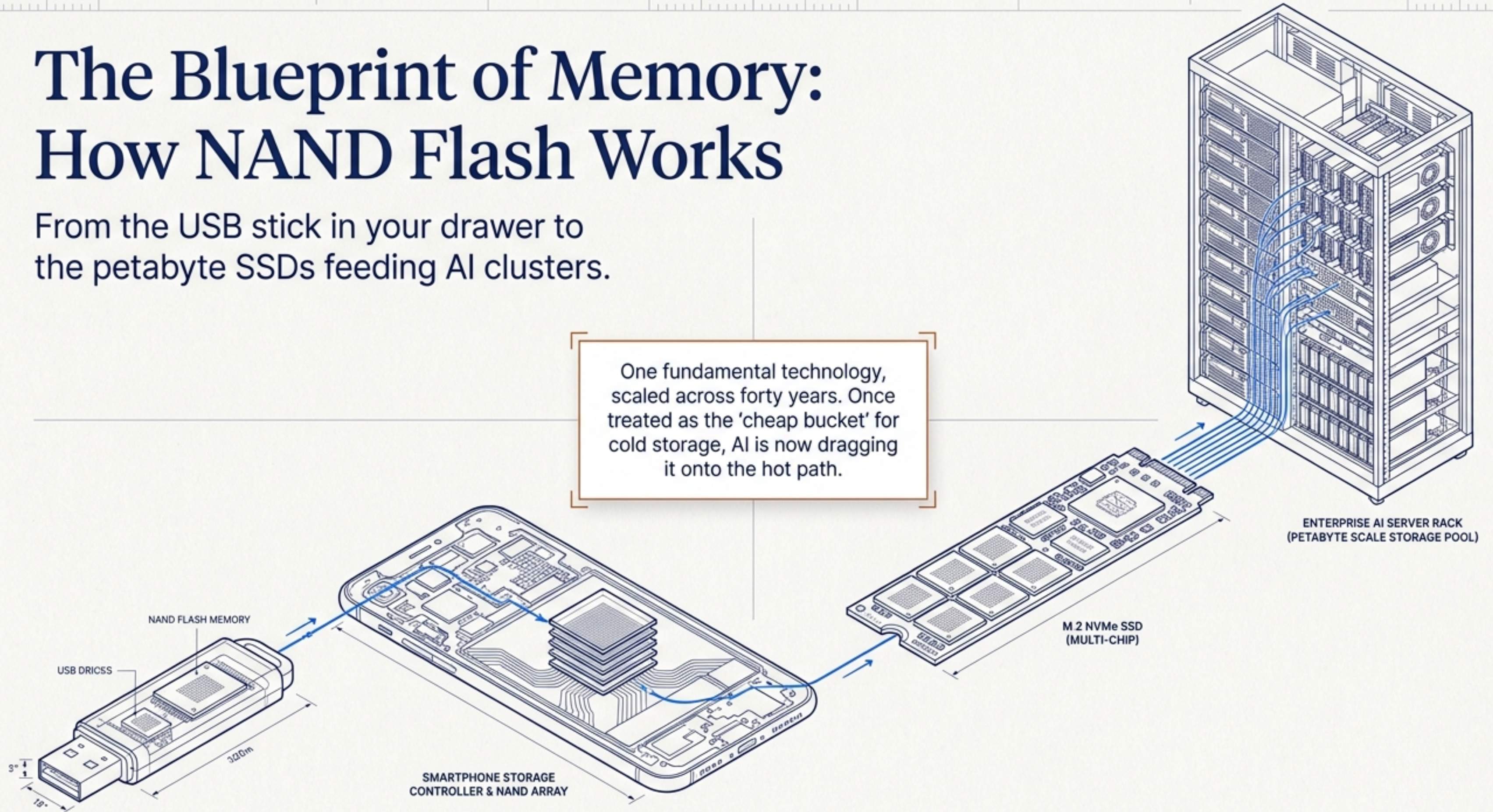


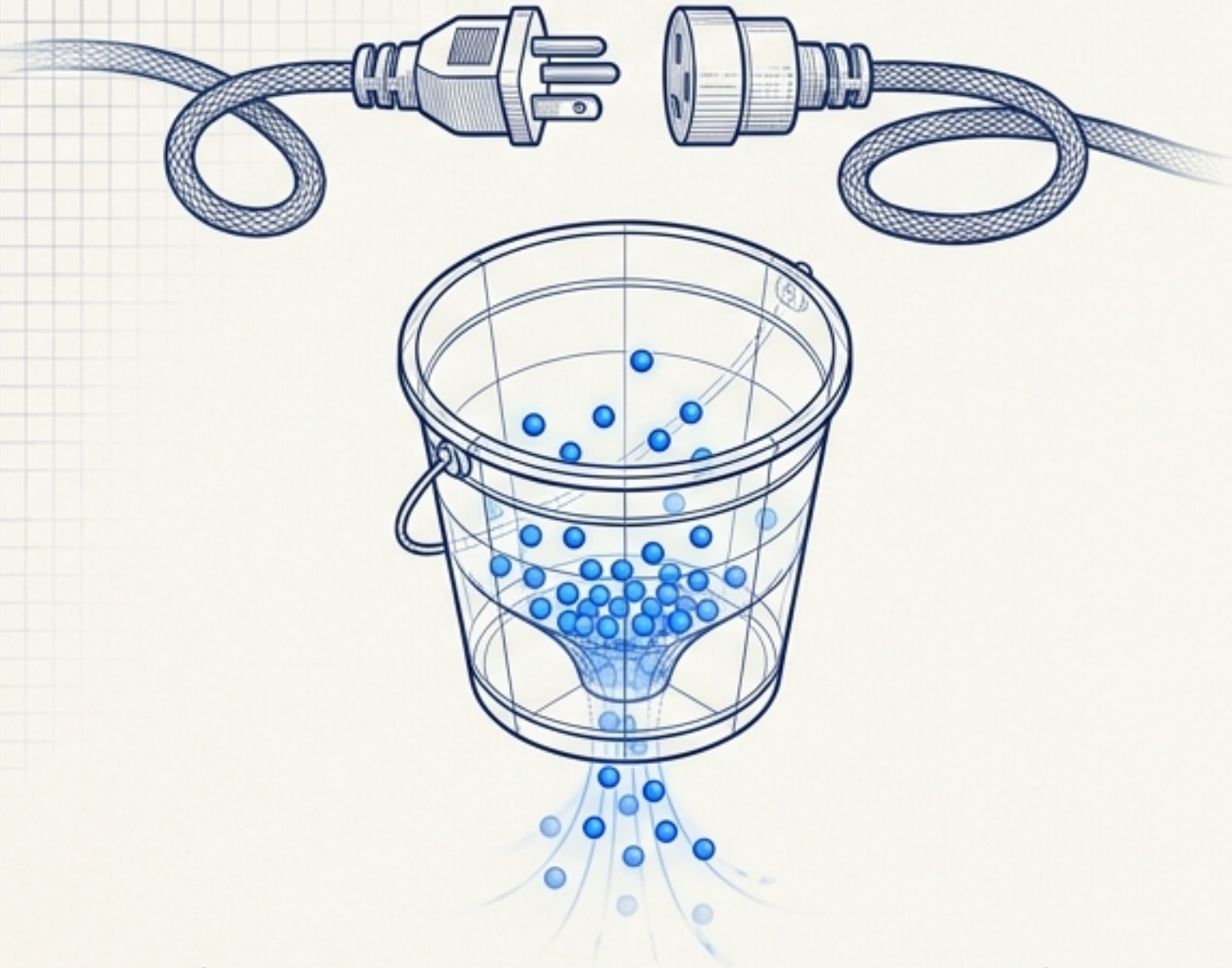
The Blueprint of Memory: How NAND Flash Works

From the USB stick in your drawer to the petabyte SSDs feeding AI clusters.

One fundamental technology, scaled across forty years. Once treated as the 'cheap bucket' for cold storage, AI is now dragging it onto the hot path.

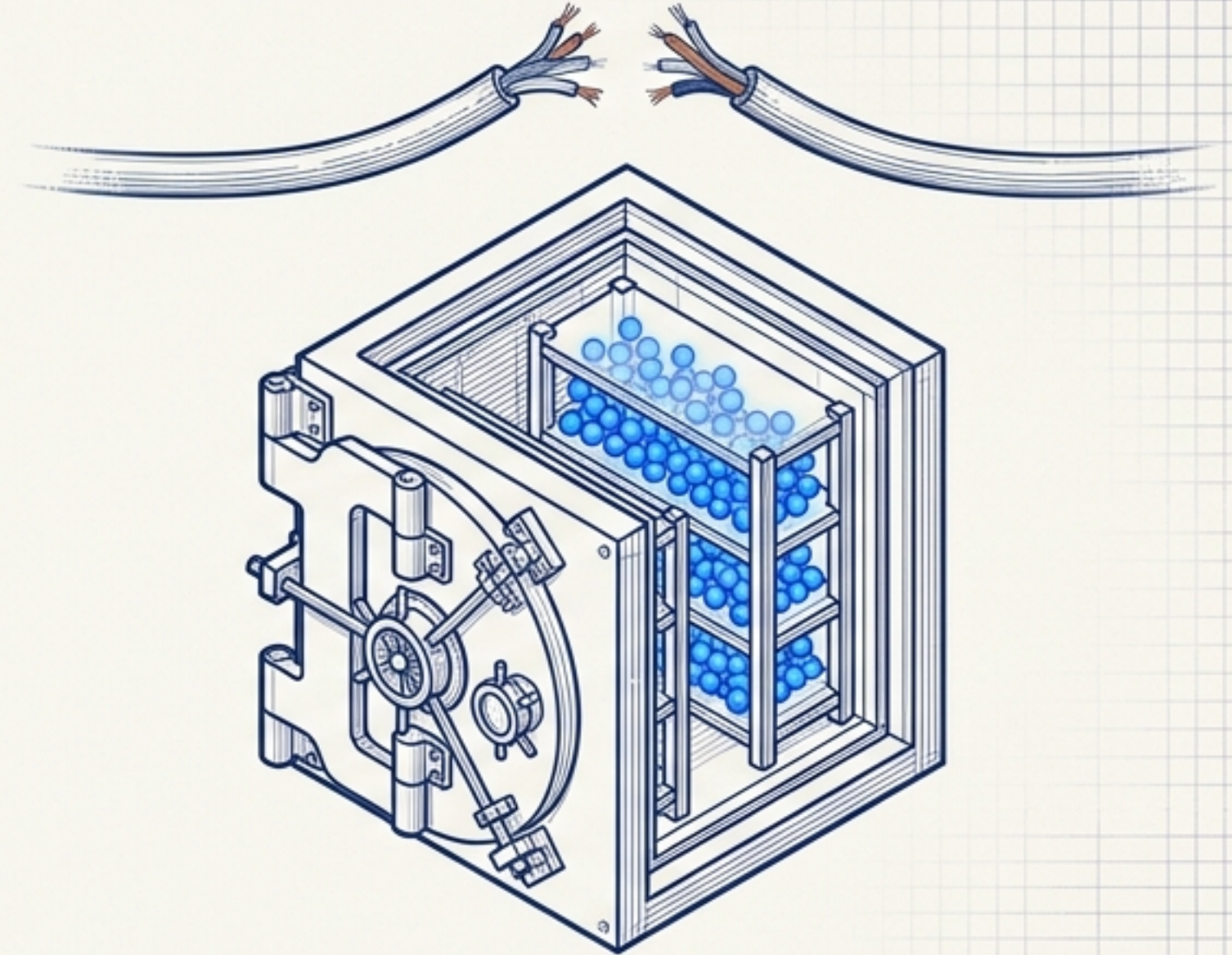


The Physics of Permanence: Non-Volatile vs. Volatile Memory



DRAM

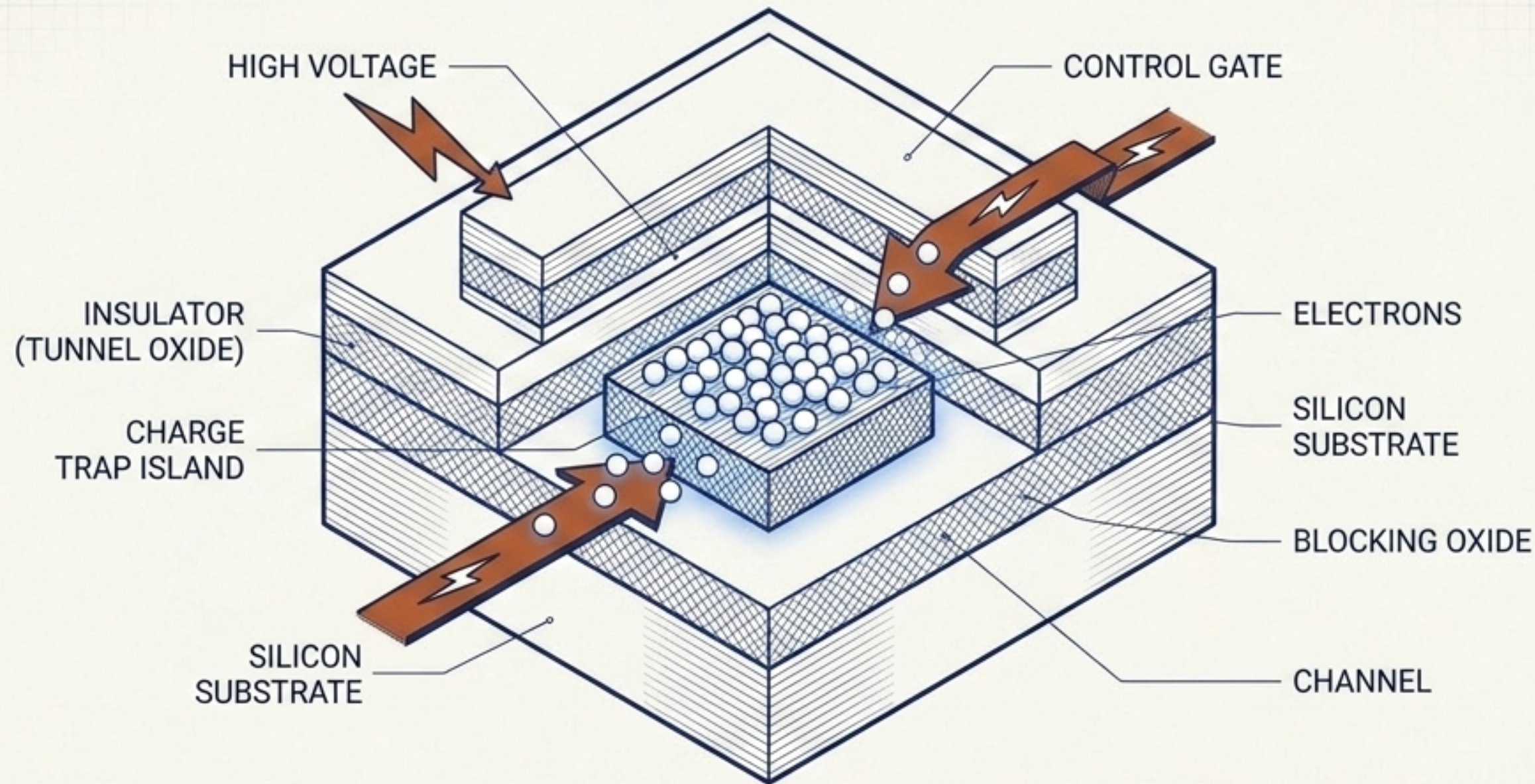
Fast, forgetful working memory. Loses everything in a fraction of a second without power.



NAND

Holds data for years without power. The foundation of modern storage.

Trapping the Electron: The Island Behind the Insulator



1. High Voltage Insertion



Electrons are forced across an insulating barrier using extreme electrical pressure.



2. The Isolated Island

The charge gets stranded on a completely insulated middle layer.



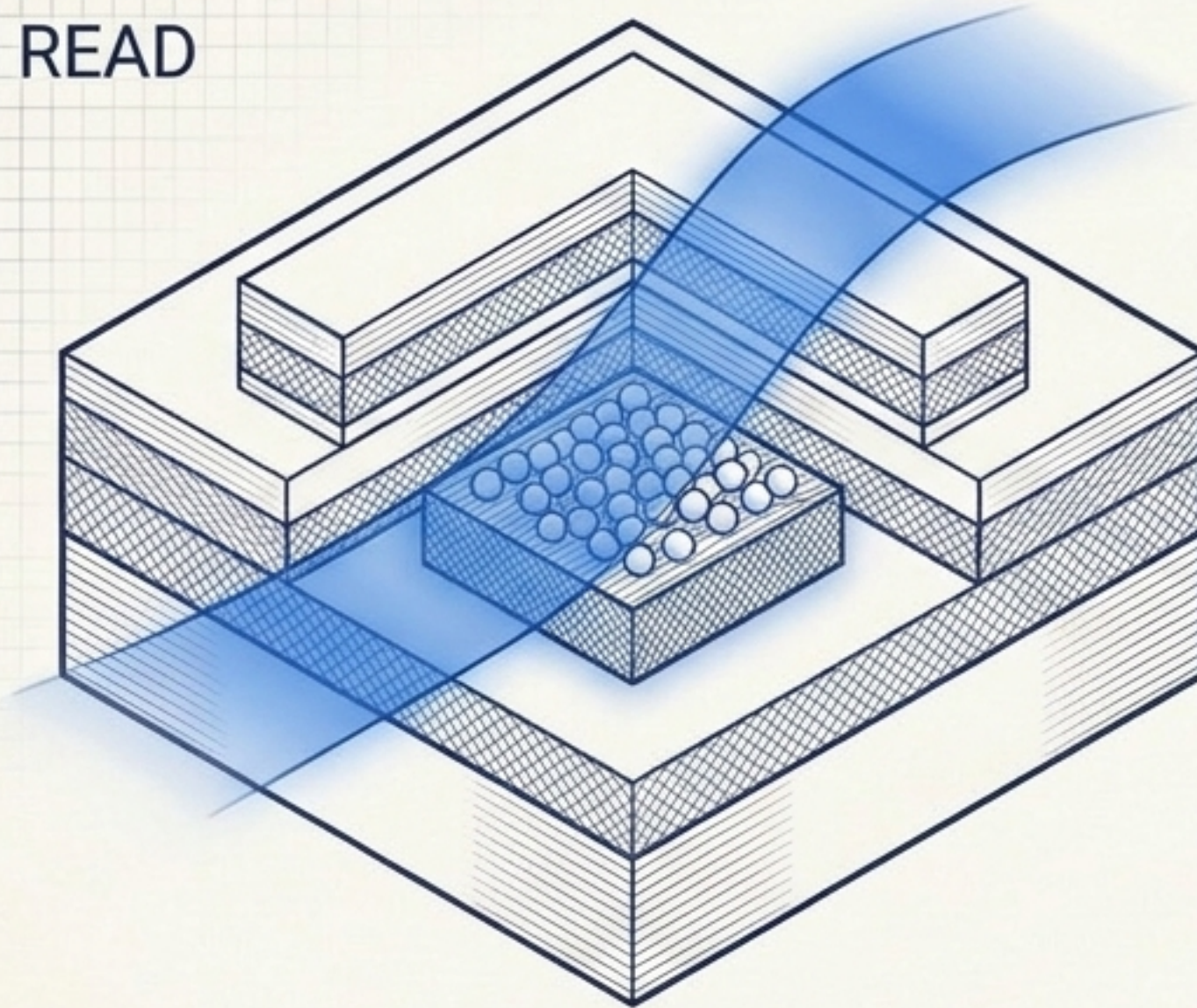
3. Permanent Memory

With nowhere to go, the exact amount of charge trapped becomes the physical memory of the cell.



The Crucial Asymmetry: Reads are Gentle, Writes are Violent

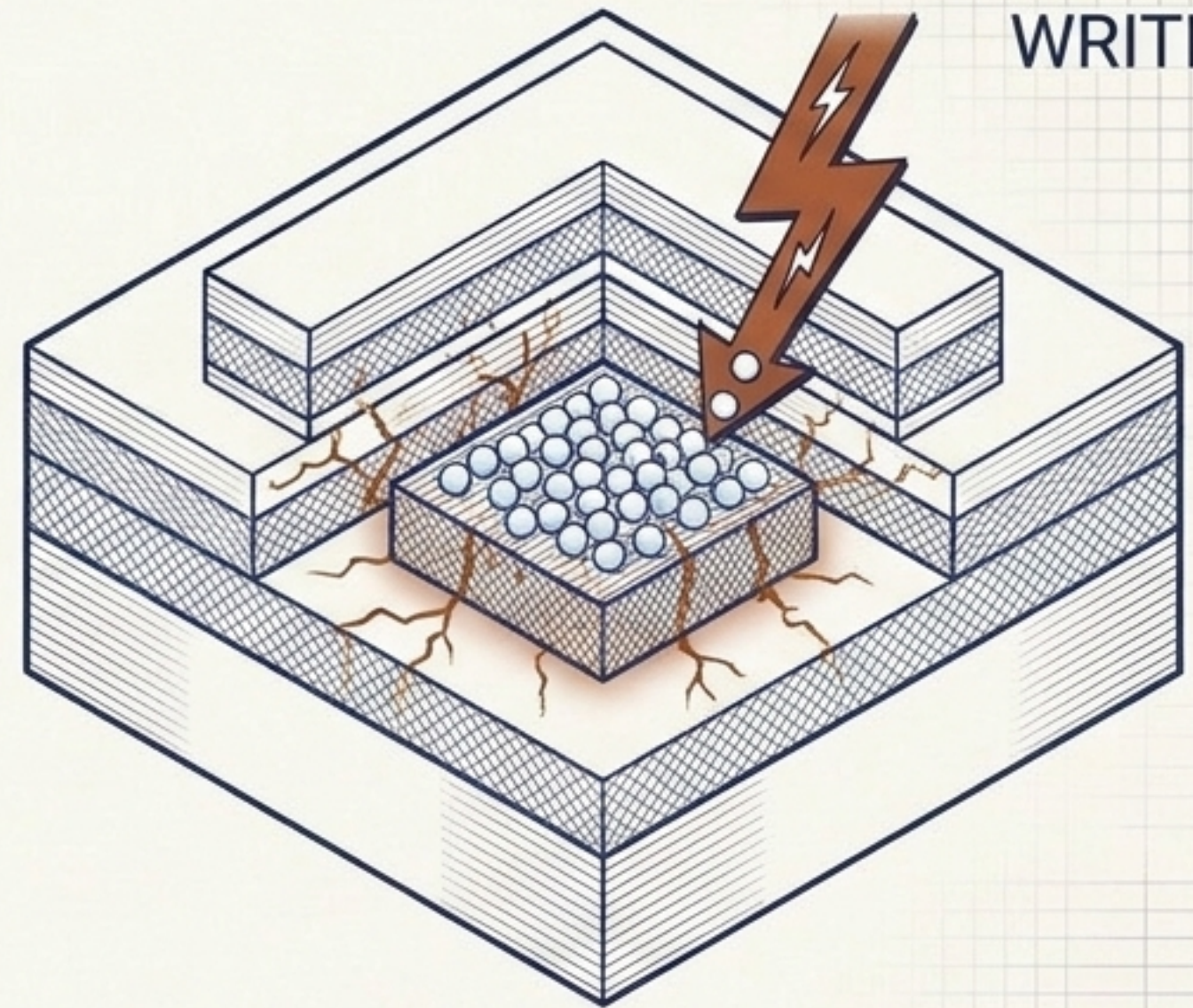
READ



READ

Cheap, fast, harmless. Measures conductivity without moving electrons.

WRITE

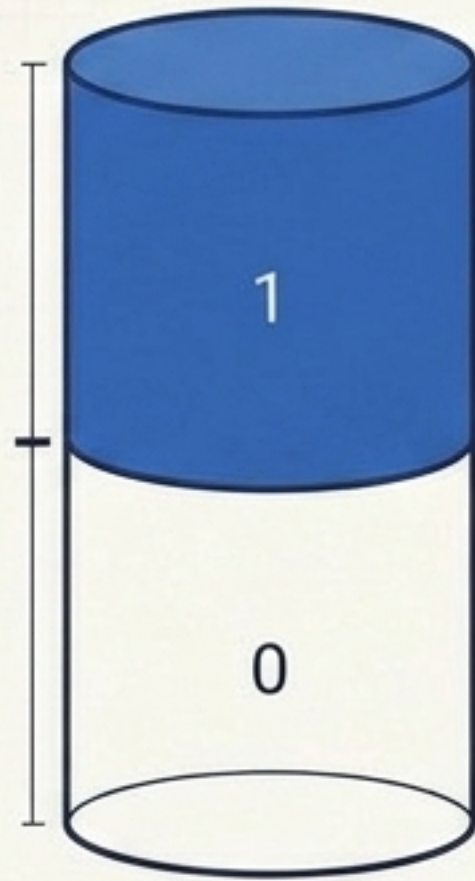


WRITE

Expensive, high-voltage forced entry via Fowler-Nordheim tunneling.

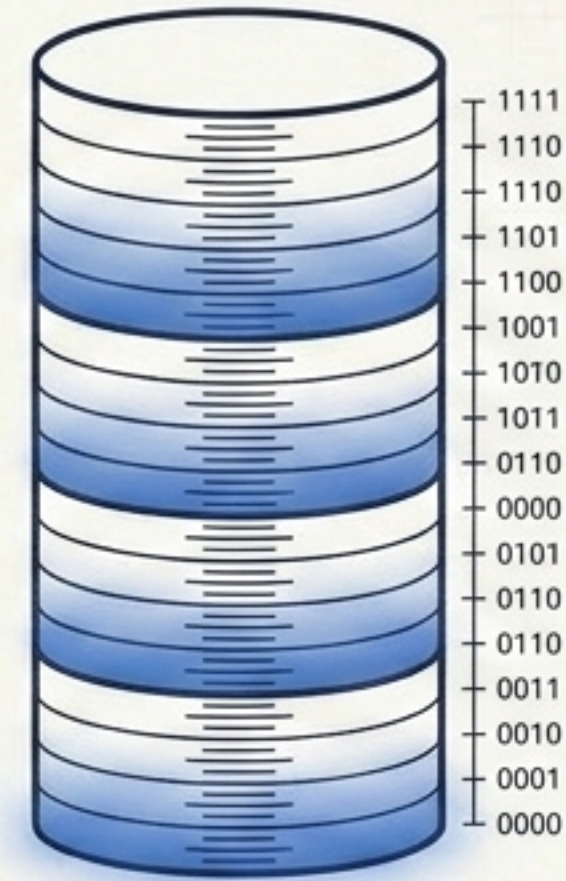
Flash doesn't wear out from sitting still.
It wears out from being written.

Cramming the Cell: From Binary to Analog



SLC (1 Bit)

Only two states to detect:
Empty or Full.



QLC (4 Bits)

Sixteen voltage levels packed into
the exact same physical space.

Why store just one bit per cell? If you can detect exactly how much charge is on the island, a single physical cell can hold multiple bits. Same silicon, more data, vastly lower cost per gigabyte.

The Endurance Geometry Matrix

Type	Bits/cell	Voltage levels	P/E Cycles (Endurance)	Role
SLC	1 Bit	2 Levels	~50,000–100k	Industrial reliability
MLC	2 Bits	4 Levels	~3,000–10,000	Premium enterprise
TLC	3 Bits	8 Levels	~1,000–3,000	Mainstream today
QLC	4 Bits	16 Levels	~100–1,000	High-capacity, cheapest
PLC	5 Bits	32 Levels	Experimental	Lab tech

Slicing the voltage window into 16 levels leaves almost no margin for error. A tiny amount of insulator damage causes the cell to read the wrong number.

The Ironclad Rule of Flash Memory



"Every bit of density you buy, you pay for in endurance."

Slower Writes

Charge must be carefully nudged and verified to hit exact micro-levels without overshooting.

Fragility

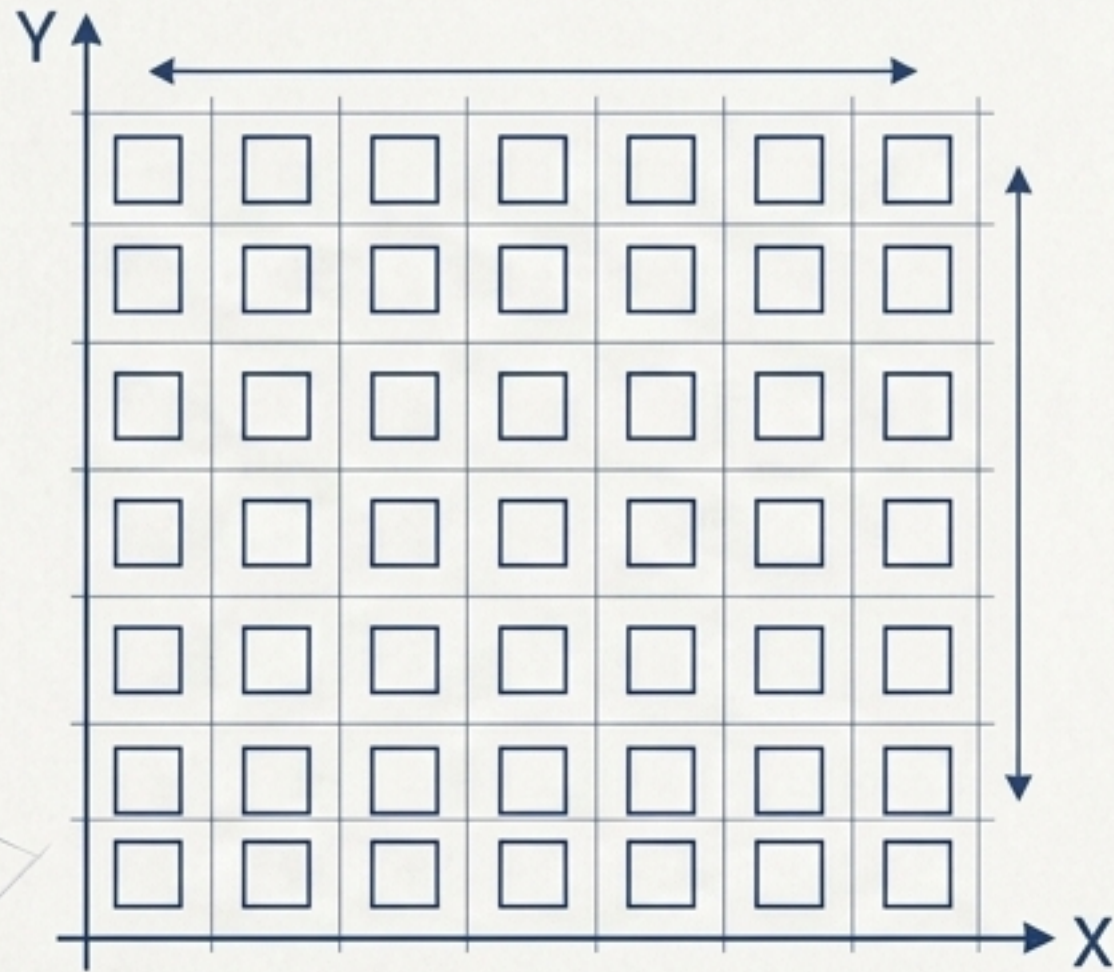
Tolerates vastly fewer writes before ultra-thin voltage windows blur past recovery.

ECC Reliance

Modern QLC is unusable without aggressive Error Correcting Codes quietly fixing bits in the background.

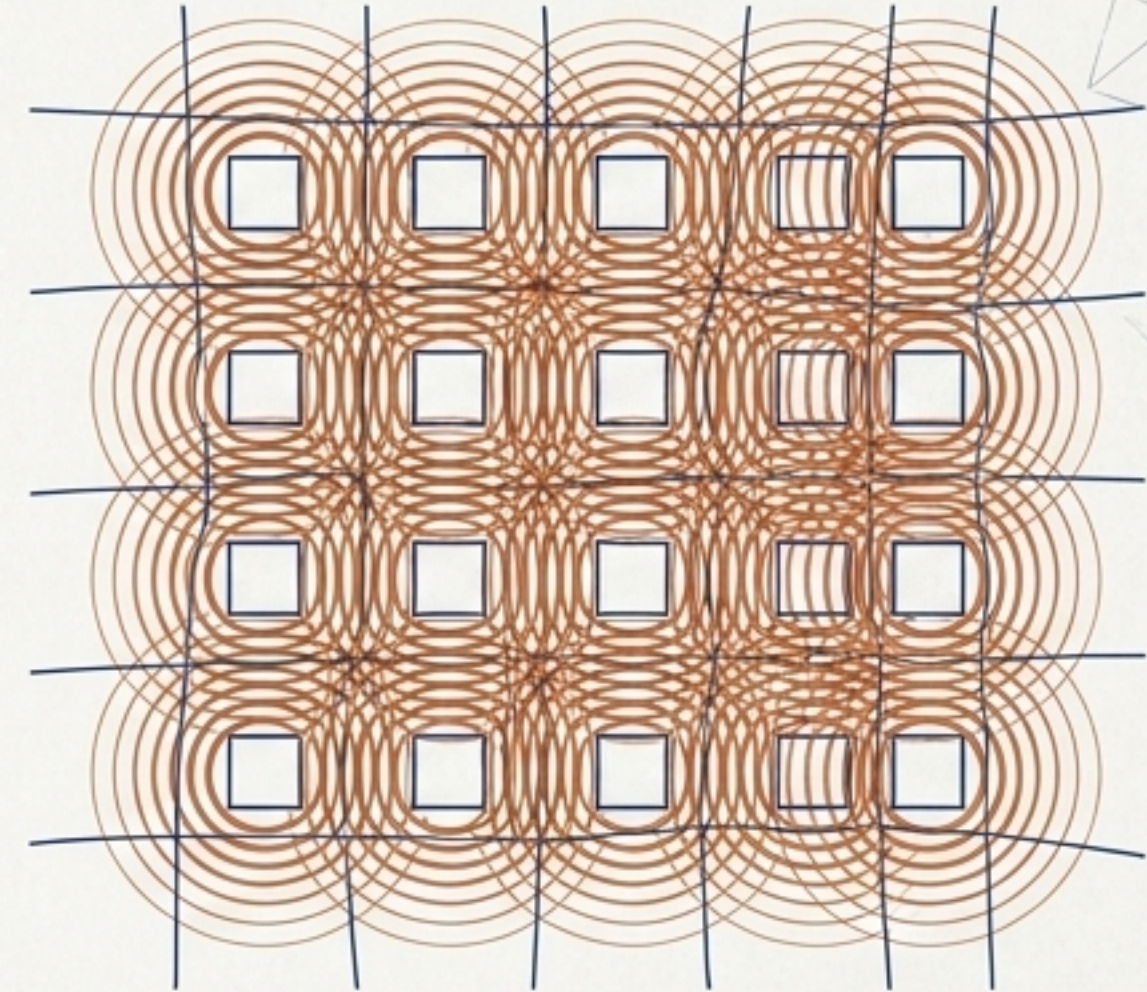
The Mid-2010s Wall: When Horizontal Scaling Broke

Pre-2015



Cells shrinking and packing neatly on a 2D plane.

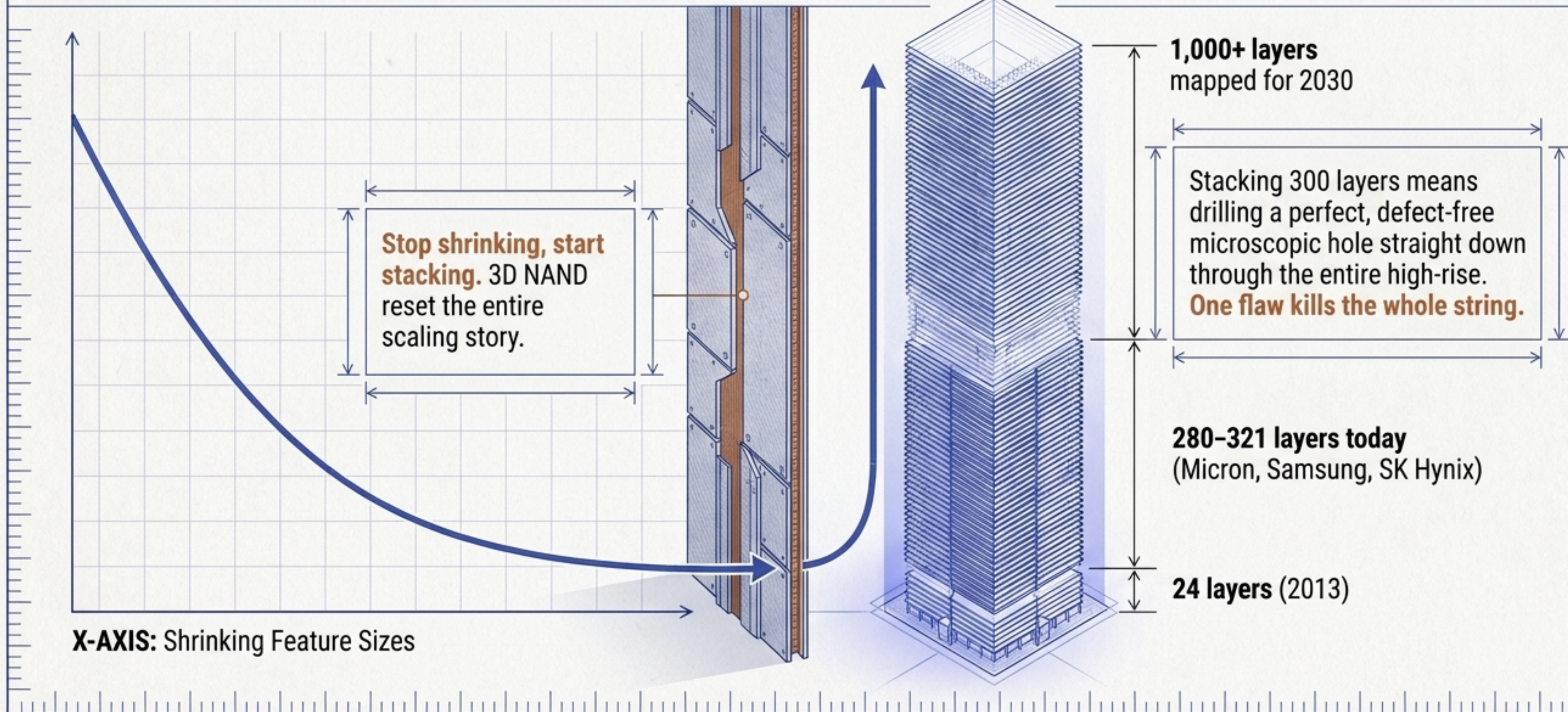
Mid-2010s Wall



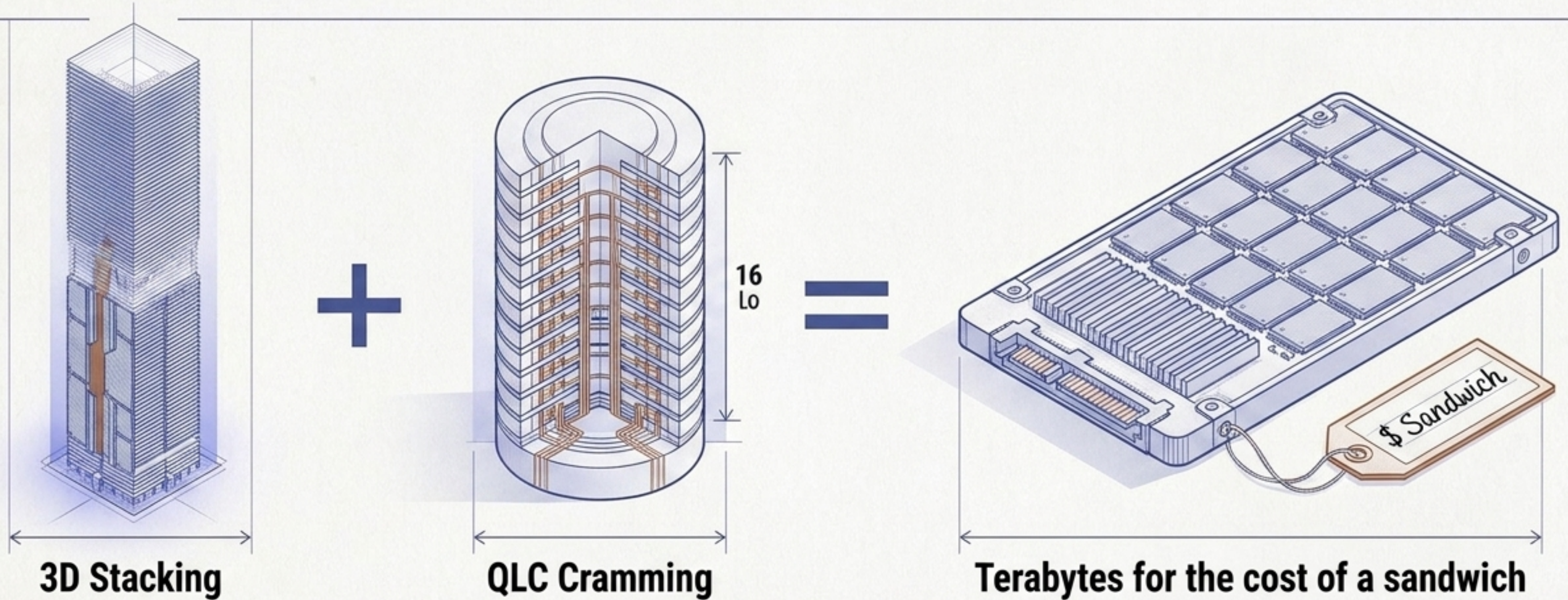
Interference. Charge on one island corrupts its neighbors.

For decades, getting cheaper meant getting smaller. But in the mid-teens of nanometers, physics intervened. Squeeze further, and planar (2D) NAND literally destroys its own data.

Moore's Law Turns 90 Degrees



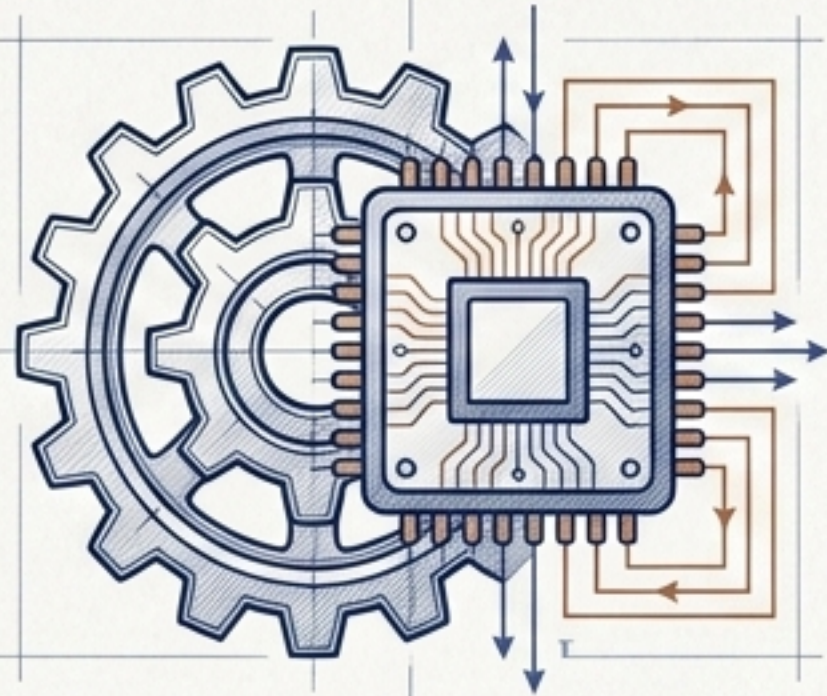
Compounding Density: The Math Behind Cheap Storage



The cost-per-bit curve didn't break; it changed axes. By compounding vertical layer stacking with aggressive bit-cramming, a modern 300-layer QLC die holds vastly more than older planar chips on the exact same patch of silicon.

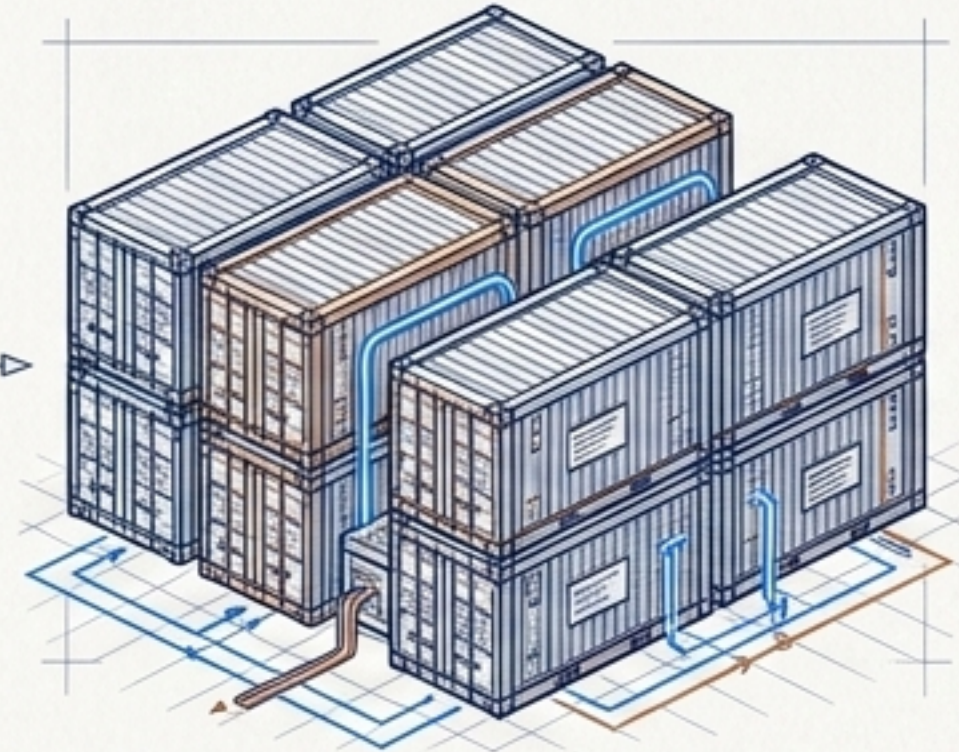
Distinguishing the Cousins: NOR vs. NAND

NOR Flash



- Random-access memory.
- Supports execute-in-place directly.
- Fast, expensive, small capacities.
- Runs the code (boot ROM, firmware).

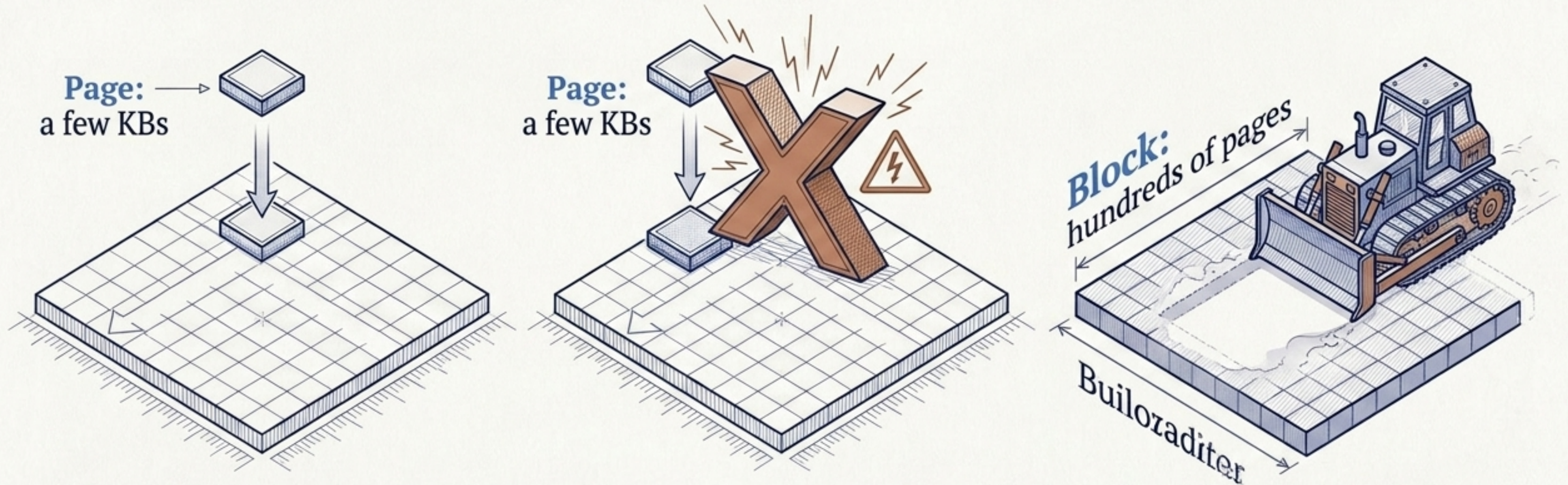
NAND Flash



- Organized in pages and massive blocks.
- Byte-by-byte reading is impossible.
- Vast density, minimal cost.
- Holds the data (mass storage).

NAND won the mass storage war
entirely on cost per gigabyte.

The Original Sin of SSDs: Pages vs. Blocks



You write in small, precise units (Pages).

You cannot overwrite a single page in place.

You can only erase in massive units (Blocks).

To change one byte, the drive must read a whole block, modify it, write it somewhere fresh, and erase the old block.

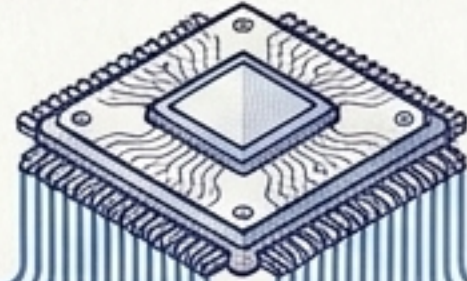
The Invisible Brain: Wear-Leveling & Amplification

Wear-Leveling

If an OS repeatedly overwrote the same physical block, QLC would die in days.

The controller artificially spreads writes evenly across all cells.

Flash Translation Layer (FTL Controller)



Write Amplification

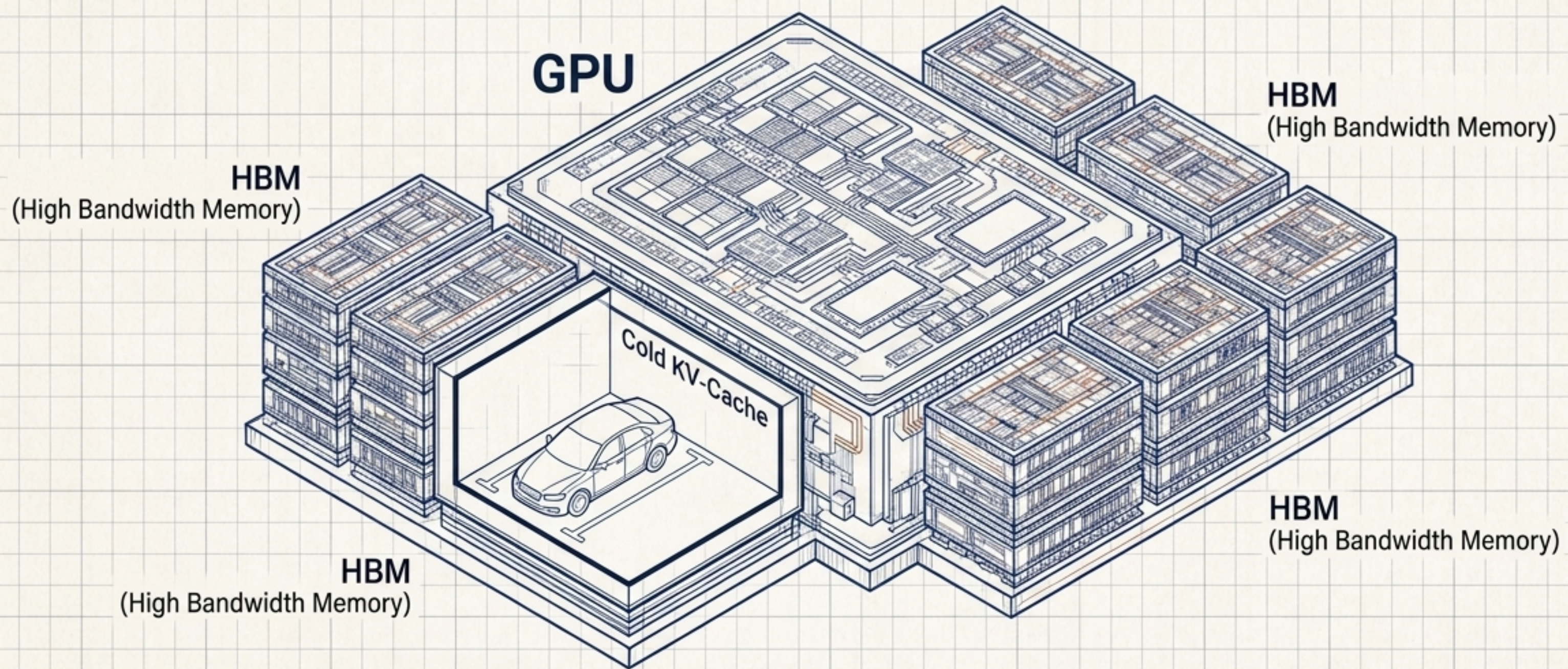
Because of the erase mismatch, one logical write triggers several physical shuffles.

A write-amplification of 3x means 1GB of OS data causes 3GB of physical flash wear.



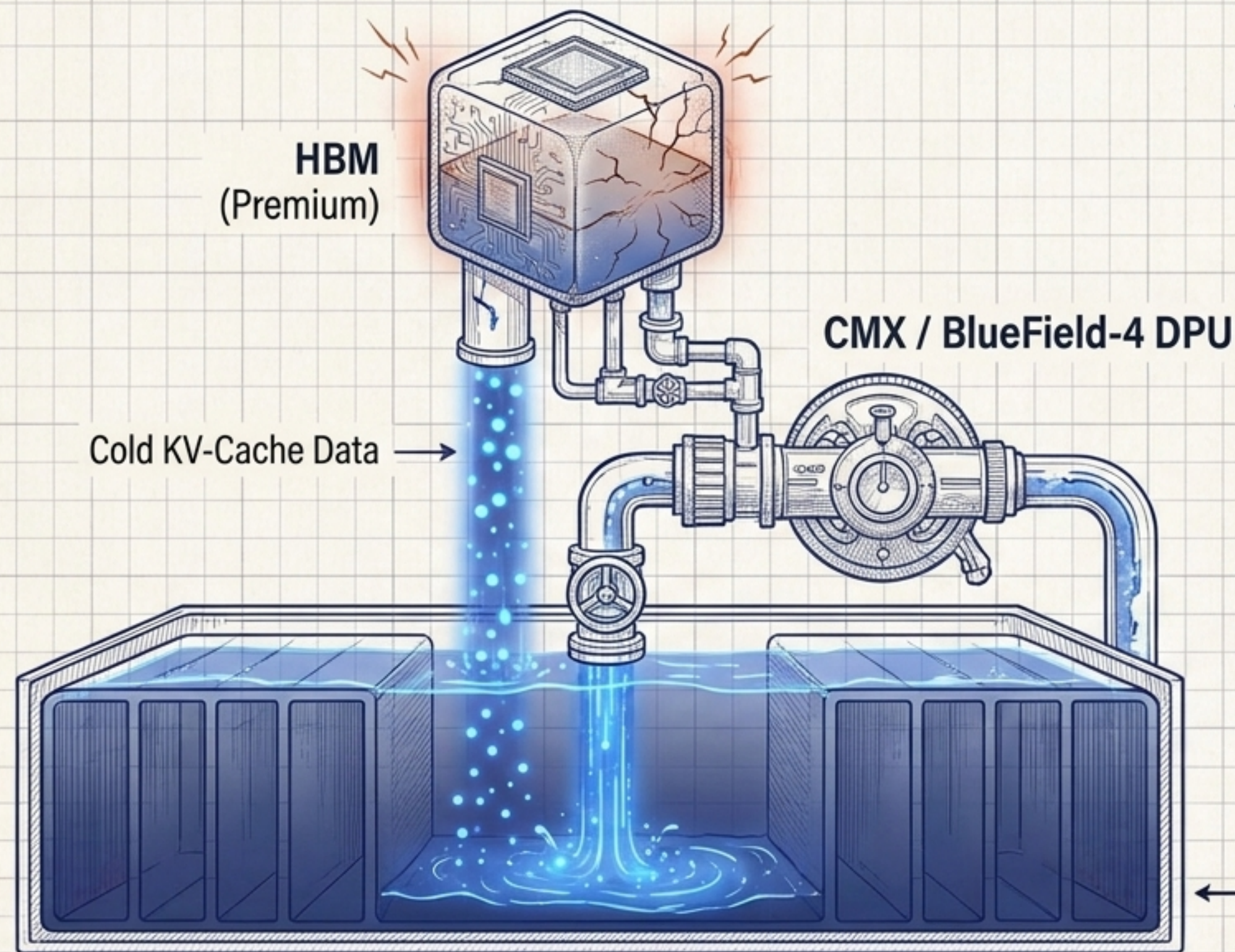
The smartest part of a modern SSD isn't the flash—it's the controller keeping the fragile flash alive.

The AI Memory Wall: Wasting the Most Expensive Real Estate



HBM is the priciest, most supply-constrained component in an AI server. In inference at scale, a massive amount of HBM is consumed by the KV cache (the model's running memory). Holding "cold" context you don't need right this second in HBM is like parking a car you won't drive until next week in the most expensive garage in the city.

The Arbitrage Opportunity: Spilling the Cache

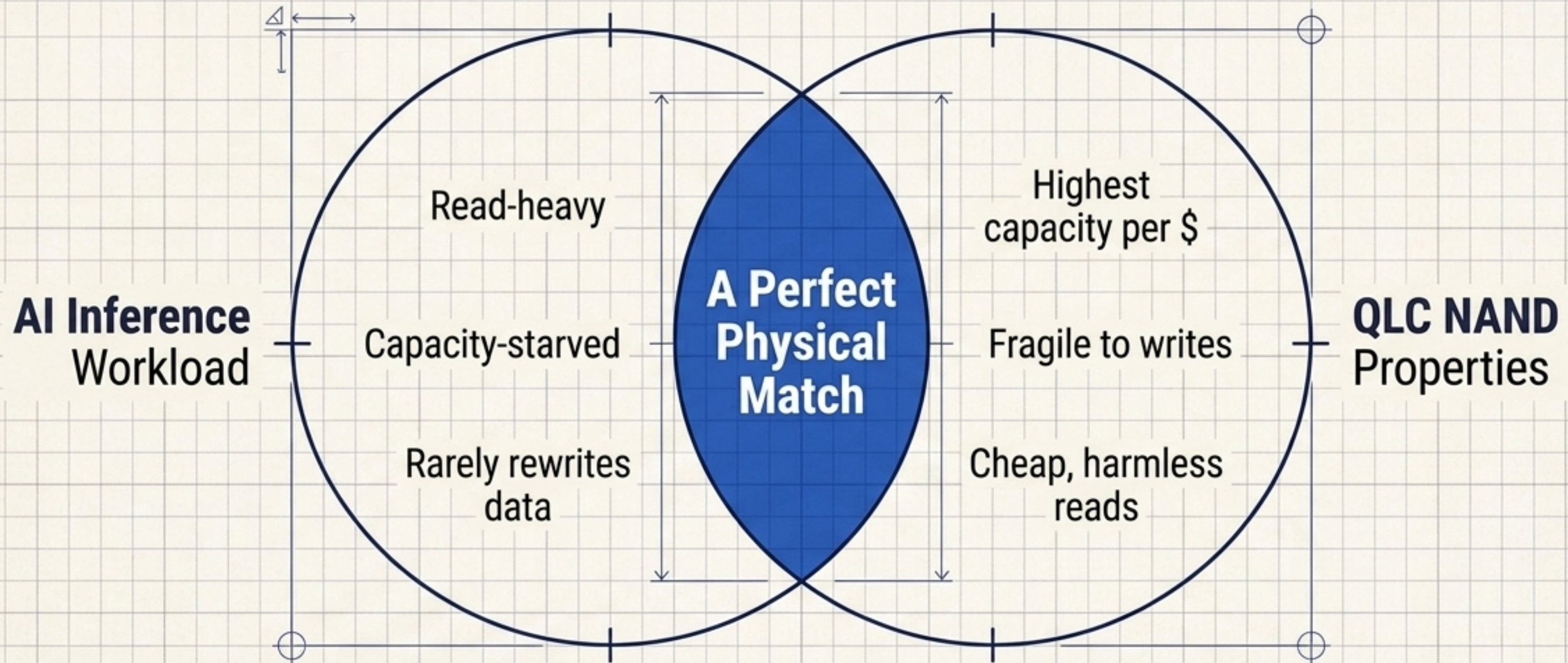


The economics are offensive, so the industry built an arbitrage.

Technologies like NVIDIA's CMX extend HBM by spilling cold KV-cache down to cheap, dense NAND, pulling it back up only when needed.

NAND just got promoted from the bottom shelf to the overflow tank for the most expensive memory in the building.

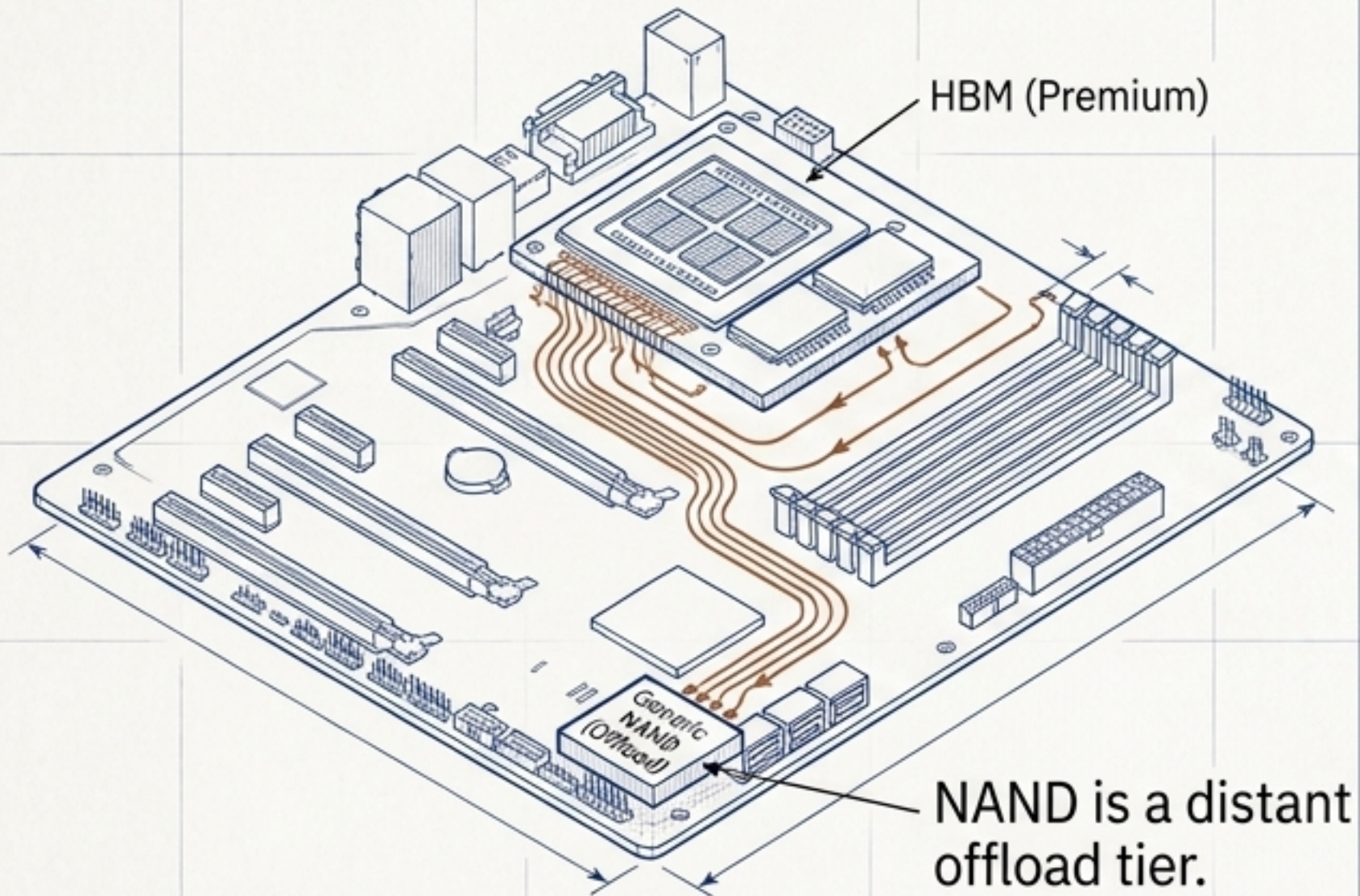
Why QLC Wins AI Inference



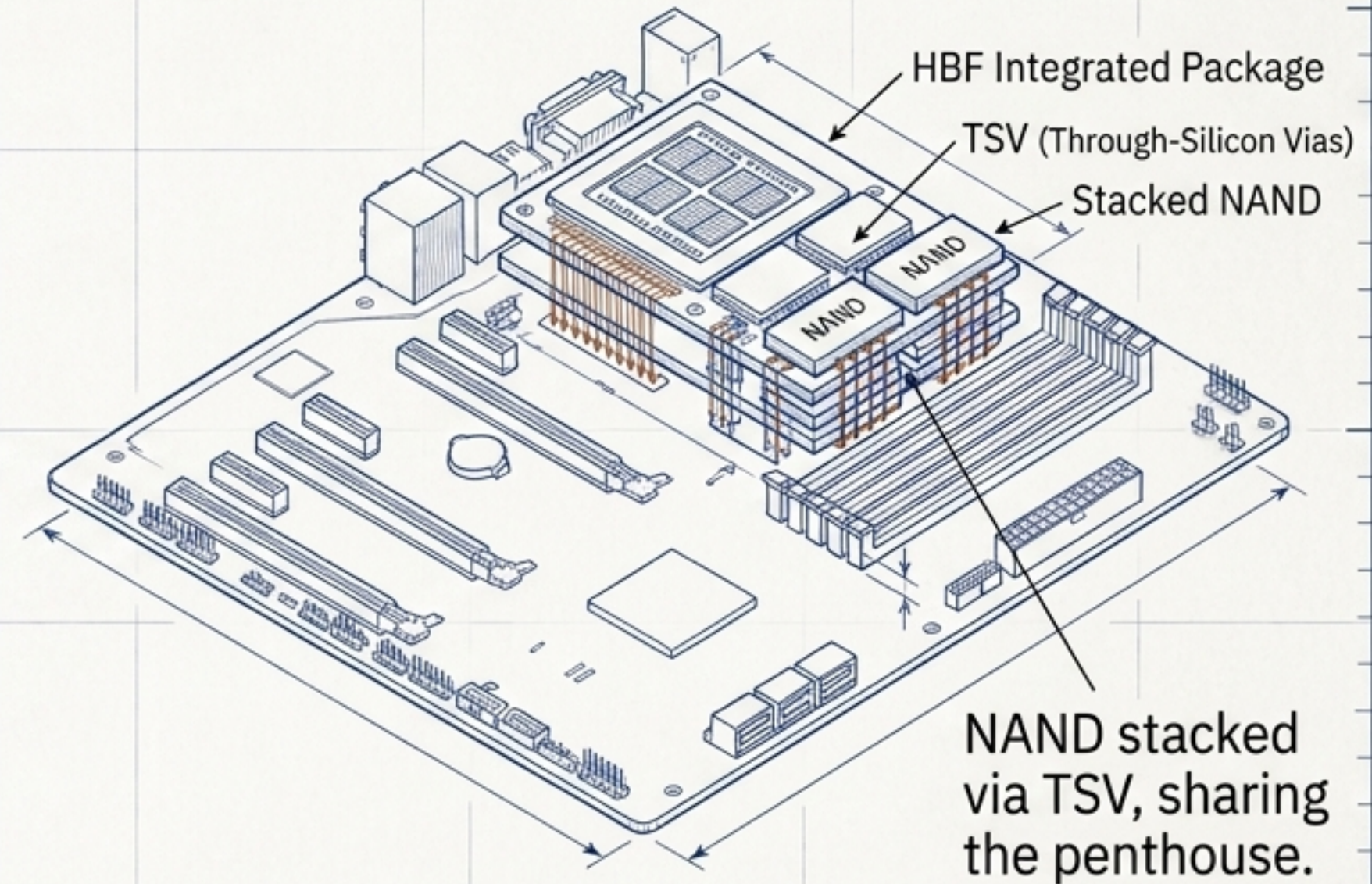
Recall the asymmetry of physics: writes kill flash, reads are harmless. AI inference is overwhelmingly read-heavy. It streams cache continuously without triggering QLC's fatal weakness (write-wear). The weakest, cheapest flash turns out to be the exact perfect match for the most demanding workload of the decade.

The Bold Future: High Bandwidth Flash (HBF)

Traditional Architecture

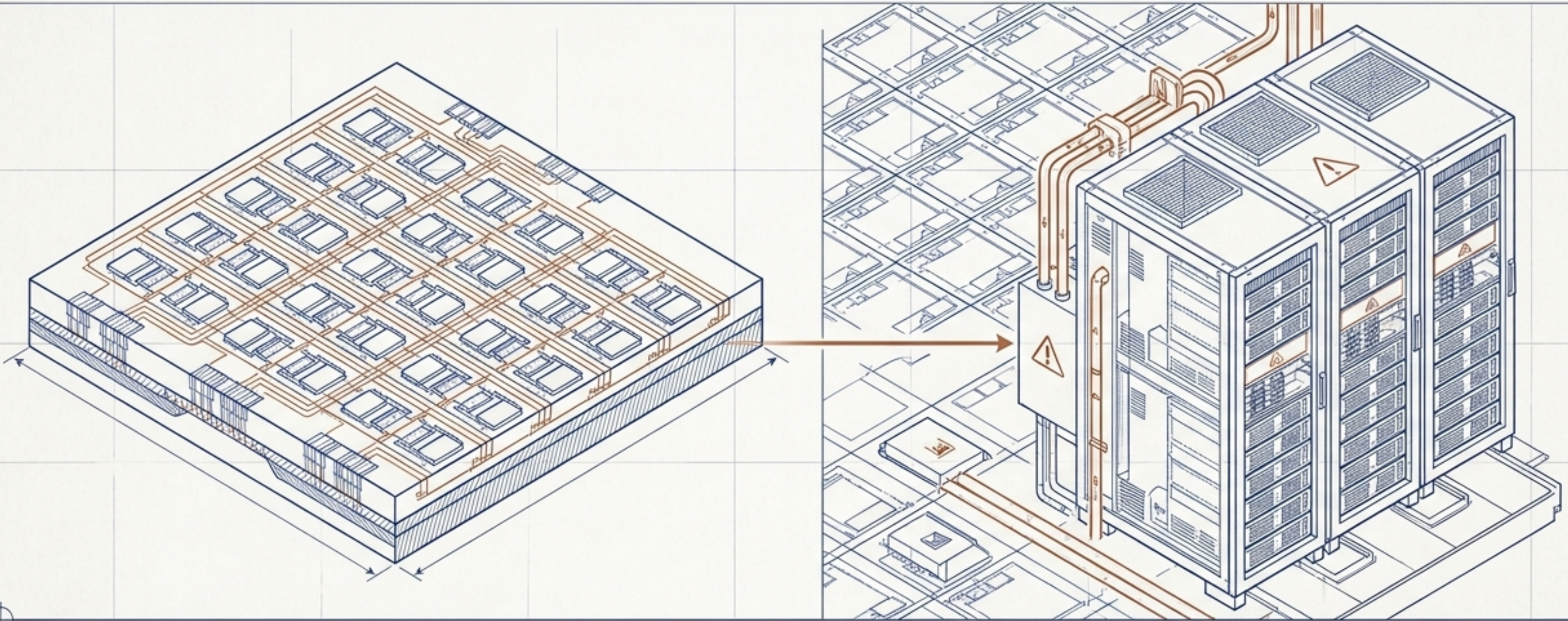


HBF Architecture (SanDisk)



The ultimate escalation. SanDisk's HBF proposes using TSV wiring to stack flash directly next to the GPU. The pitch: 8 to 16 times the capacity of HBM at comparable total system cost. NAND stops being a distant offload tier and moves into the premium real estate.

The Misunderstood Hero of the AI Stack



NAND didn't become faster. It didn't become more durable. The system simply learned to build offload paths that route read-heavy, capacity-hungry data directly to the memory built for exactly that. The chip that got cheap by trading endurance for density is winning because the new boss doesn't mind the scars.