Compressed Enabled Managed MRAM Memory Technology

Nilesh Shah
VP Business Development
ZeroPoint Technologies

Executive Summary

- **Problem #1:** LLMs memory bound. HBM-based GPUs <60% utilization.
- **Problem #2:** Existing lossy compression methods degrade accuracy
- **Opportunity:** LLM inference is memory-bound 6:1 read-to-write ratio
- **Solution:** Lossless compression-enabled MRAM memory chiplet subsystem on UCIe interface.
- **Impact:** MRAM delivers HBM-like bandwidth at 30-50% lower power, AI-specific chiplet augmentation to GPU-based inference
- **Impact:** Lossless compression squeezes models by 1.5X, leading to bandwidth, power and Tokens/s gain

Compression enabled MRAM Chiplet deliver “Better together” AI inference memory solution

LLM model layers: Memory Bound

LLAMA 2.0 7B example

2 stages : Prefill and Decode

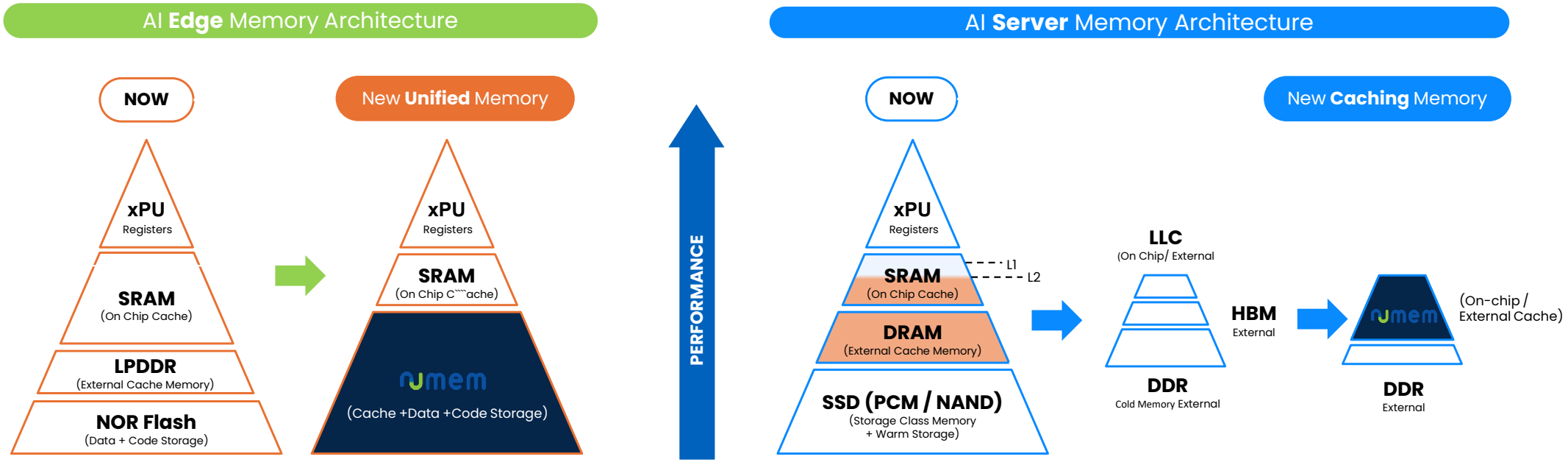
- Prefill is Compute bound
- Decode is Memory Bound
- Decode time dominates Prefill

LLM inference MEMORY BOUND

or layers in Llama-2-7b using the Roofline model of Nvidia A6000 GPU. In this example, the sequence length is 2048
e is 1.

Layer Name	OPs	Memory Access	Arithmetic Intensity	Max Performance	Bound
Prefill					
q_proj	69G	67M	1024	155T	compute
k_proj	69G	67M	1024	155T	compute
v_proj	69G	67M	1024	155T	compute
o_proj	69G	67M	1024	155T	compute
gate_proj	185G	152M	1215	155T	compute
up_proj	185G	152M	1215	155T	compute
down_proj	185G	152M	1215	155T	compute
qk_matmul	34G	302M	114	87T	memory
sv_matmul	34G	302M	114	87T	memory
softmax	671M	537M	1.25	960G	memory
norm	59M	34M	1.75	1T	memory
add	8M	34M	0.25	192G	memory
Decode					
q_proj	34M	34M	1	768G	memory
k_proj	34M	34M	1	768G	memory
v_proj	34M	34M	1	768G	memory
o_proj	34M	34M	1	768G	memory
gate_proj	90M	90M	1	768G	memory
up_proj	90M	90M	1	768G	memory
down_proj	90M	90M	1	768G	memory
qk_matmul	17M	17M	0.99	762G	memory
sv_matmul	17M	17M	0.99	762G	memory
softmax	328K	262K	1.25	960G	memory
norm	29K	16K	1.75	1T	memory
add	4K	16K	0.25	192G	memory

NuRAM: Managed MRAM “SmartMem” Technology

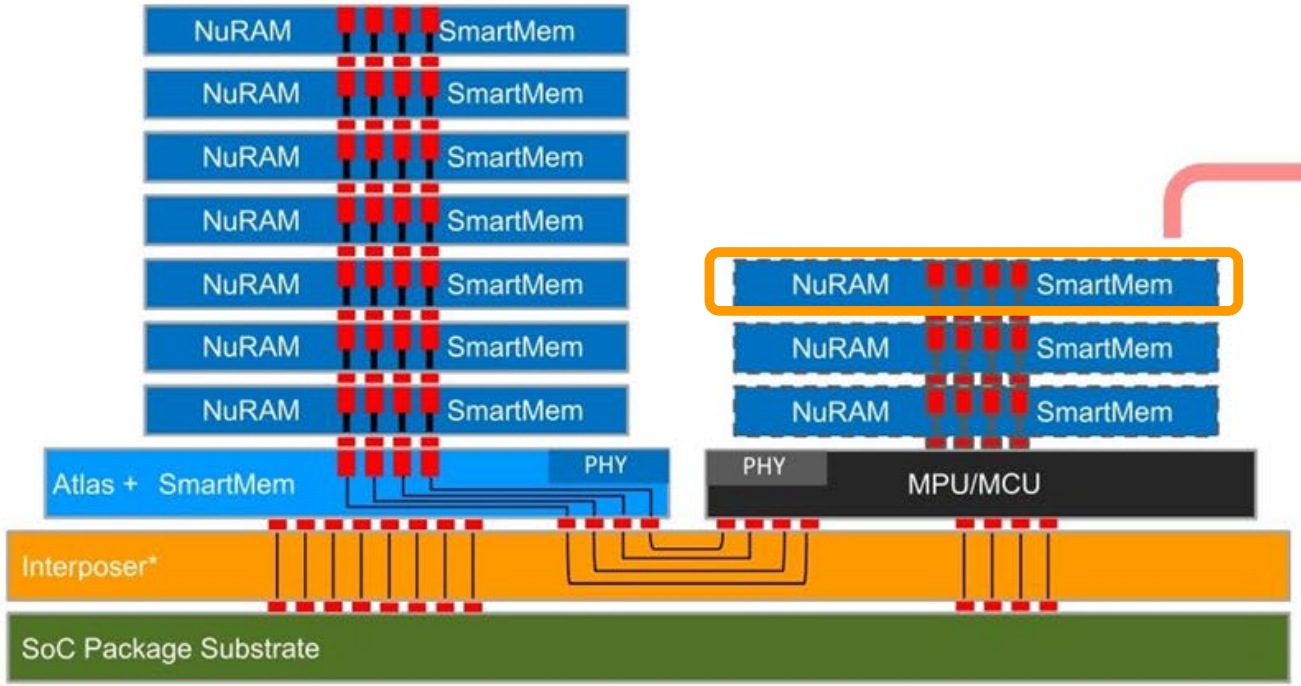


Fewer Memory Components (Space),
Less Power, & Better Performance

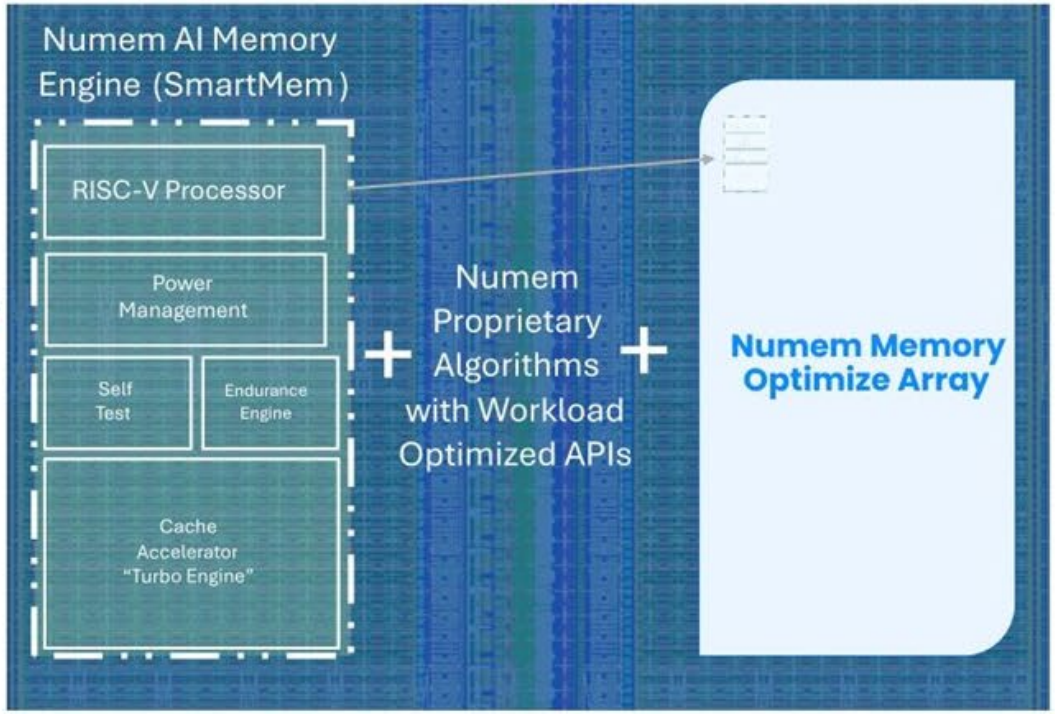
NuRAM is replacing the current tier, eliminating
SRAM/DRAM Bottleneck (Space/Power/Performance)

3X HBM BW and DRAM like GB’s capacity via stacking,
1000X lower standby and 2.5X denser than SRAM

AI Memory engine : deep dive



Numem SoC example (I/F : UCIe, HBM, LPDDR, etc.)



Numem Each Chiplet

Gen 1 12nm: 2-6TB/s, 3-9GB/stack

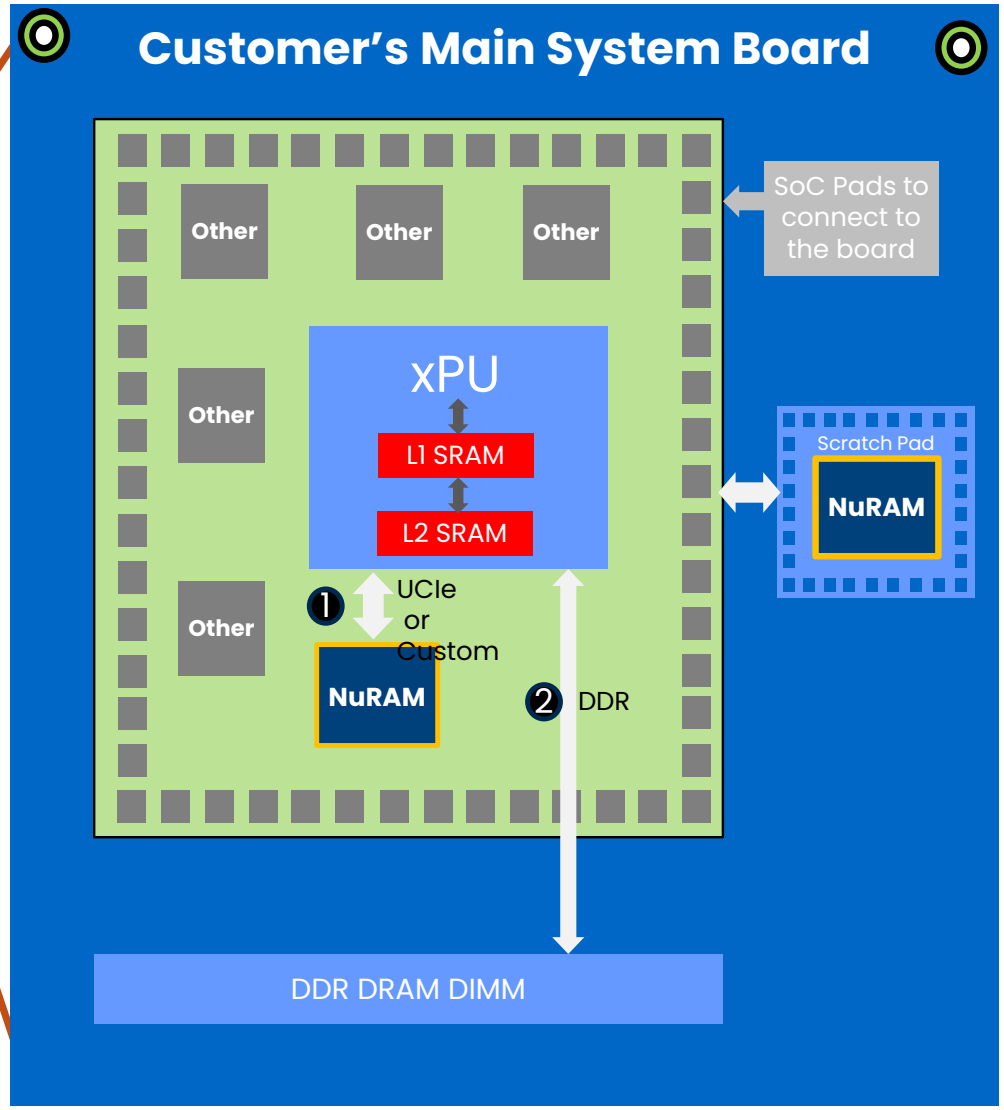
Gen 2 5nm: 4-12TB/s , 8-16GB/stack

Three Key Ingredients

1. AI Memory Engine
2. Numem Algorithms
3. Numem Array Design

NuRAM: Server SOC integration

CPU Server Chassis



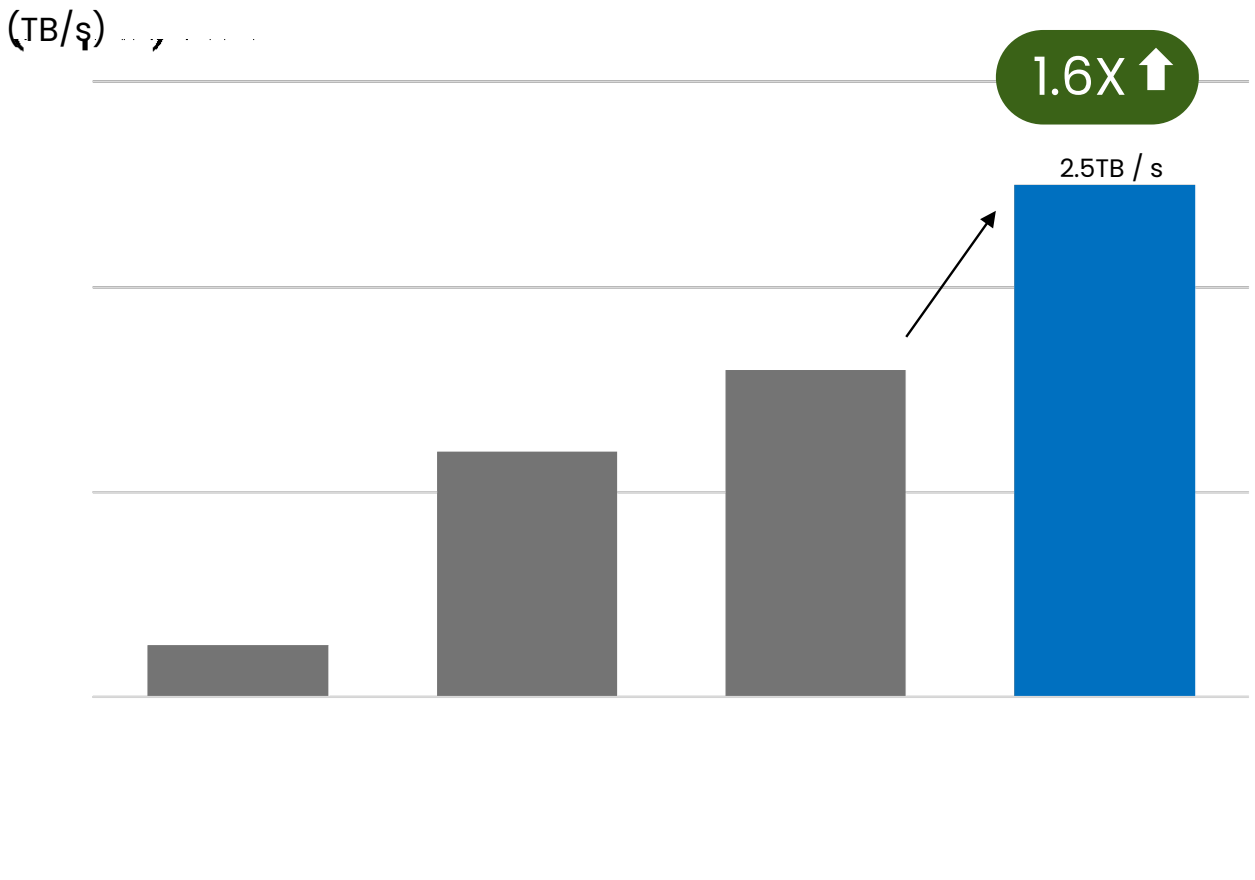
Numem Solution Benefits

1. **Use NuRAM as the fast LLM (Last Level Memory)**
   - Keep the NuRAM in SoC
   - Fast Interface (UCIe, SRAM, etc.)
   - Much faster latency than DRAM
   - Ultra low power (stand by)
2. Instant-on by non-volatility

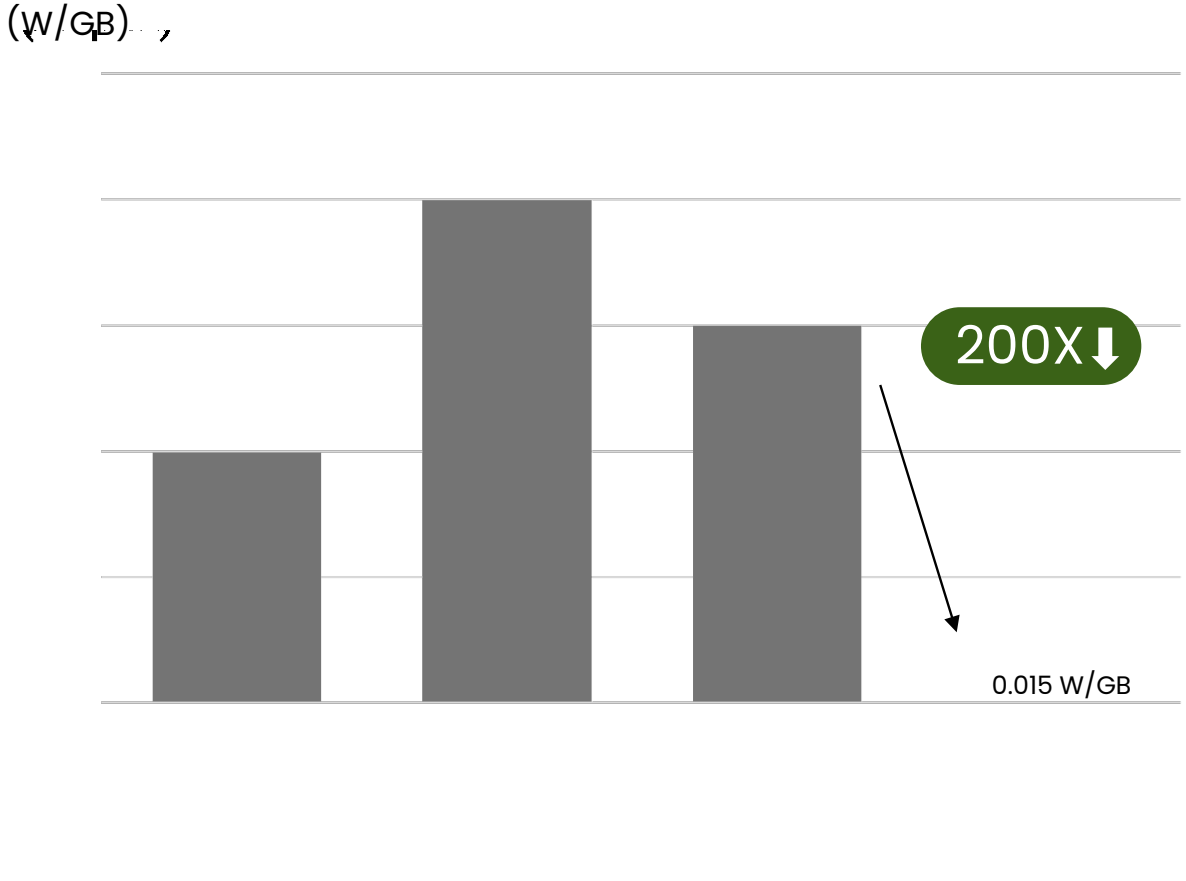
- **Less Access to DRAM DIMM**
  - Less interaction with DIMM (less wait time)
  - Potentially less DIMM size -> lower power

Benchmarks

Numem Bandwidth against AI Memory Modules



Numem Standby Power against AI Memory Module



Are Lossy Quantized , Pruned LLM models compressible ?

Block based Compression Algorithms

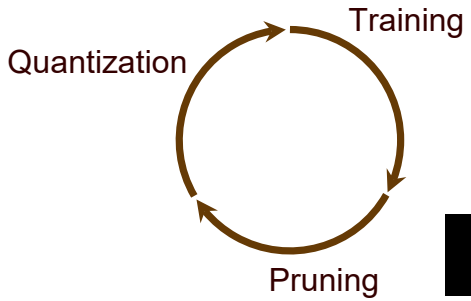
Industry Standard Algorithm	Compression Ratio	Block size
LZ4	1.0X (no compression)	64Kb
ZSTD	1.25X (us Latency)	
Deflate	1.25X (us Latency)	
Snappy	0.99X (no compression)	

Cacheline Algorithm

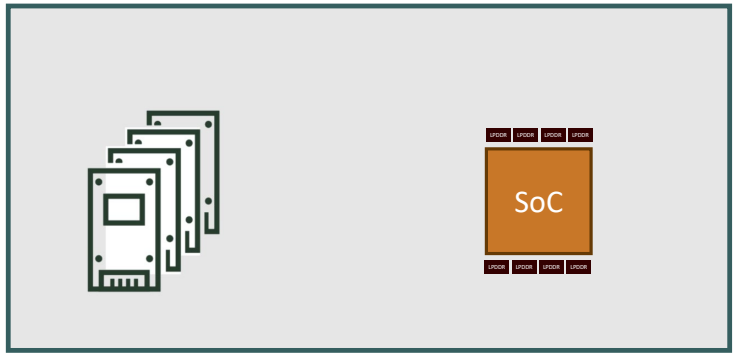
	Compression Ratio	Block size
ZeroPoint	1.5X +	64 byte

Cacheline compression on Foundational Models :
1.5X real time (de)compression with nanosecond latencies

Addressing the LLM cost of transferring and storing on NVMe and HBM/LPDDR



Offline compression
ZeroPoint algorithm



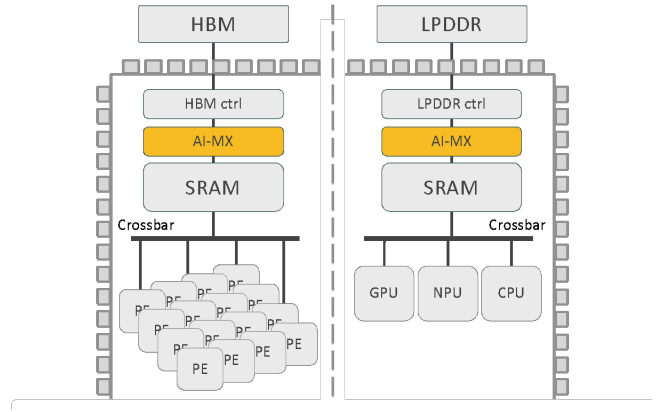
Opportunity

- Foundational models are compressible
- ZeroPoint algorithms show CR: 1.5+
- What if we add offline compression before we store the Inference model and we add inline decompression when we read the model from LPDDR?

Value proposition

- Increased available LPDDR capacity by up to 50%
- Increased available transmission and LPDDR BW
- Increased available storage capacity

Inline decompression

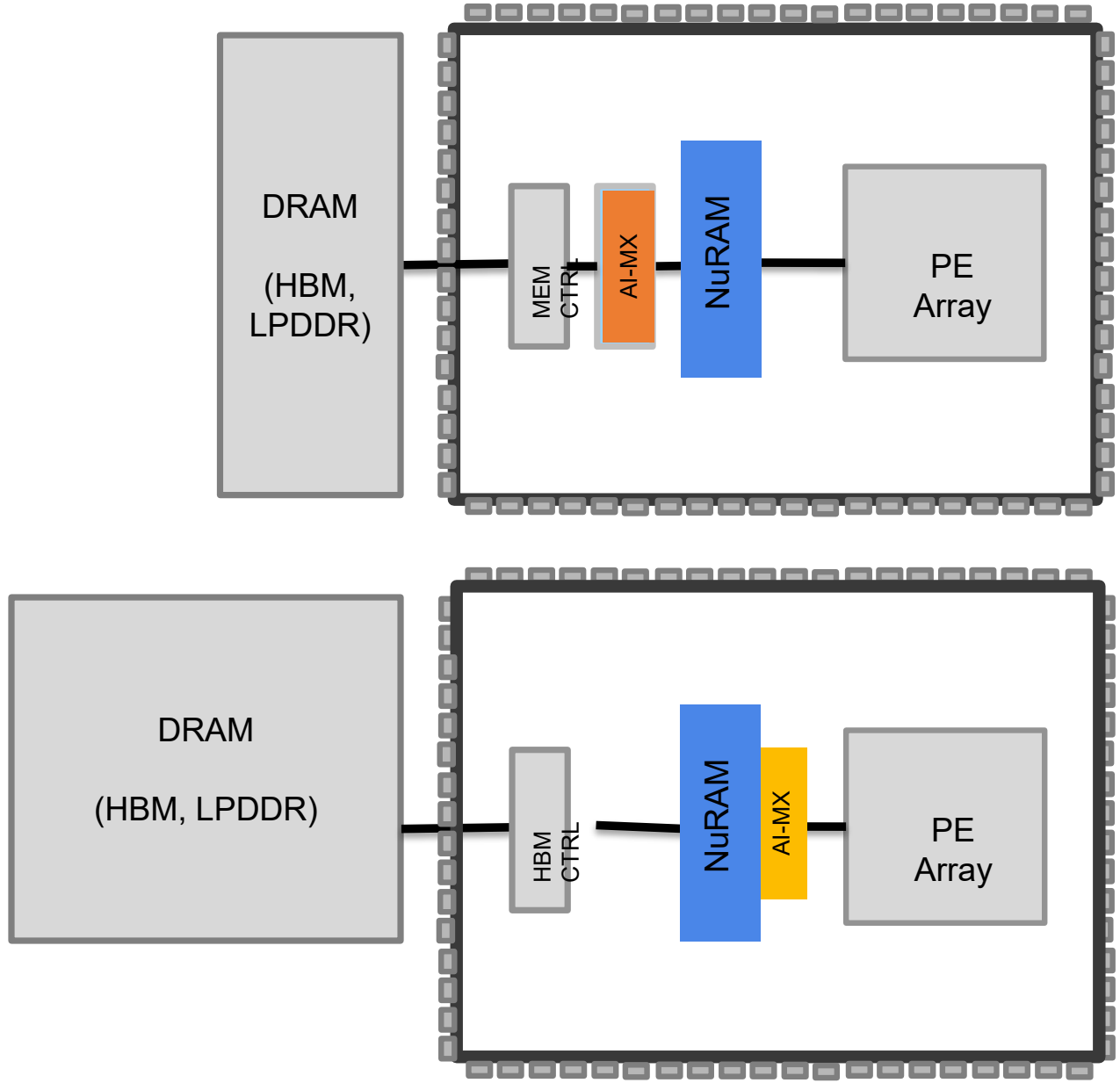


AI-MX – How does this work

1. AI-MX IP must be integrated on the accelerator SoC
   (in the memory subsystem, or close to the PE)
1. Model data is compressed in software before being loaded to the system
2. Model data is loaded to the system (along with compression metadata) & AI-MX is configured
3. Model data is requested by PE from DRAM, being inline decompressed by AI-MX

Compression approach

- Through the ZeroPoint proprietary and patented lossless compression X21 algorithm
  (provided as a software library)

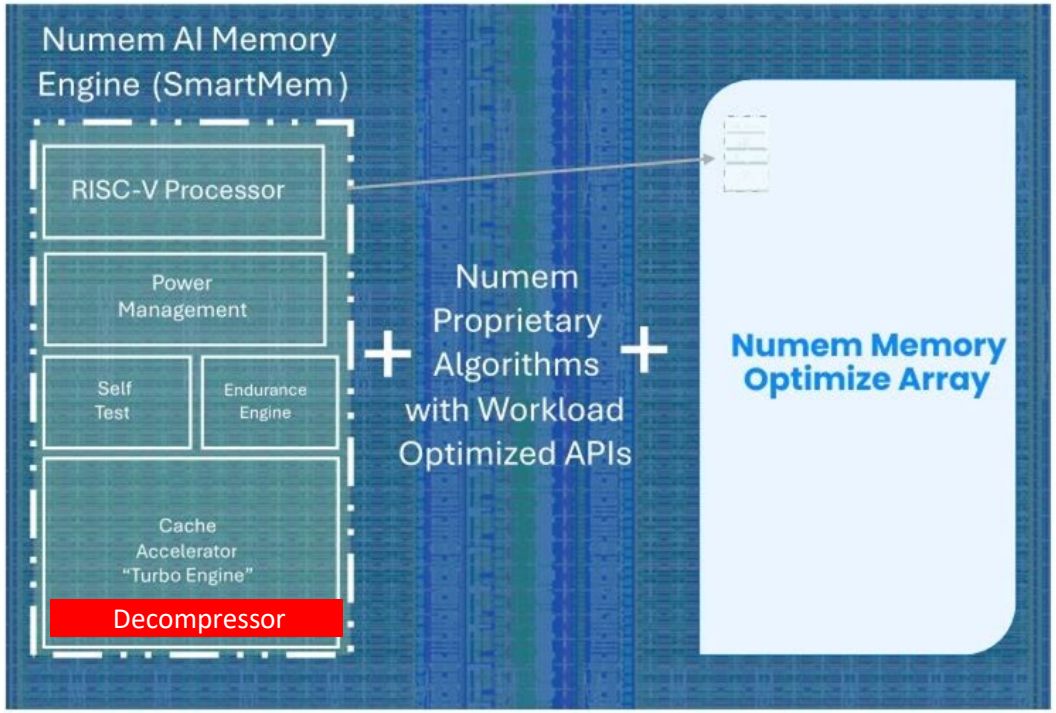


Solution: integrated compression engine on managed MRAM SmartMem controller

Fit larger models

Amortize cost of MRAM via larger effective capacity

Increase effective BW



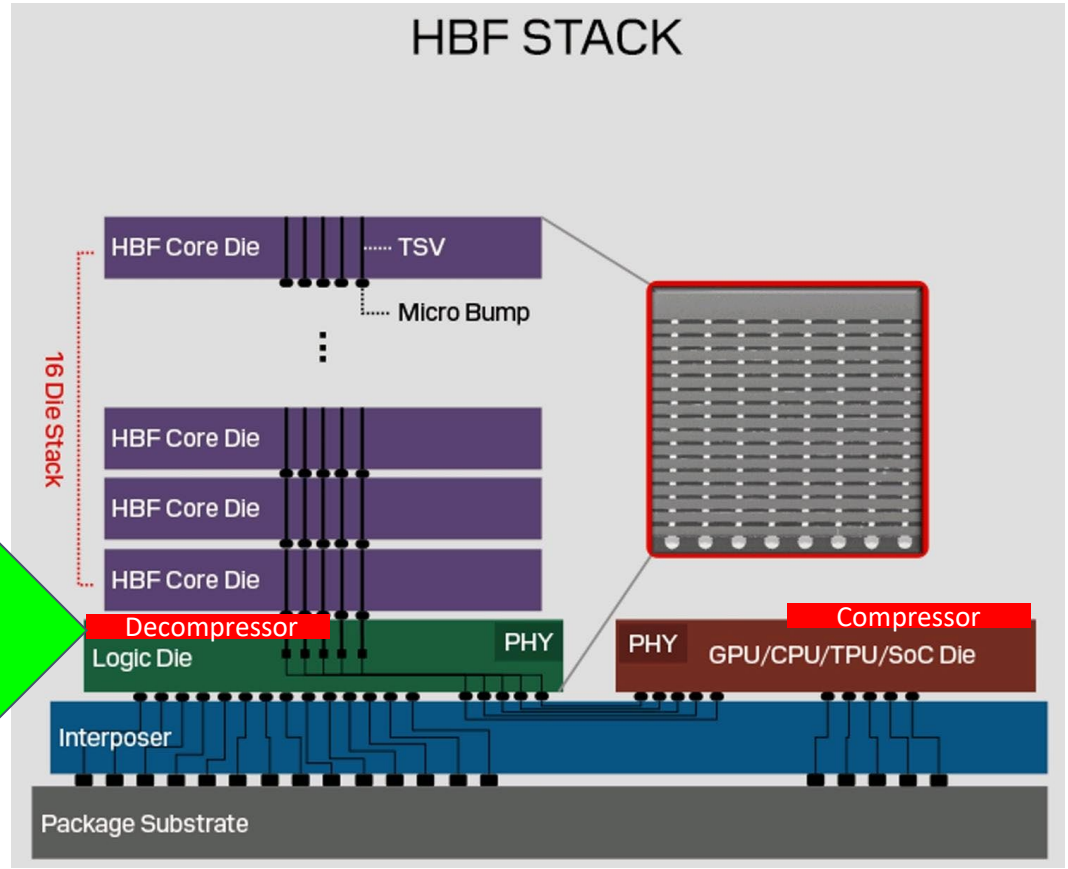
Inline compression + Managed MRAM have potential to deliver 5X effective capacity over HBM

Future possibilities: integrated compression engine on HBF?

Boost HBF Bandwidth, capacity by 2X

Proposal: Lossless (de) compression engine on Logic Die

~~Compressor in SW, Compute die~~



Base Image Source: SanDisk
https://www.sandisk.com/company/newsroom/press-releases/2025/2025-07-24-sandisk-forms-hbf-technical-advisory-board-to-guide-development-and-strategy-for-high-bandwidth-flash-memory-technology

Decompressor inserted speculatively

Inline compression +HBF potential to deliver 2X effective capacity over HBF

Datasheet– AI-MX_Gen1 (Pool of decompressors)

Characteristics	Value
Compression granularity	Only 32B or 64B
Clock frequency	Up to 1.75 GHz (Samsung 4nm / TSMC N5 technology)
Bandwidth	Matches the B/w of 1 HBM3e channel (@1.2GHz)
Decompression latency	Pipelined decompressor; Deterministic latency; Default: 21/ cycles +1/ foreach next decompressed 64B block of the compressed package (or stream)
Area and peak power* / IP instance	Area = 0.04mm2 Peak Power = 0.057W (Samsung 4nm, Low Power Plus (LPP)) @1.2GHz
Compression ratio (geomean) 100% lossless	1.5x for Llama model data at bf16 1.25x - 1.43x for Llama model data at OFP8 (e4m3 / e5m2)

Area and Power efficient IP for LLM model (de) compression

Summary/ Call to Action

Summary:

- LLM model inference memory bound

- Compression enabled Managed MRAM technology offers path to differentiate, compete with higher Tokens-per second- per watt

Call To Action:

- Partner to sample combined IP within your SoC Architecture
- Collaboratively evaluate performance via memory trace
- Collaboration to develop use cases,benchmarks