

Deleting the Memory Wall Entirely Trades Bandwidth for Capacity

Wafer-Scale Compute

46,225 mm² of silicon (dinner-plate size) housing 4 trillion transistors.

Zero Off-Chip Hops

44 GB of SRAM sits directly on the wafer. HBM is deleted entirely.

Devastating Speed

~21 PB/s bandwidth. Roughly 7,000x the bandwidth of an H100's HBM. Lightning-fast single-stream decode.

The Capacity Problem

44 GB cannot hold a 70B parameter model in FP16 (requires ~140 GB). Escaping the bandwidth bottleneck creates a massive capacity wall.



Promoting NAND to Active Memory Frees Up Expensive HBM Real Estate



KV-Cache Bloat

Long AI contexts generate massive cache footprints. Dormant data monopolizes expensive HBM space while waiting to be reused.

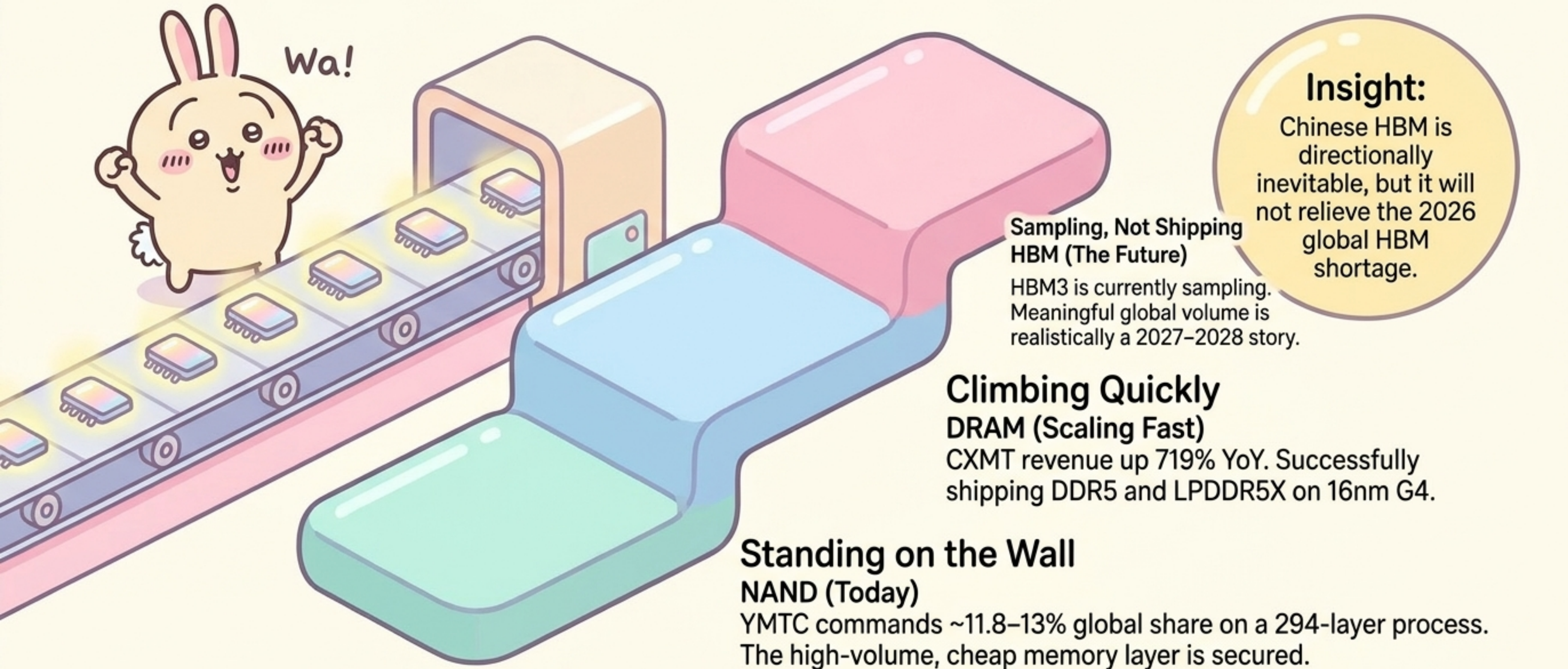
Active NAND Tiering (CMX)

A DPU keeps hot cache in HBM and spills cold cache down to NVMe flash, prefetching it back exactly when needed.

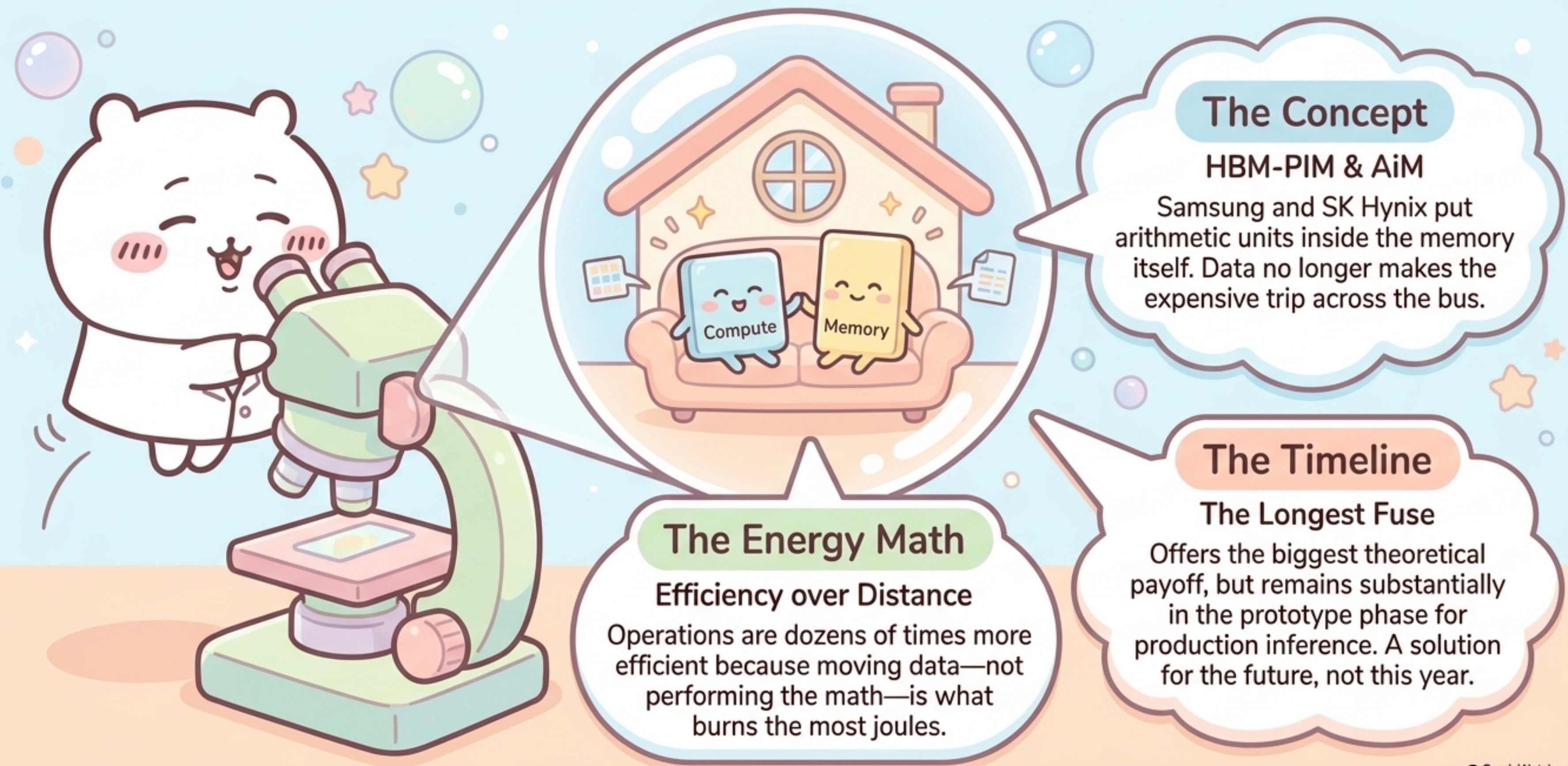
The Impact

Time-to-first-token plummets from 65s down to 3s.
Yields 5x better power efficiency. NAND is no longer just cold storage.

Chinese Domestic Supply Cures the NAND Deficit Today, but HBM Relief Waits Until 2028



Moving Compute to the Data Offers Massive Efficiency but the Longest Wait



The Concept

HBM-PIM & AiM

Samsung and SK Hynix put arithmetic units inside the memory itself. Data no longer makes the expensive trip across the bus.

The Energy Math

Efficiency over Distance

Operations are dozens of times more efficient because moving data—not performing the math—is what burns the most joules.

The Timeline

The Longest Fuse

Offers the biggest theoretical payoff, but remains substantially in the prototype phase for production inference. A solution for the future, not this year.

Hybrid Interposers Bring Vast Capacity to the Hot Loop Until the 2028 Cycle Turn

Why HBM? (DRAM)

- **Fast, infinite rewrites.** Perfect for active per-token KV-cache.
- **Trade-off:** High cost, lower capacity.

The Ultimate Hybrid

SK Hynix + SanDisk MOU:
Fast HBM for writes + vast
HBF for reads, sharing the
exact same interposer package.

Why HBF? (Flash)

- SanDisk stacks NAND like HBM. Delivers 8–16x the capacity of HBM at comparable system cost.
- Perfect for read-heavy, static model weights.



The Cycle Turn Conclusion

AI demand is relentless and TSMC CoWoS packaging remains sold out. The memory wall holds, and prices stay elevated, until new fab supply catches up to demand in 2028.