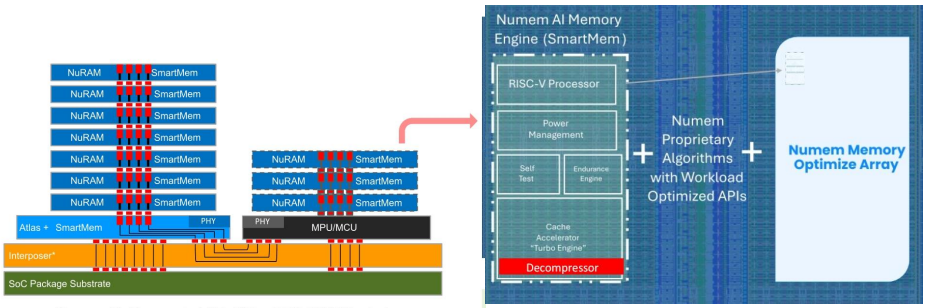
OCP FUTURE TECHNOLOGIES SYMPOSIUM | 29 APRIL, 2025 DUBLIN, IRELAND

Compression enabled MRAM memory chiplet subsystems for LLM Inference Accelerators

Angelos Arelakis , Nilesh Shah, Evangelos Vasilakis
ZeroPoint Technologies Gothenberg, Sweden

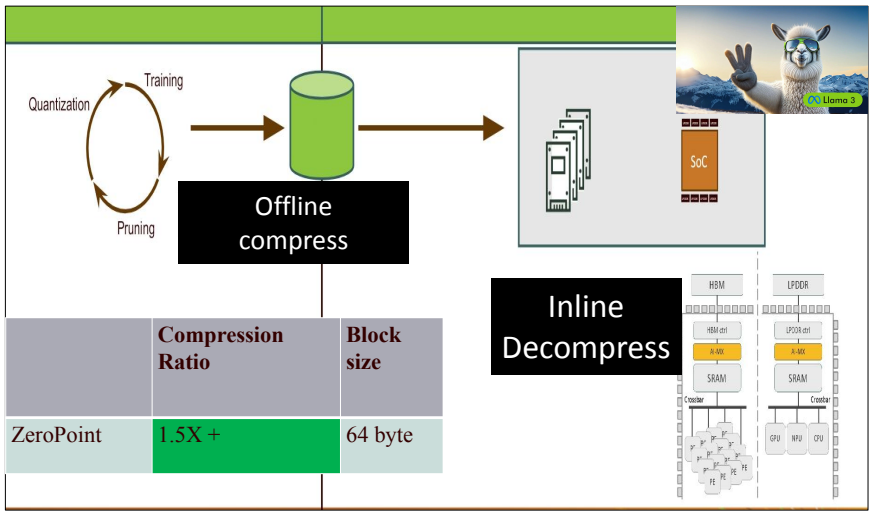
Max Simmons , B Nataraj, Rama Gorti
Numem , Sunnyvale, CA

- **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



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

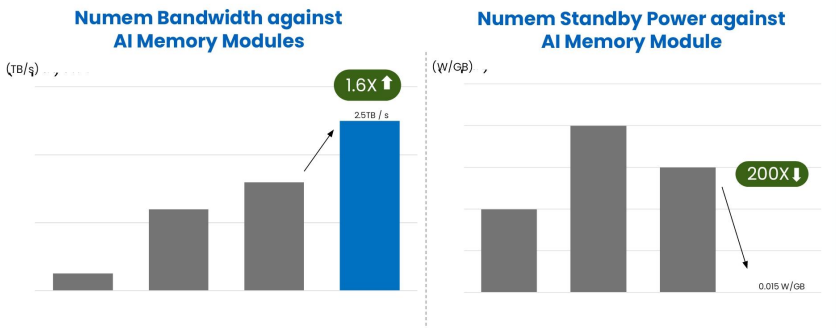
Decompressor IP	SPEC
Compression granularity	Only 32B or 64B
Clock frequency	Up to 1.75 GHz (Samsung 4nm / TSMC N5 technology)
Bandwidth	Matches the B/w of 1 HBM3e channel (@1.2GHz)
Decompression latency	Pipelined decompressor; Deterministic latency; Default: 21/ cycles +1/ foreach next decompressed 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)

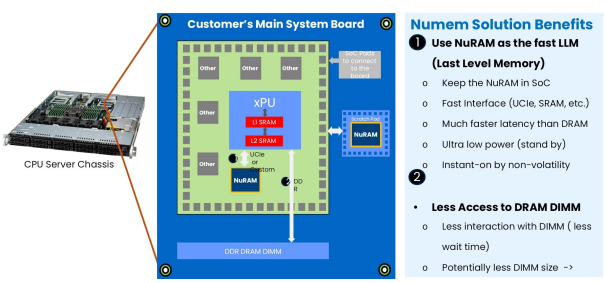


	Compression Ratio	Block size
ZeroPoint	1.5X +	64 byte

NuRAM Managed MRAM SmartMem: 3X HBM BW and DRAM like GB's capacity via stacking, 1000X lower standby and 2.5X denser than SRAM

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





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)
   - Instant-on by non-volatility
2. Less Access to DRAM DIMM
   - Less interaction with DIMM (less wait time)
   - Potentially less DIMM size ->

1st generation
Model (LLM) memory SW compression and HW decompression
- Compressing LLMs in SW before deployment, reducing the LLM memory footprint
- Memory capacity is increased by up to 50%. The model is decompressed in HW when read from memory, improving bandwidth by up to 50%
- Significantly reduces storage requirements and network transfer times while maintaining model quality

2nd generation
AI memory bandwidth acceleration
- Compression and decompression in HW
- Increasing total memory bandwidth by more than 50%
- Enables compression and decompression of model (LLM), key-value cache data, and other data, further expanding memory capacity and improving inference speed

3rd generation
AI memory capacity expansion and bandwidth acceleration
- Compression and decompression in HW
- Increasing memory capacity and bandwidth by more than 50%
- Enables compression and decompression of both the model (LLM) and key-value cache data, further expanding memory capacity and improving inference speed