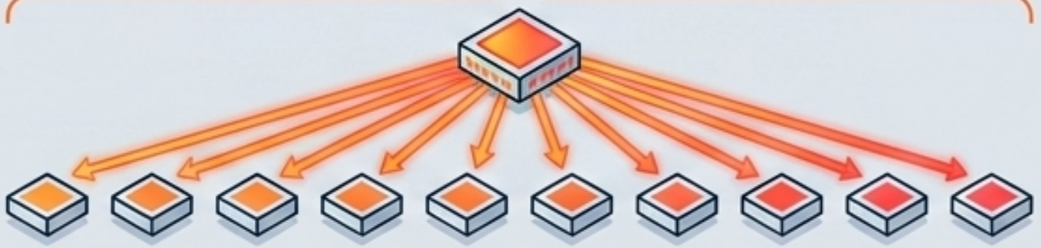
A CPU carries water by the bucket.
A GPU commands a thousand soldiers in lockstep.

Summing 1 to 100

CPU: Few mighty cores. Pulls one piece of data, processes it, writes it back, then fetches the next.



Brighten every pixel in a photo simultaneously



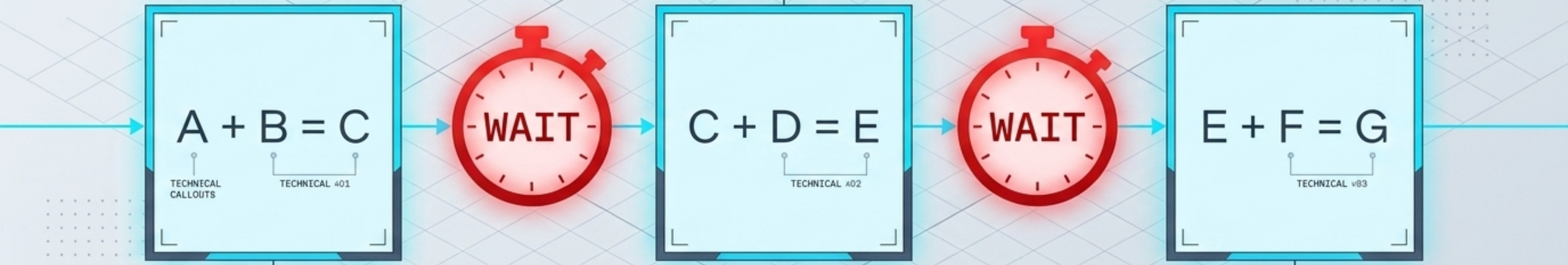
GPU: Thousands of weak cores. The controller hands out work, and on one command, they all dash to memory, compute, and write back together.

The Diverged Families of System Memory

	Main Memory (RAM)	Video Memory (VRAM)
Partner Processor	Central Processing Unit (CPU) 	Graphics Processing Unit (GPU) 
Workload Type	Serial (Sequential)	Parallel (Concurrent)
The Primary Metric	Obsesses over Latency	Obsesses over Bandwidth
Bus Width	Narrow (Forever 64-bit) 	Massive (384 to 512-bit) 
Latency Tolerance	Glacial (50-80ns is too slow)	Highly Tolerant (>200ns is fine)

CPUs do strictly serial work where bandwidth is useless and latency is everything

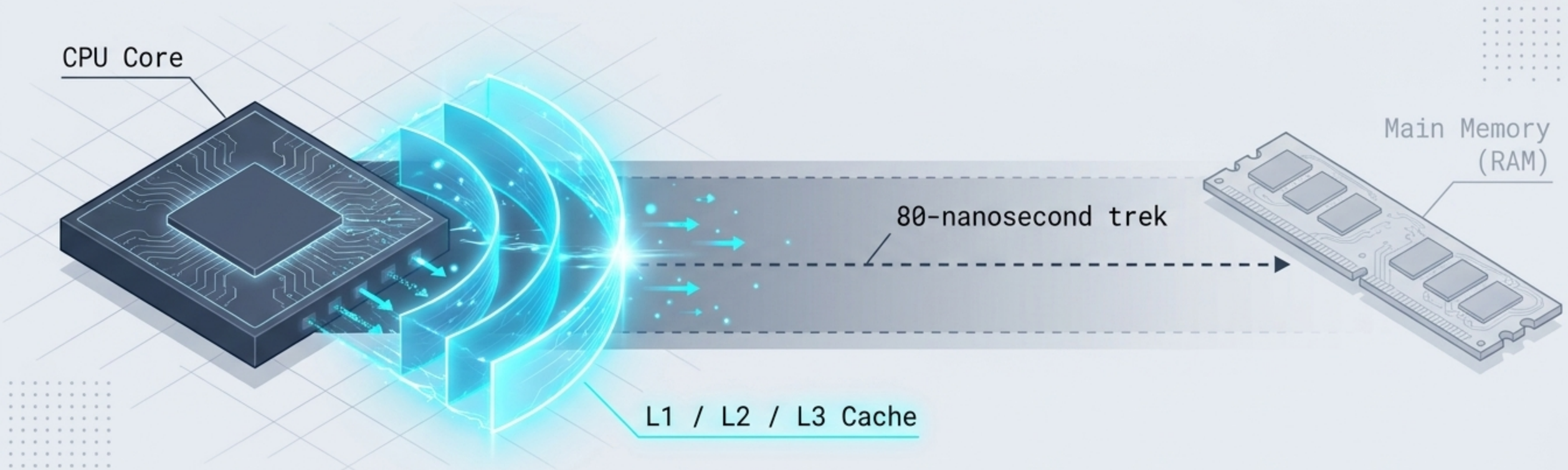
If every step requires the result of the previous step, you cannot split 99 math problems across 99 cores.



Latency: The gap between asking for data and getting it. For a CPU, **50 to 80 nanoseconds** is a glacial wait.

Because the data payloads are tiny, widening the road (bandwidth) doesn't make the trip any faster.

Because main memory is too slow, processors bolt on massive caches to intercept the trip.



To avoid the 80-nanosecond trek to standard memory, modern CPUs hoard immediate data in ultra-fast, on-die SRAM cache.

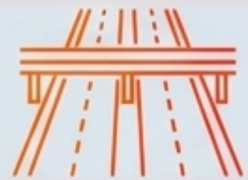
Main memory's design remains stagnant—sticking to a 64-bit bus and merely nudging clock frequencies—because the CPU relies on cache to dodge memory trips entirely.

Video memory operates on the physics of shipping freight.

Bandwidth
[Cargo Delivered]

=

Bus Width
[Number of Roads Built]



×

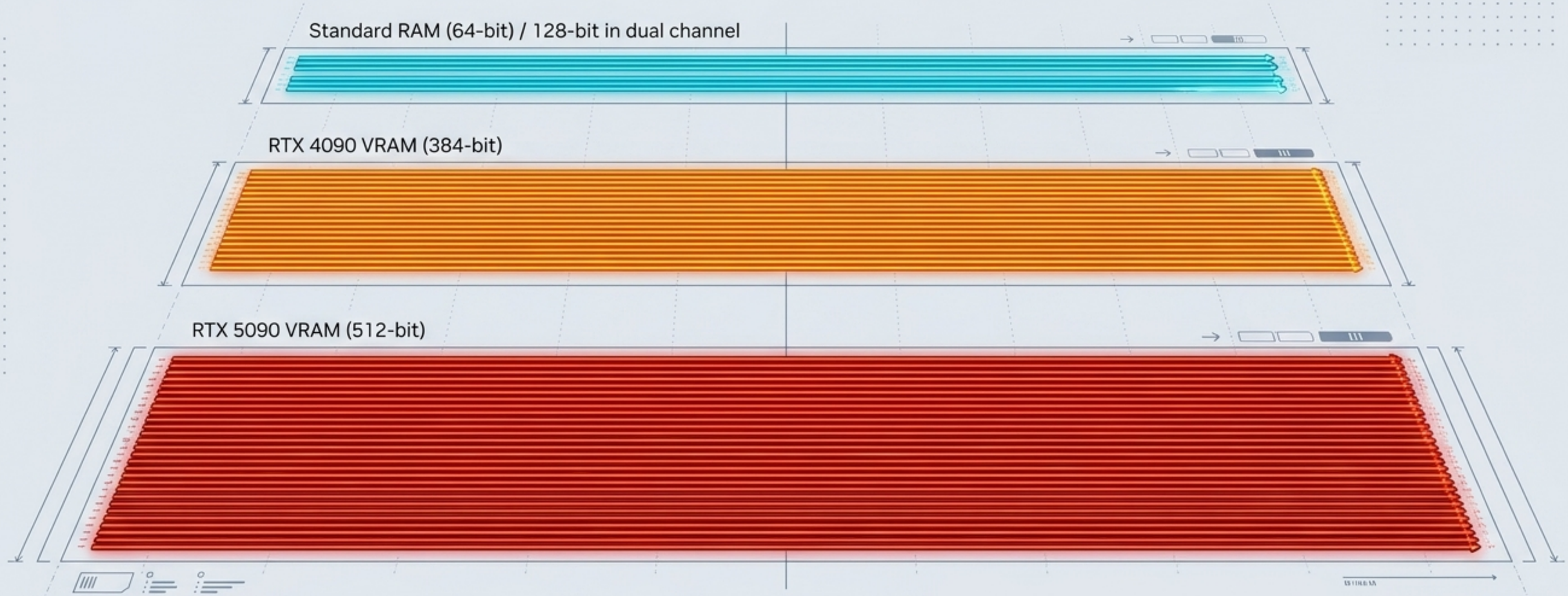
Transfer Rate
[Speed of the Trucks]

×

Signaling
[Cargo per Truck]

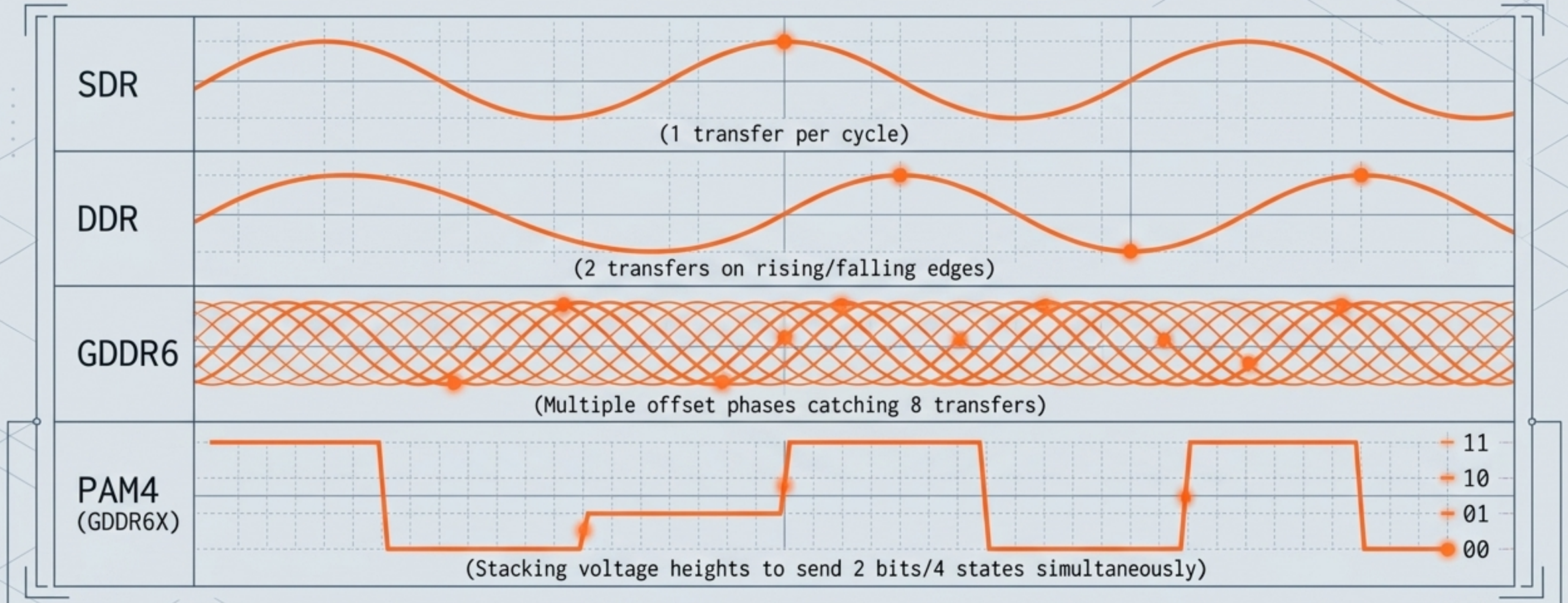
A GPU's 'grab it all at once' appetite requires moving massive volumes of data simultaneously.
To achieve this, VRAM engineers manipulate all three variables of the highway.

Variable 1: Build a wider highway.



While main memory has been frozen at a 64-bit bus width for decades, video memory routinely crushes it by building hundreds of physical parallel roads.

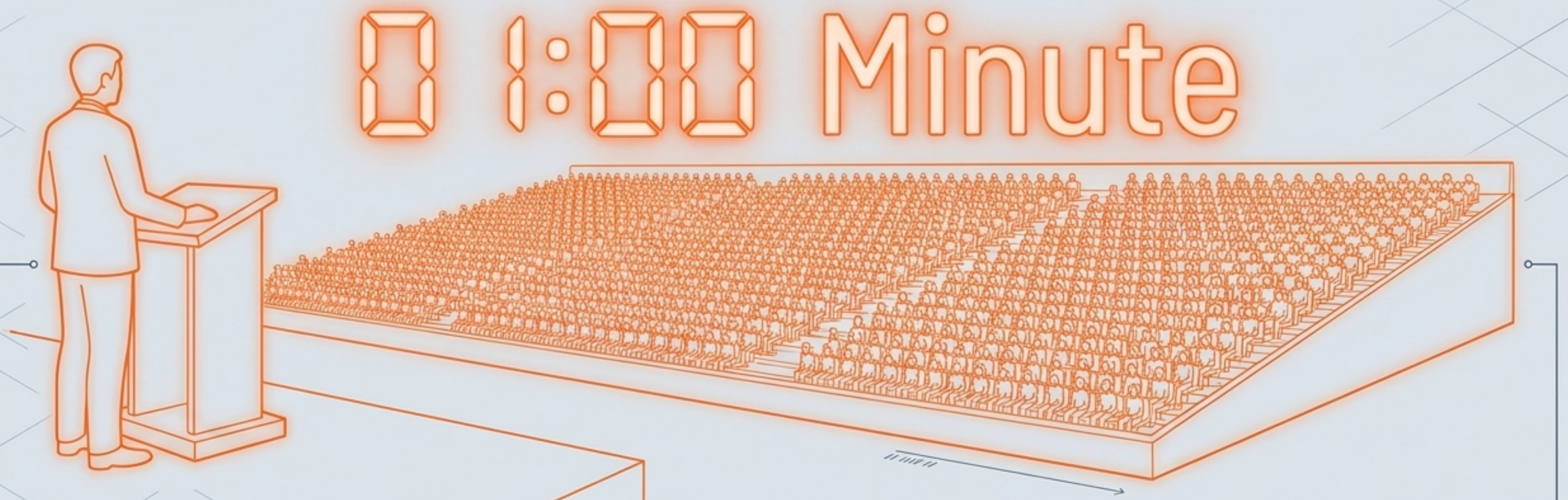
Variable 2 & 3: When trucks max out their speed, you stack the cargo.



Trucks topping out near the speed of light can’t floor it any harder. You have to pack more trucks into the gaps.

PAM4 was so aggressive it risked splitting its pants. By GDDR7, engineers retreated to a steadier three-level signal (PAM3).

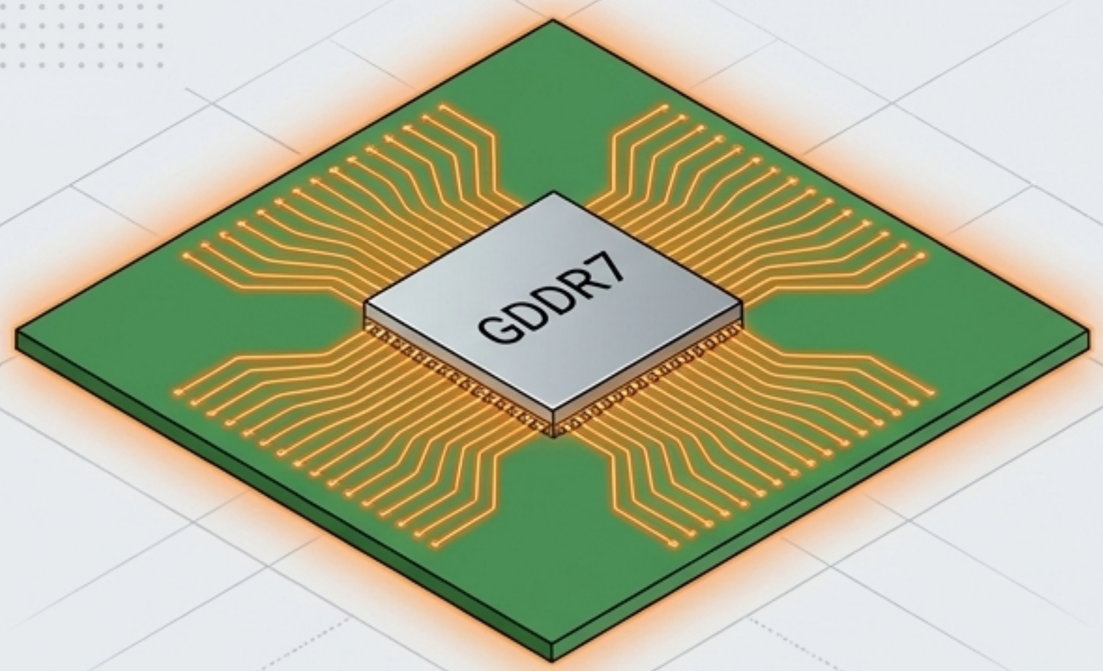
The Latency Illusion: Why a 200-nanosecond wait doesn't matter to a GPU.



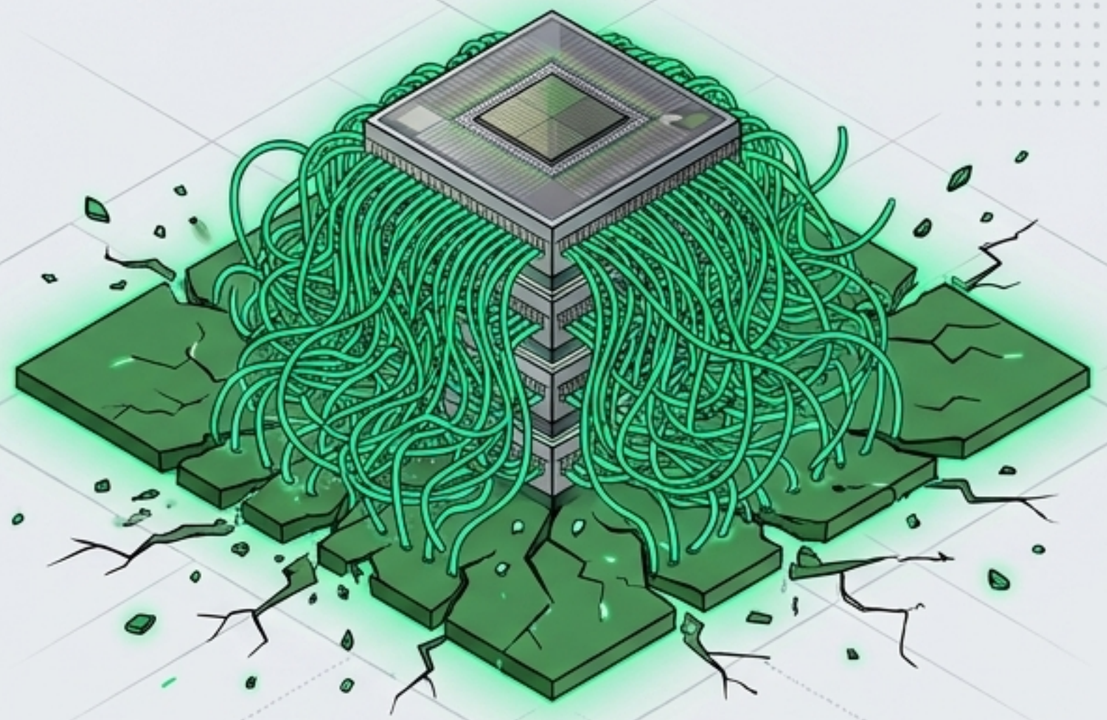
The Old Teacher's Math: "Me talking for 1 minute wastes 1 minute for all 45 classmates, so 45 minutes are wasted." False. It is just 1 minute.

If 10,000 GPU cores all request data concurrently, a 200ns latency penalty is absorbed simultaneously. They all wait together, making the delay irrelevant in the face of massive bandwidth.

Artificial Intelligence demanded more data lanes than a flat 2D board could physically hold.



A regular GDDR7 chip is 32 bits wide. One HBM3E stack requires 1,024 data bits.



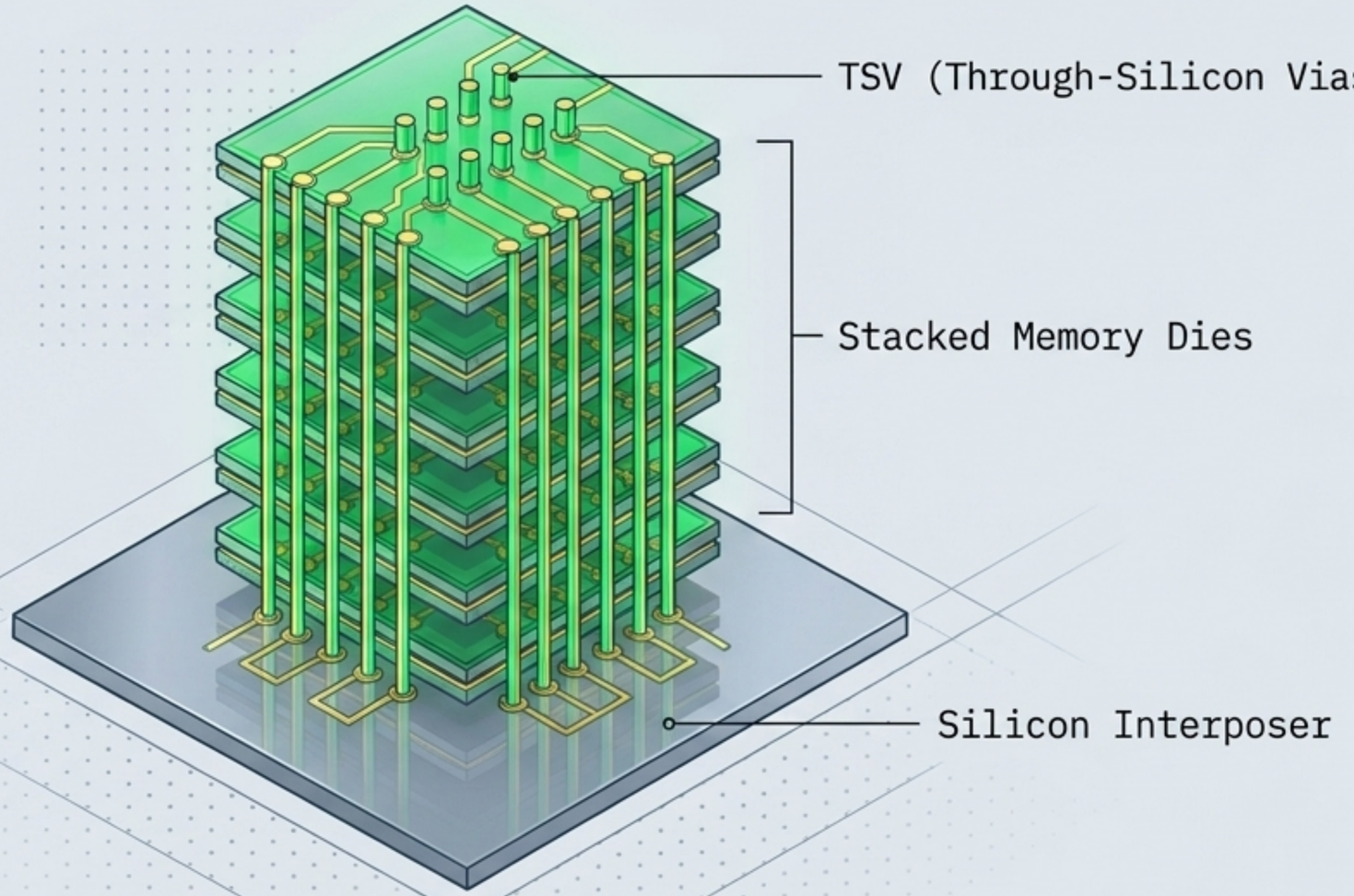
Factor in power, address, and clock lines, and HBM needs nearly 4,000 wires. Standard fiberglass substrates simply cannot fan that out.

High Bandwidth Memory solves the bottleneck by abandoning the motherboard for the Z-axis.

The fix: Swap the fiberglass for a slab of silicon **polished smoother than a mirror.**

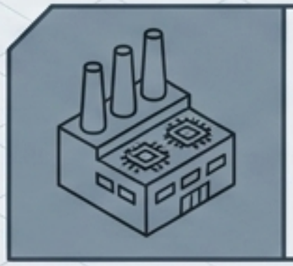
Instead of placing chips side-by-side, dies are stacked vertically and drilled through with microscopic copper pillars (TSVs) to bolt the memory directly onto the GPU's roof.

Cost of Scale: One GB300 server node holds >1TB of HBM and costs 2 million RMB—largely due to the lump of HBM on its roof.



When artificial intelligence consumes global manufacturing capacity, the consumer pays the difference.

AI Server Farms Demand Massive HBM Output.



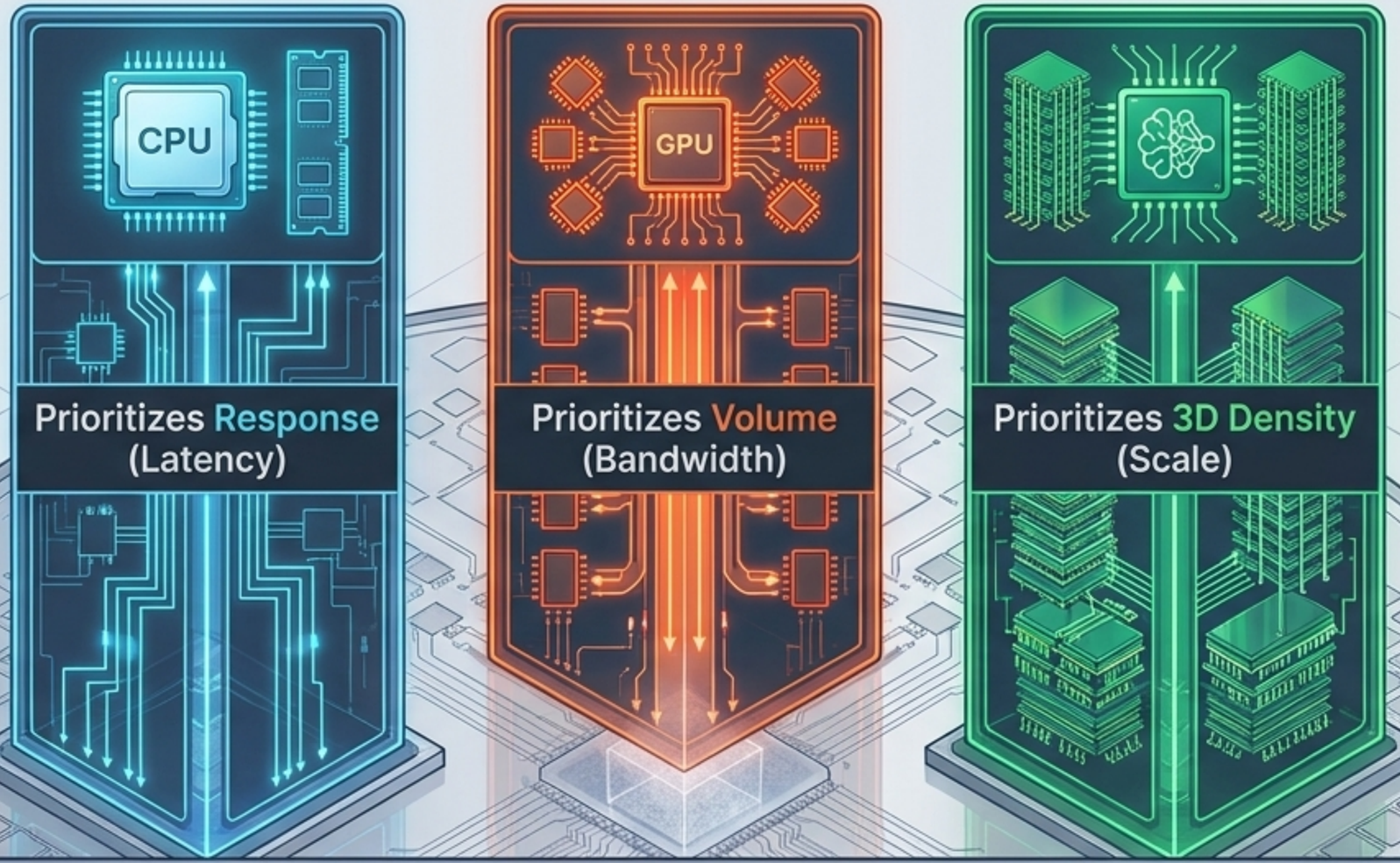
HBM Production Lines Overlap Heavily with Standard RAM.

Consumer Supply Shrinks → Prices Rise.

HBM isn't created in a vacuum. It shares the exact same silicon fabrication lines as ordinary main memory.

The extra money you paid on your last RAM upgrade was partly footing AI’s bill.

The master dictates the architecture.



Whether it’s the solitary sprinter of the CPU, the marching army of the GPU, or the massive data models of AI, all memory starts as a leaky capacitor on a slab of silicon. Physics provides the building blocks. The master dictates the architecture.