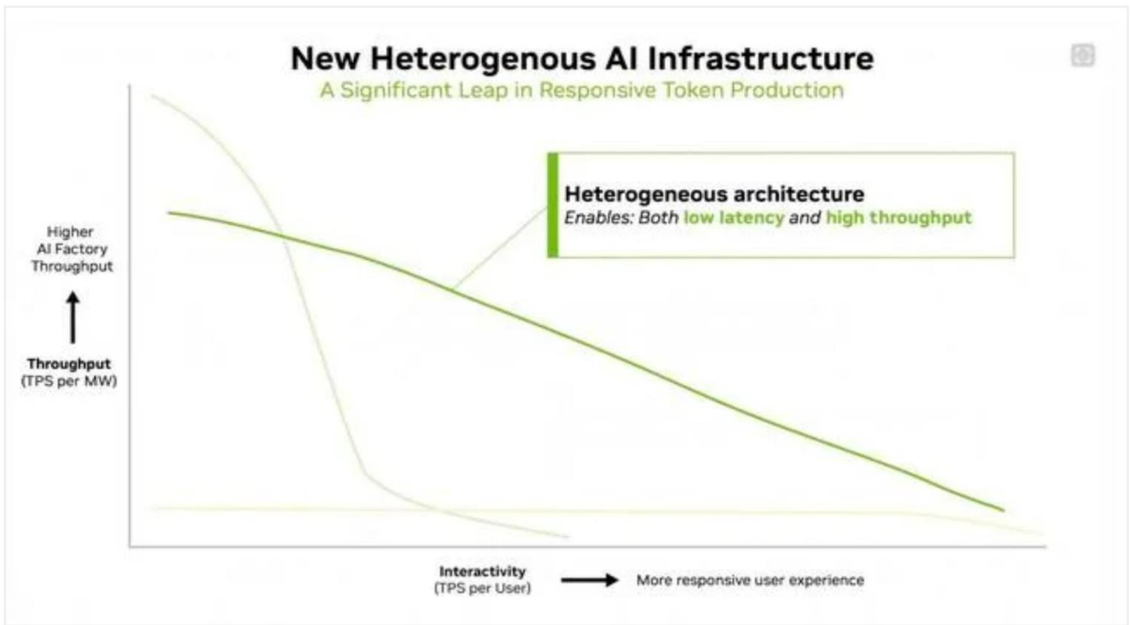
Enhancing AI Data Center Efficiency: OCP OFP8 Lossless Compression for LLM Inference

AI Infrastructure Has Standardized, Not Optimized

- OCP OFP8 (E4M3, E5M2) becoming standard
- Next: OCP MX formats (FP4 / INT8 mix)
- Inference now heterogeneous: GPU + LPU + CPU + storage
- Performance depends on data movement across system
- OCP formats optimize precision, not movement



In this graphic, the faint green and yellow lines show GPU and LPU scaling. By combining the two, Nvidia aims to deliver the best of both worlds — high throughput and interactivity. Image

Precision Reduced , Memory Pressure Persists

- Inference decode = memory bandwidth dominated
- Streaming weights + KV cache every token
- SRAM fast but small, HBM large but constrained

Measured results:

- 1.5x compression (BF16), 1.33 to 1.43x (OFP8)
- FP8 still has redundancy:not fully optimized

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

Compression Extends OCP format gains to System Level

- Lossless, cacheline-level (64B) compression
- Inline across GPU, LPU, DRAM, CXL, storage
- No retraining, no accuracy loss

Impact:

- 1.3 to 1.5x effective bandwidth/capacity
- 20 to 50% higher tokens/sec, up to 25% lower TCO
- Compression = system-level multiplier

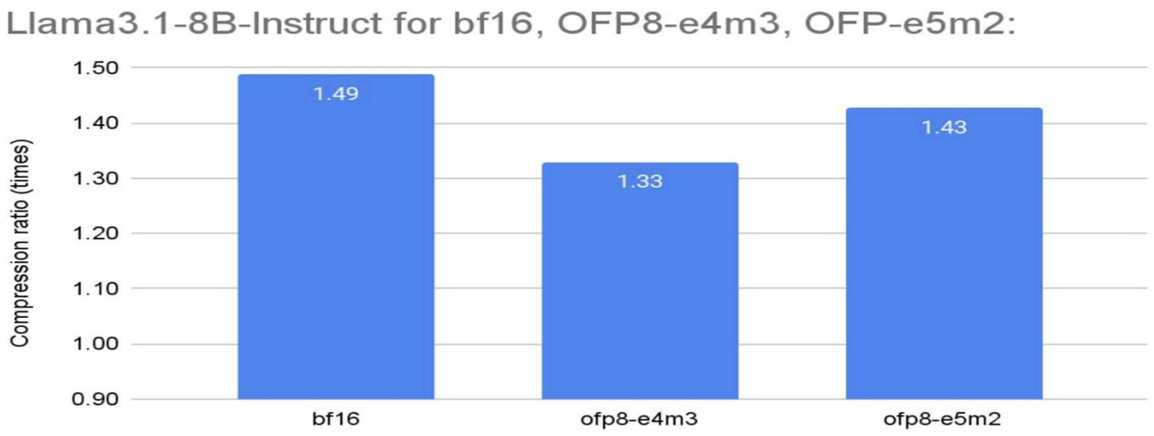


Figure 1: Compression Ratio(CR) achieved on BF16, OFP8 formats with pre-trained LLM model

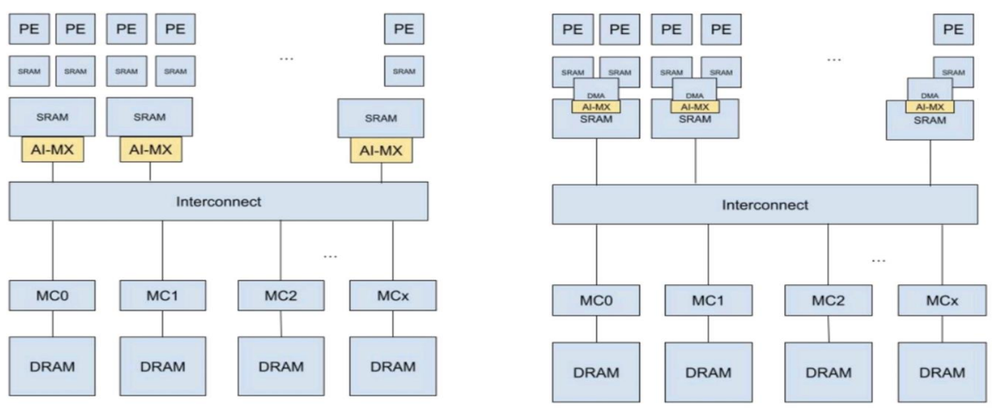


Figure 3: Hardware accelerated (De)Compression IP (AI-MX) Integration options