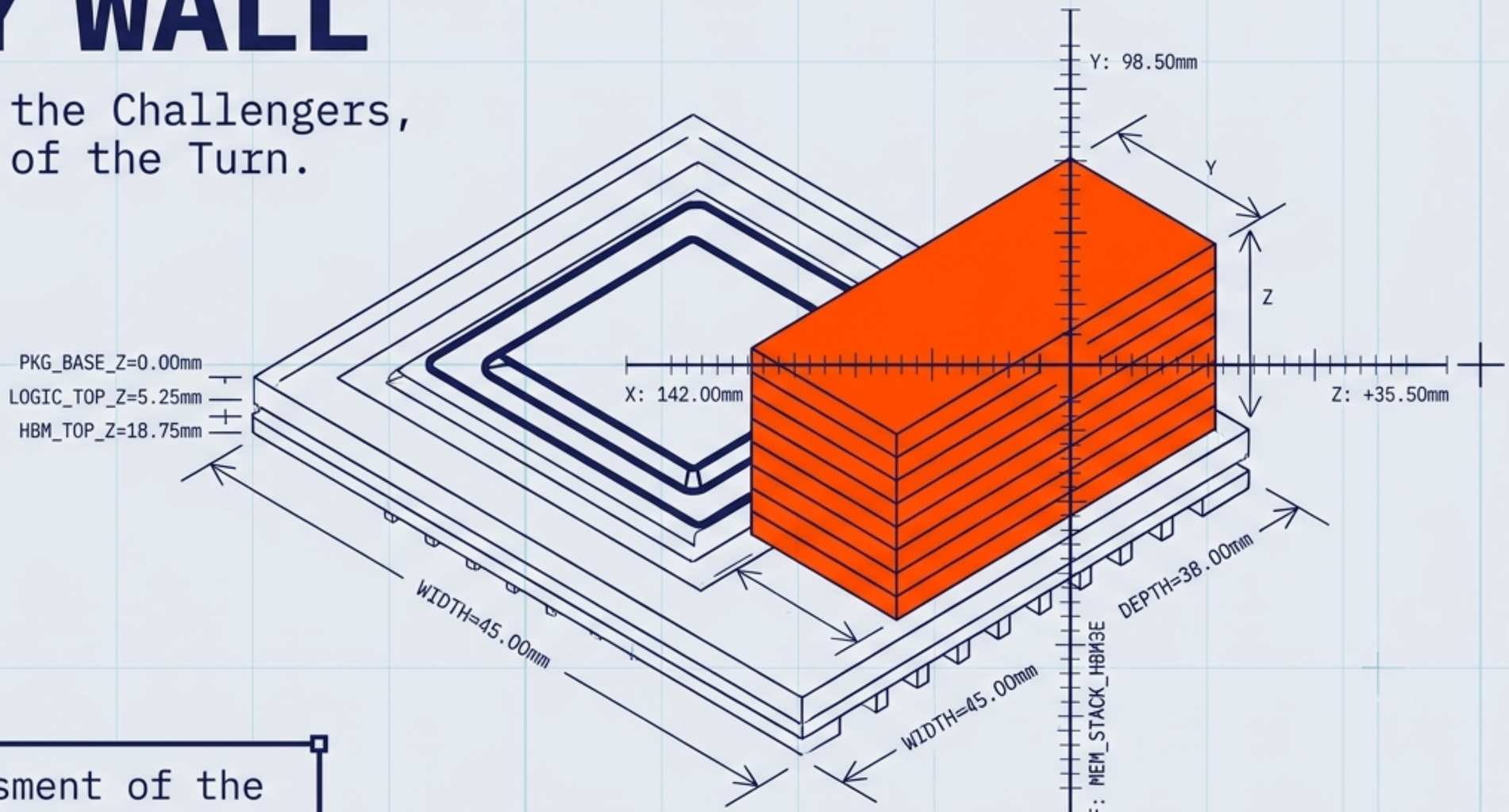
BEYOND THE MEMORY WALL

The Workarounds, the Challengers, and the Timeline of the Turn.



A Strategic Assessment of the 2026 AI Infrastructure Market.

The Anatomy of a Siege

CONSTRAINT 1: BANDWIDTH

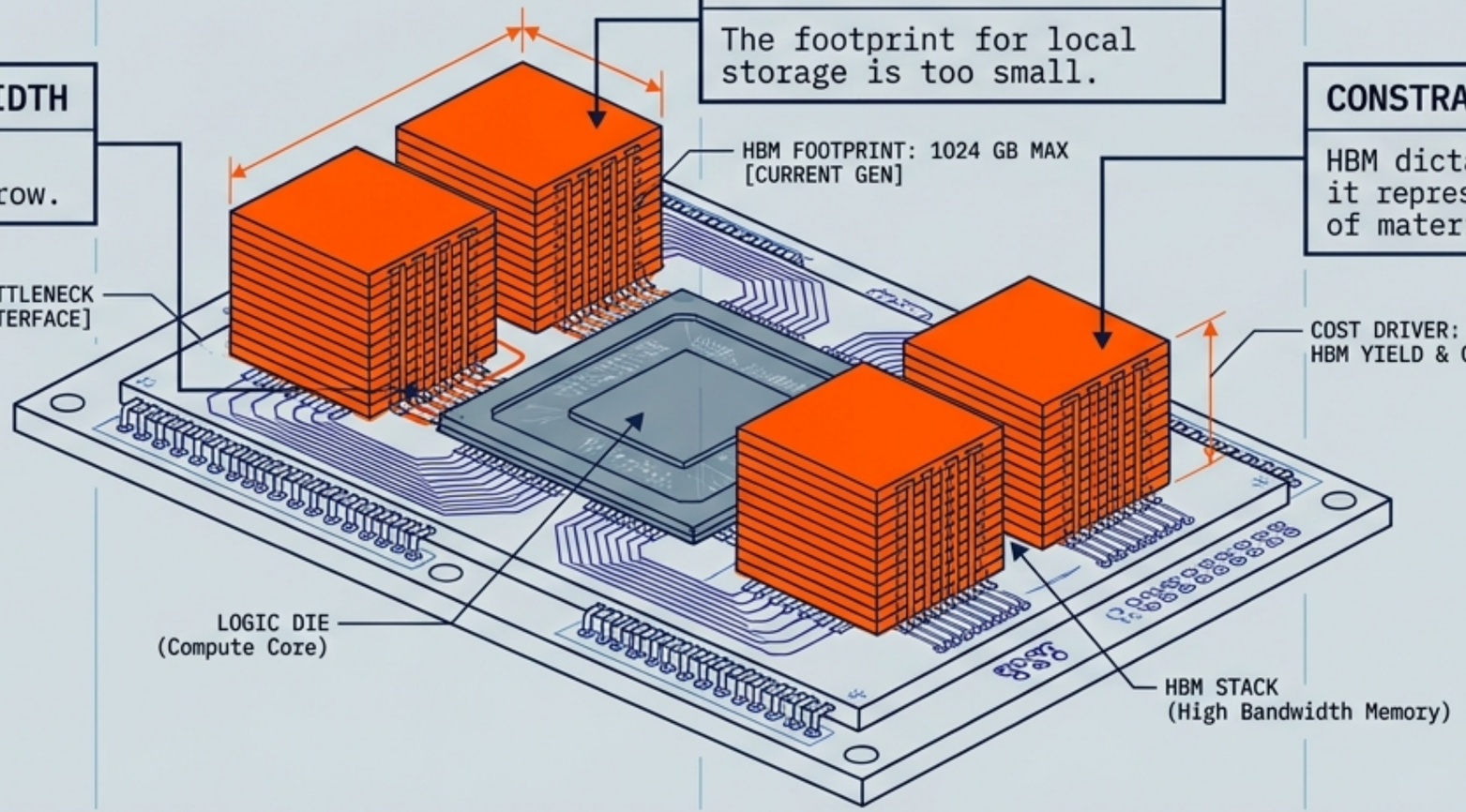
The firehose into the compute die is too narrow.

CONSTRAINT 2: CAPACITY

The footprint for local storage is too small.

CONSTRAINT 3: COST

HBM dictates the economics: it represents ~50% of the bill of materials for an AI GPU.



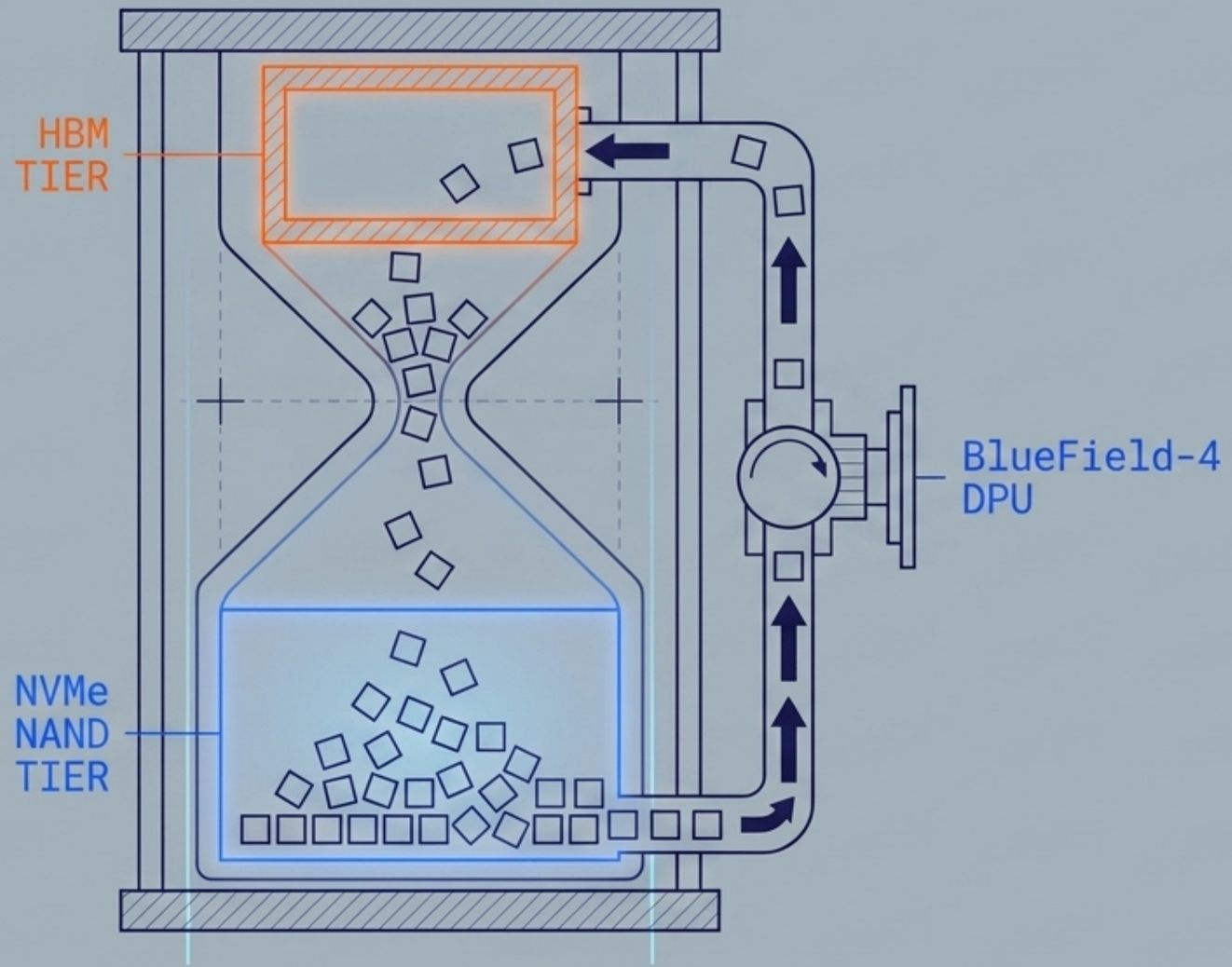
SYNTHESIS: When a single component dictates half the cost of the most important chip in the world, the industry attacks it from every side.

THE ARSENAL: FIVE VECTORS OF ATTACK

VECTOR	PRIMARY ARCHITECT	CONSTRAINT ATTACKED	STRATEGIC APPROACH	2026 EFFICACY
CMX	NVIDIA	Capacity / Cost	**EXTEND** HBM by tiering to NAND	High (Deploying)
CXL	Cloud Hyperscalers	Capacity	**POOL** memory over PCIe fabric	High for capacity / Low for latency
WSE-3	Cerebras	Bandwidth	**DELETE** HBM entirely for SRAM	Niche (Model size constrained)
PIM	Samsung / SK Hynix	Bandwidth	**INTEGRATE** compute into memory	None (Prototype phase)
HBF	SanDisk / SK Hynix	Cost / Capacity	**STACK** NAND for read-only weights	Low (Sampling H2 '26)

VECTOR 1: CMX AND THE TELL

KEY INSIGHT: NVIDIA is not trying to sell you less HBM. They are trying to make each gigabyte carry a larger workload by promoting NAND to an active participant.

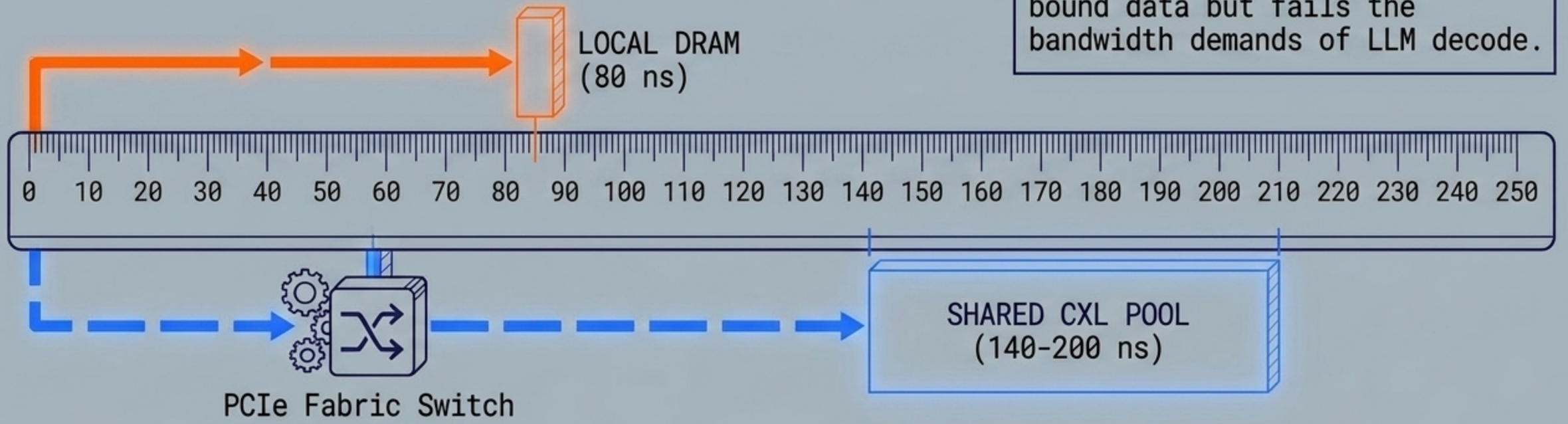


1	MECHANISM
	Spills cold KV-cache out of HBM down to flash. Prefetches ahead of decode.

2	PERFORMANCE
	Yields 5x tokens per second. Time-to-first-token on long contexts drops from 65s to 3s (VAST Data).

VECTOR 2: THE CXL LATENCY TAX

KEY INSIGHT: CXL widens the back of the pyramid; it doesn't move the top. It works for capacity-bound data but fails the bandwidth demands of LLM decode.



1	THE SPECS
	CXL 4.0 specs: 128 GT/s on PCIe 7.0, bundling toward 1.5 TB/s.

2	THE VERDICT
	Google published "A Case Against CXL Memory Pooling" — when the hyperscaler with the most to gain argues against it, the market must update.

VECTOR 3: DELETING THE WALL ENTIRELY

→ 46,225 mm² of silicon.

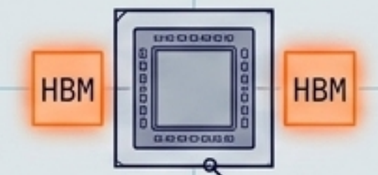
→ 44 GB of SRAM.

→ Zero off-chip memory hops.

→ Yields ~21 PB/s bandwidth (7,000x an H100's HBM).

→ Runs Llama 4 Maverick at ~2,500 tokens/sec per user.

THE PACKAGE

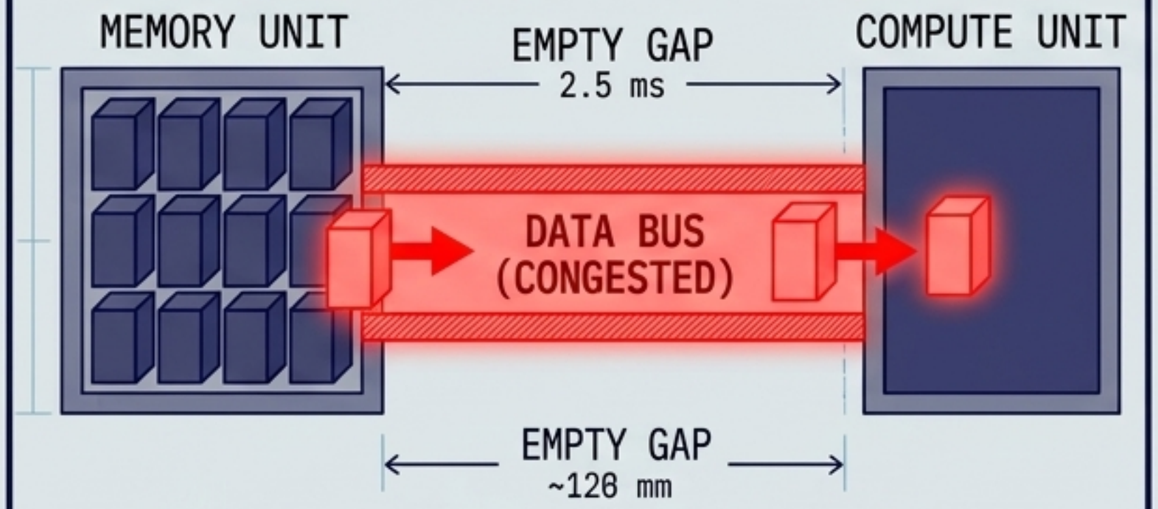


→ Holds ~140 GB of HBM capacity.

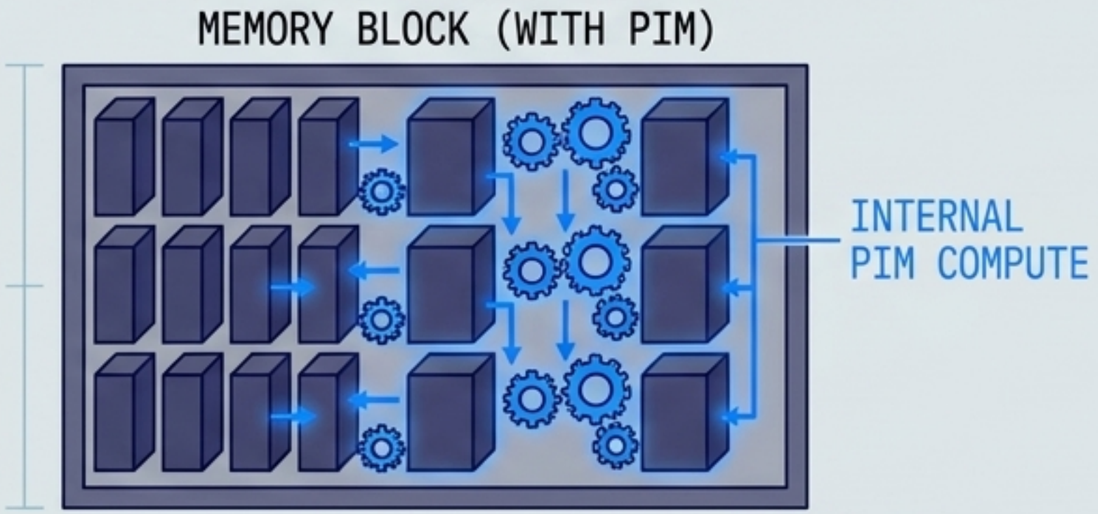
THE TRADE-OFF: Deleting HBM doesn't beat the wall. It trades a bandwidth problem for a capacity problem. 44 GB of wafer-scale SRAM cannot hold a 70B parameter model in FP16—that requires lashing four systems together.

VECTOR 4: THE PIM LONG FUSE

TRADITIONAL



PIM ARCHITECTURE



1	**MECHANISM**	2	**THE PAYOFF**	3	**THE REALITY CHECK**
	Stop moving data to the compute. Move the compute to the data (e.g., Samsung HBM-PIM, SK Hynix AiM).		Dozens of times more efficient on memory-bound GEMV operations.		As of early 2026, it remains prototype-and-patent. LPDDR6-PIM standardization is just starting.

KEY INSIGHT: PIM is the answer if the wall is permanent. It is **not* the answer to the wall in front of us this year.

VECTOR 5: REBUILDING THE WALL WITH FLASH

THE LOGIC

Inference is heavily read-biased (weights, cold context). Why pay the DRAM premium for fast, infinite rewrites when static weights don’t need them?

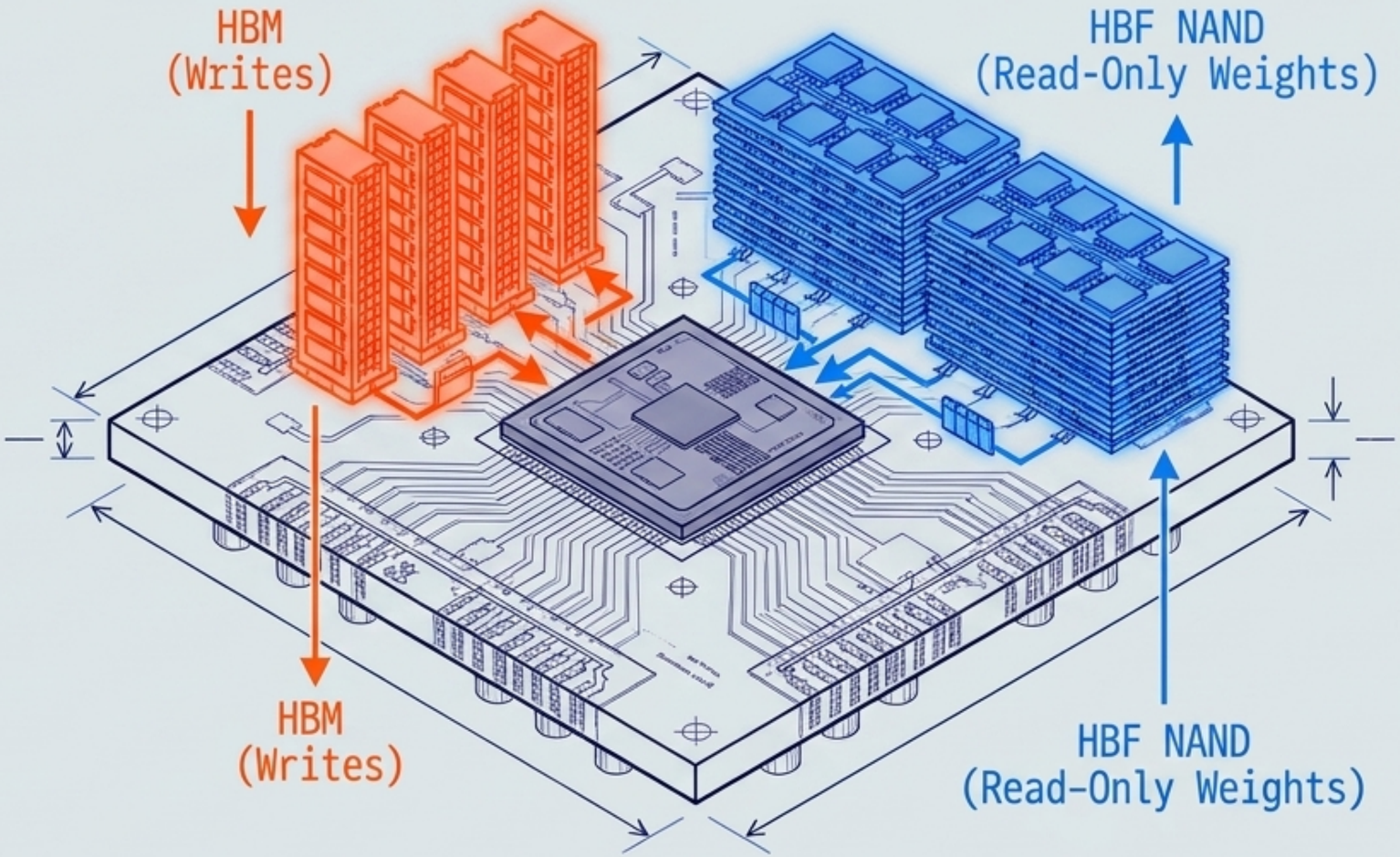
THE SPECS

SanDisk’s HBF stacks 16-die NAND via TSVs. Yields 8–16x the capacity of HBM4 (~512 GB vs 48 GB) at comparable system cost.

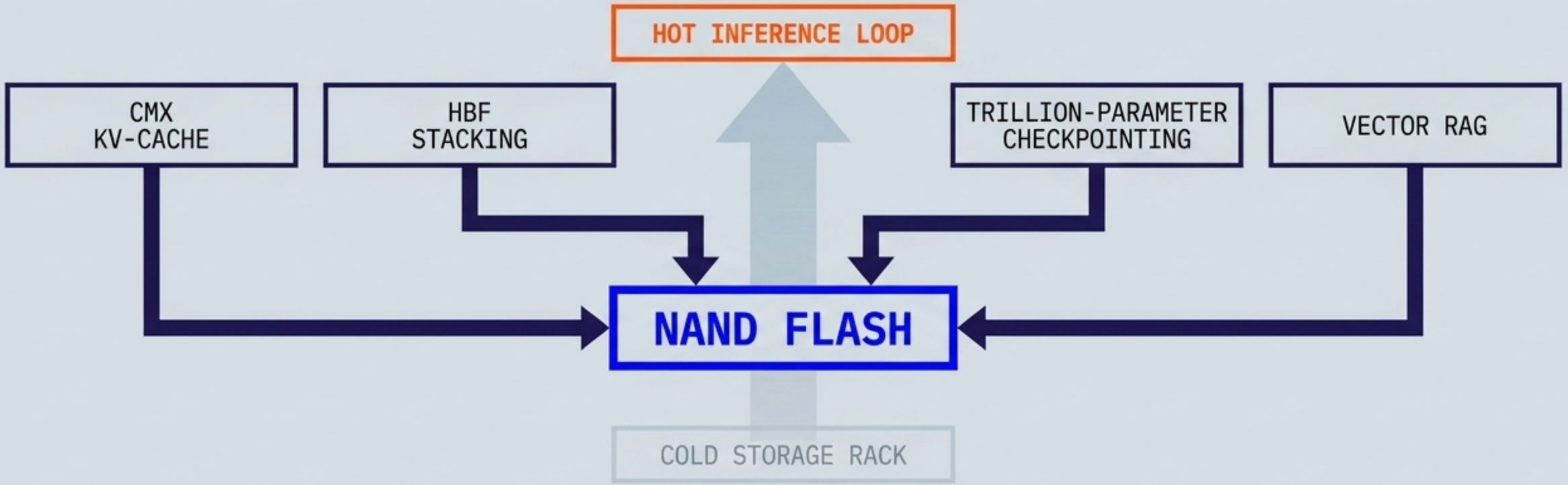
TIMELINE LIMIT

Pre-production samples H2 2026. Devices early 2027.

THE HYBRID INTERPOSER MAP



THE UNINTENDED CONSEQUENCE: NAND'S PROMOTION



SYNTHESIS INSIGHT

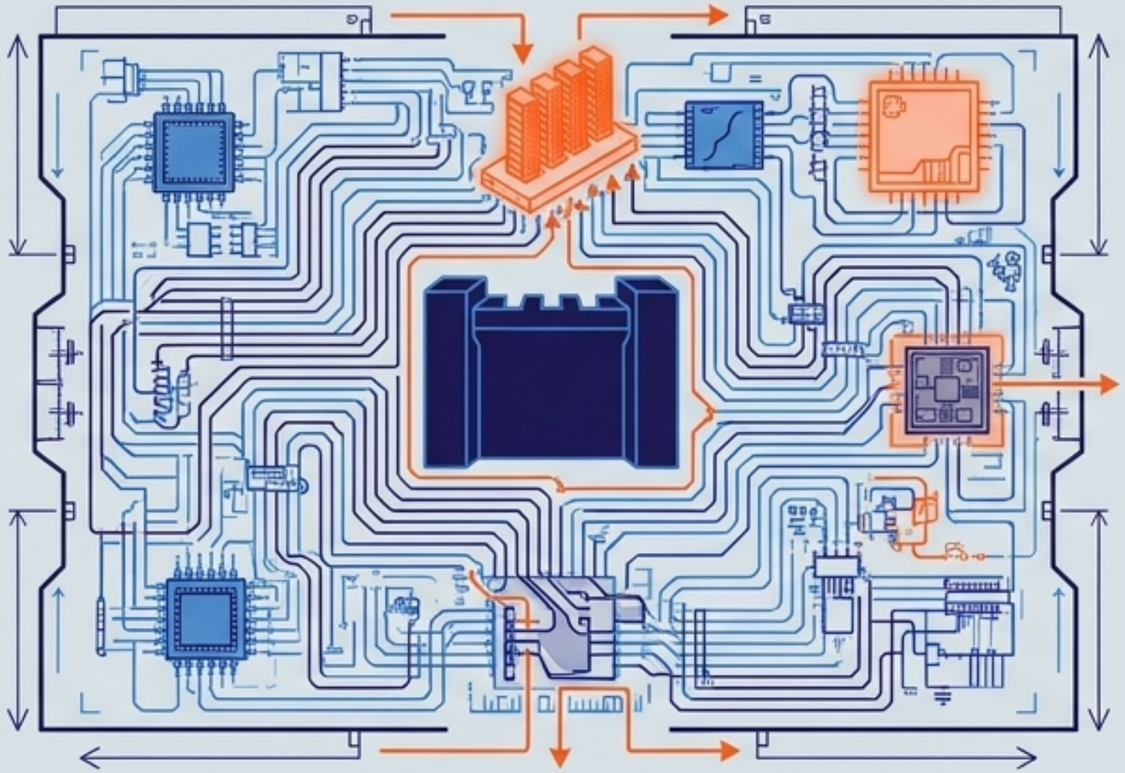
Every attempt to route around the HBM wall ends with NAND doing inference work it was never built for.
The cold storage chip has been dragged into the hot loop.

THE RECEIPTS:

- Q1 2026 Top-5 NAND revenue: +83.7% QoQ to $38.9 billion.
- 2026 total NAND market forecast: $270.6 billion (+280% YoY).
- 30TB Enterprise SSD price: ~$3,062 (April 2025) → ~$9,318-$11,000 (Q1 2026).

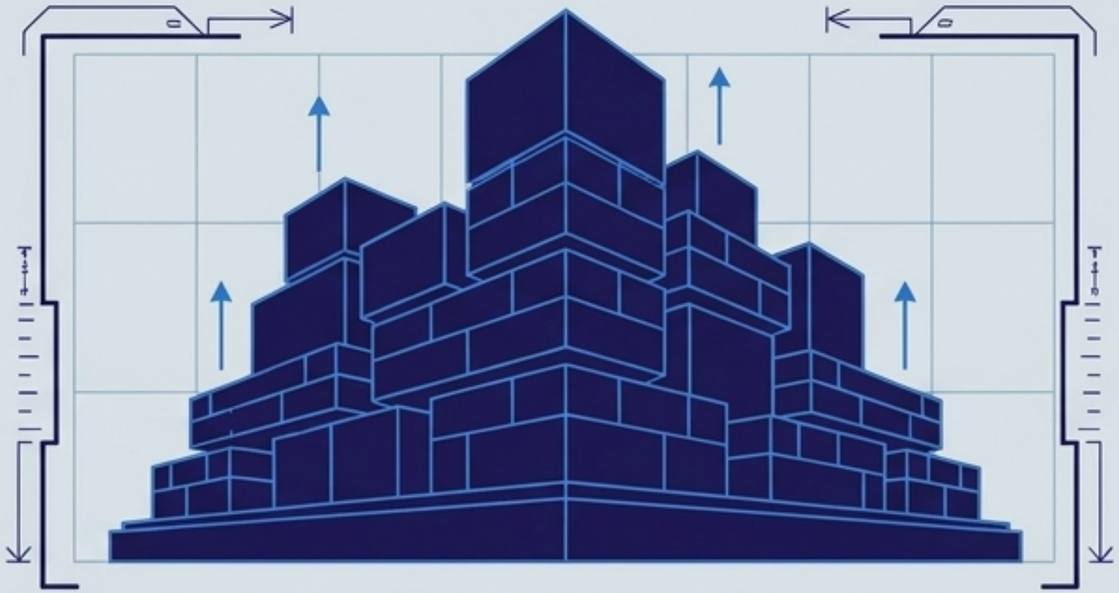
THE TWO FRONTS: ARCHITECTURAL VS. STRUCTURAL

THE WEST: ROUTING AROUND THE TOP



SUBTEXT: Relies on HBM dominance, advanced packaging (CoWoS), and highly complex architectural tiering.

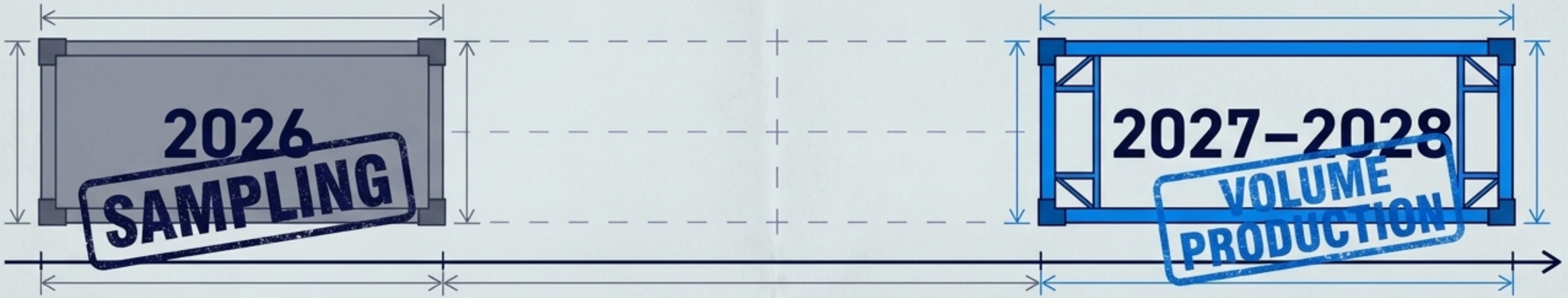
CHINA: BUILDING FROM THE BOTTOM



SUBTEXT: Scaling under sanctions by dominating the high-volume foundation layers.

- YMTC (NAND Front): Standing on the wall. 11.8–13% global market share on 294-layer process. Targeting $30B+ IPO (May 2026).
- CXMT (DRAM Front): Climbing fast. Q1 2026 revenue up 719% YoY. Shipping DDR5-8000 and LPDDR5X-10667.

THE CHINESE HBM REALITY CHECK



THE CONTEXT

CXMT has sent HBM3 samples to Huawei, heavily fueling market narratives of an imminent supply glut.

THE REALITY

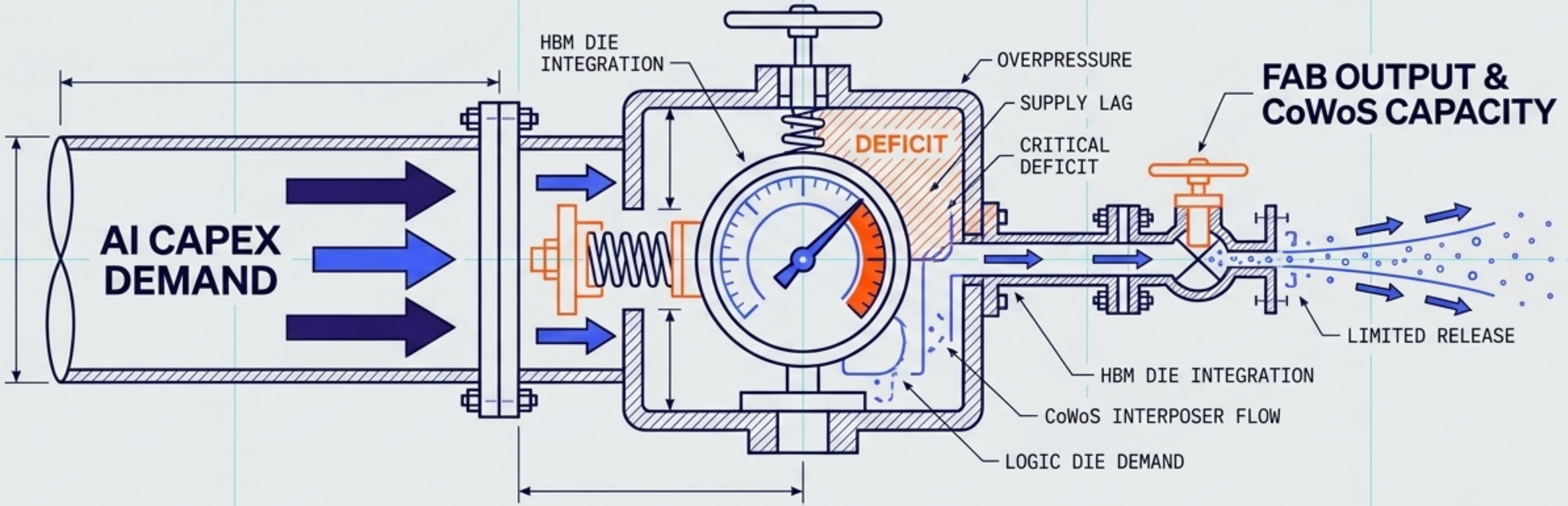
HBM volume production is a multi-year ramp. Scaling complex 3D packaging under tooling sanctions guarantees severe time lags.

STRATEGIC TAKEAWAY

A real, sanctions-proof NAND business exists *today*, but Chinese HBM is directionally inevitable rather than immediately disruptive.

Anyone claiming Chinese HBM relieves the 2026 shortage is selling a 2028 fact dated to this year.

THE MECHANICS OF THE CYCLE TURN



THE RULE

What ends a memory cycle is never demand softening—AI demand is the most non-discretionary capex on earth right now.

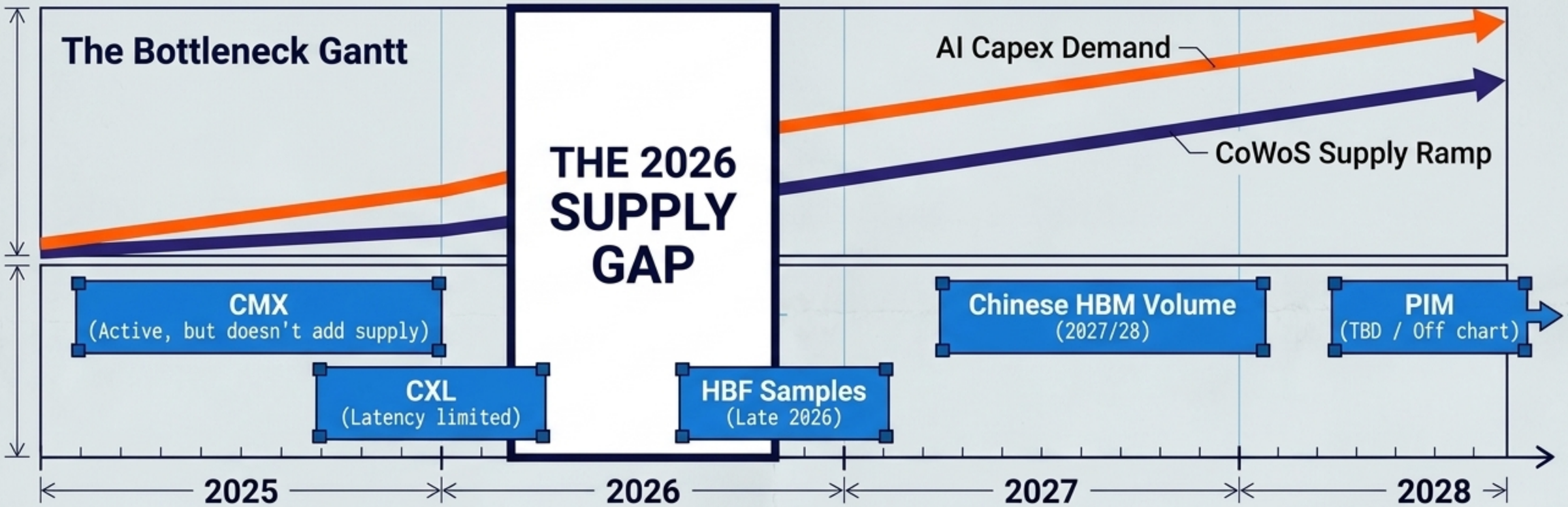
THE CATALYST

Cycles break (as in 2019 and 2022) only when supply catches up while demand digests.

THE FORMULA

[New HBM Bit-Supply] + [New CoWoS Capacity] > [Hyperscaler Absorption]

THE ANATOMY OF THE TURN



SYNTHESIS:

None of the workarounds add meaningful HBM-grade supply in 2026. They reduce the **need* for HBM at the margin, but they do not flood the market.

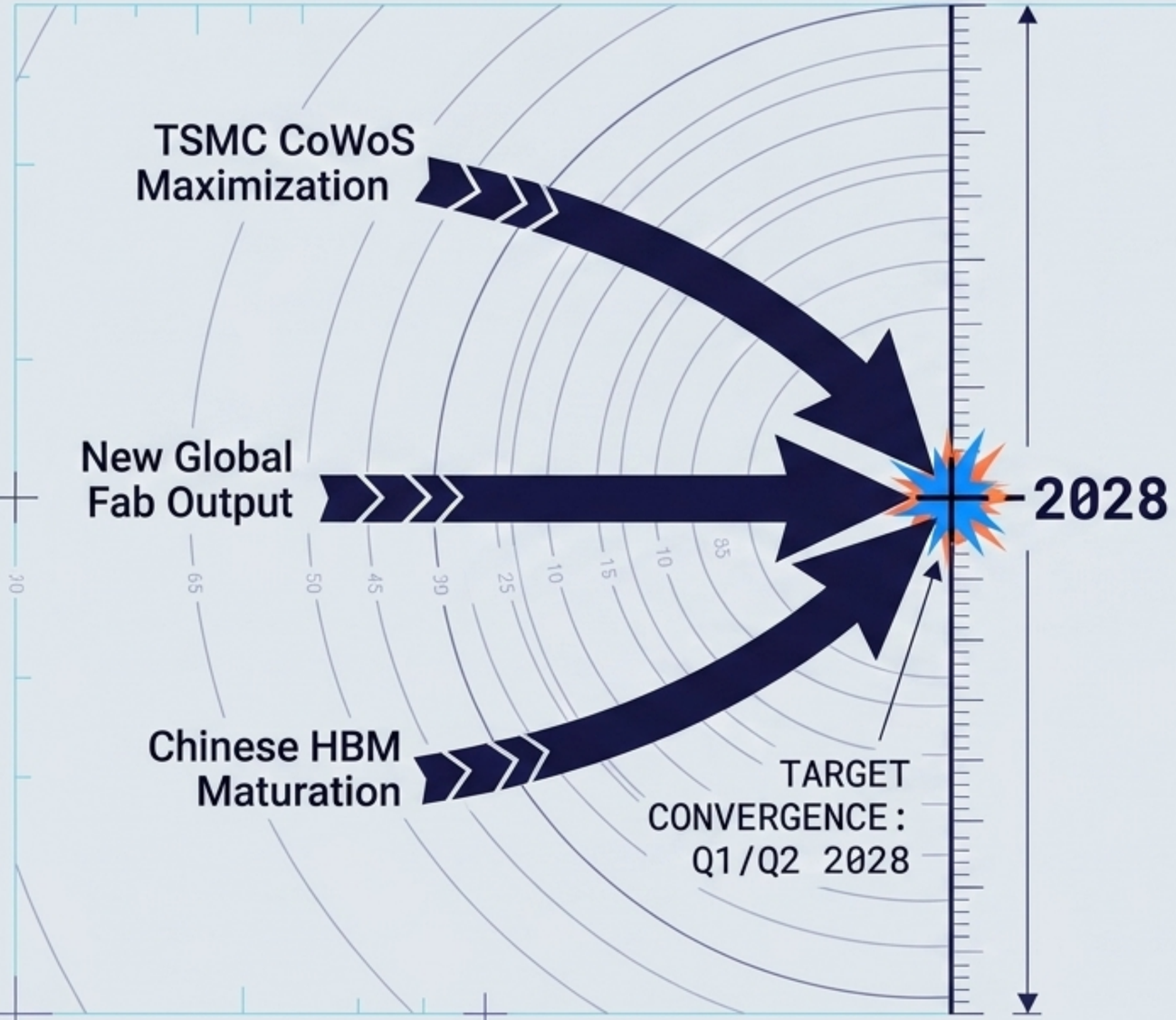
TSMC CoWoS REALITY

Ramping from 75k to 120k+ wafers/month by end of 2026, but remains completely sold out against the demand curve.

THE CONCLUSION

Because supply won't break the wall in 2026, a market turn this year can ***only*** come from demand.

THE TRUE GLUT RISK IS A 2028 EVENT



MARKET TRAJECTORY

The rate of acceleration cools in late 2026. Absolute prices remain elevated through 2027.

ANALYST CONSENSUS

IDC, Counterpoint (Q4 '27 inflection), DBS, and Intel's Lip-Bu Tan completely align: no return to baseline pricing until 2028.

THE WATCHLIST

The crack will not show up in memory price prints. It will show up in hyperscaler capex guidance. As long as capex rises and CoWoSS stays sold out, the cycle runs.

THE STRATEGIC BASELINE

TAKEAWAY I

The binding constraint inside compute has permanently moved. For two decades it was FLOPS. Today, it is the bandwidth and capacity of the memory sitting next to the logic.

TAKEAWAY II

Every workaround detailed in this briefing represents capital that saw the wall clearly and moved early. That is the only edge there is.

The Memory Wall doesn’t reward the people who climb it.
It rewards the people who price it first.