Optimizing Foundational Models: Hardware-Accelerated Memory Compression & Interconnects

Nilesh Shah
ZeroPoint Technologies
&
Rohit Mittal
Auradine

Gen AI: Trends

Inference spend dominates

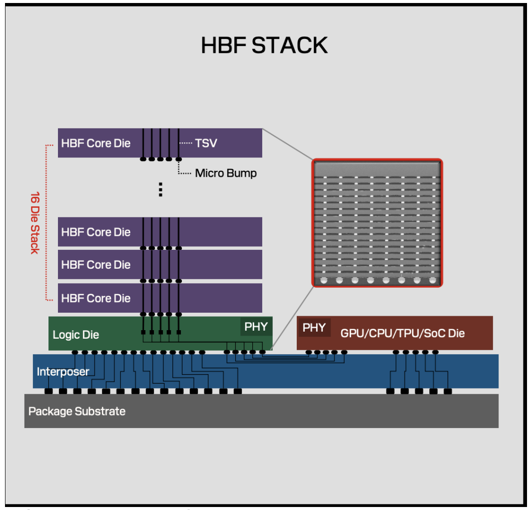
Source: https://menlovc.com/2023-the-state-of-generative-ai-in-the-enterprise-report/

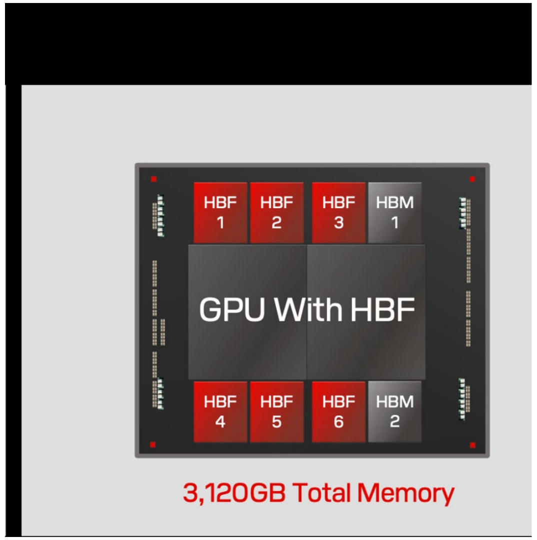
Insatiable model memory needs: Scale Memory

High Bandwidth Flash (HBF), targeting 8× HBM capacity for AI inference at similar cost

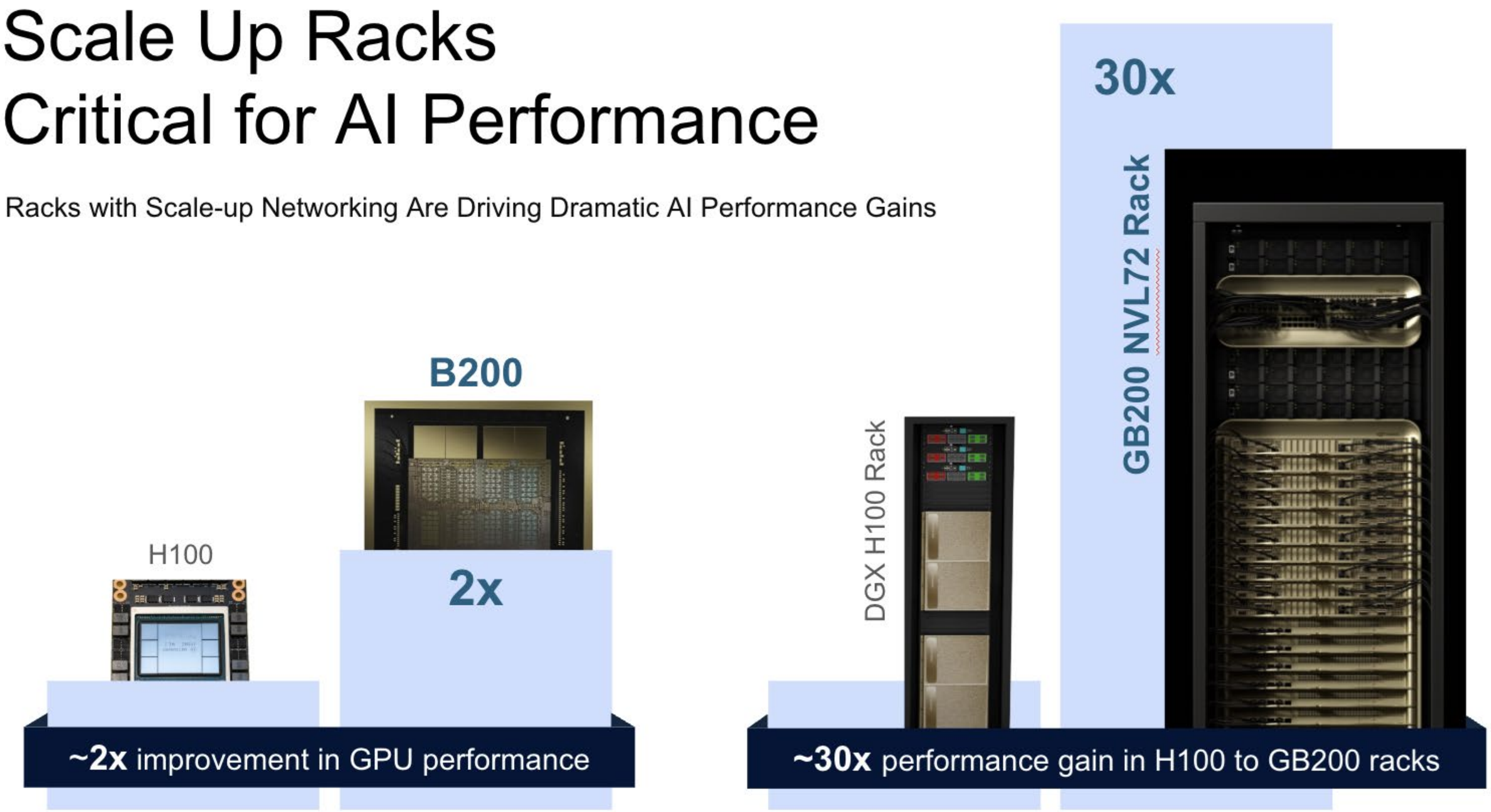
Leverages **BiCS and wafer bonding 8-16X**

Source: https://investor.sandisk.com/static-files/79481580-ada2-4e08-bdeb-4b440d08f4ab



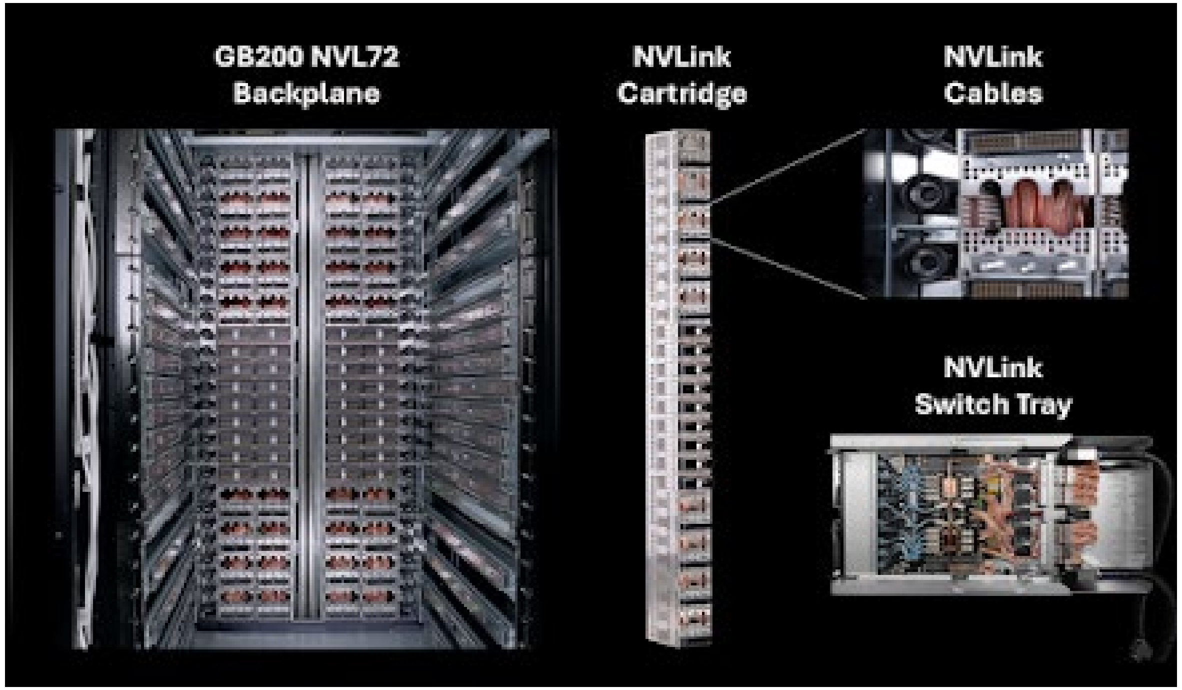


GenAI: Importance of Rack level interconnects



Source: semianalysis, industry sources

Scale Up Interconnects for AI factories



Rack is the differentiator

Key Takeaway: Silicon by itself is not a solution. Rack (with scale up) is the unit of AI

Source: semianalysis, industry sources

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

[LLM Inference Unveiled: Survey and Roofline Model Insights]

Custom Accelerator Racks: Memory Hierarchy as differentiator

Accelerator	Total Memory per Rack (TB)	Notes
Cerebras CS-3	1.2–1200	MemoryWall architecture with external memory fabric (**SRAM** + Flash)
NVIDIA DGX H100	2.56	8x H100 GPUs per system, 4 systems/rack, 80GB HBM2e each
AMD MI300X	4.5	8x GPUs/rack, 192GB **HBM3** per GPU
Biren BR104	4.5	Assumed 8x 564GB systems per rack
FuriosaAI RNGD	3.0	Estimated from 80 chips x 32–40GB **DRAM/SRAM**
Rebellions REBEL	3.0	Similar assumption to Furiosa
Groq LPU	0.2	On-chip **SRAM** only, no external DRAM or HBM
d-Matrix M1200	1.0	64 SoCs per rack x 16GB **eDRAM** each
SambaNova SN40L	10	**CBA+HBM+Flash** hybrid; storage-rich inference rack

Cerebras NVIDIA

Table 1: Rack ISO Space - CS-3, DGX H100, and DGX B200 Components

Component	WSE-3	H100	B200
Chip Size	46,225 mm^2	814 mm^2	~1600 mm^2
# Cores/Chip	900000	16896 FP32	-
On-Chip Memory/H100	44 GB	0.05 GB	-
System	CS-3	DGX H100	DGX B200
System Dimension	15U	8U	10U
# Chips/System	1	8	8
On-Chip Memory/H100	44 GB	0.4 GB	-
Memory Capacity	1.2-1,200 TB	0.64 TB	1.5 TB
System Power	23 kW	10.4 kW	14.3 kW
Price	$2.5M (est.)	$0.35M	$0.5M
Rack Dimension: ISO Space	30-32U	30U	32U
# Systems/Rack	2	4	3
# Chips/Rack	2	32	24
On-Chip Memory	44 GB	1.6 GB	-
Memory Capacity	1.2-1,200 TB	2.56 TB	4.5 TB
# Cores/Rack	900000	33792 FP32	-
Rack Price	$5M (est.)	$1.4M	$1.5M
Rack Power	46 kW	41.6 kW	43.9 kW

Source: A Comparison of the Cerebras Wafer-Scale Integration Technology with Nvidia GPU-based Systems

Insatiable model memory needs: Model compression

But, aren't AI workloads already compressed?

Foundational models **LOSSY** compressed during training

- Quantization – less accurate weights
- Pruning – fewer connections between weights

PRO: Model size reduces

***CONS**: **Accuracy reduces (garbage in garbage out), incur expensive retraining**

Does Lossless compression work?

YES!

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
**proprietary	1.5X + ns latency	64 byte

*source: ISCA'25: Meta's Second Generation AI Chip: Model-Chip Co-Design and Productionization Experiences

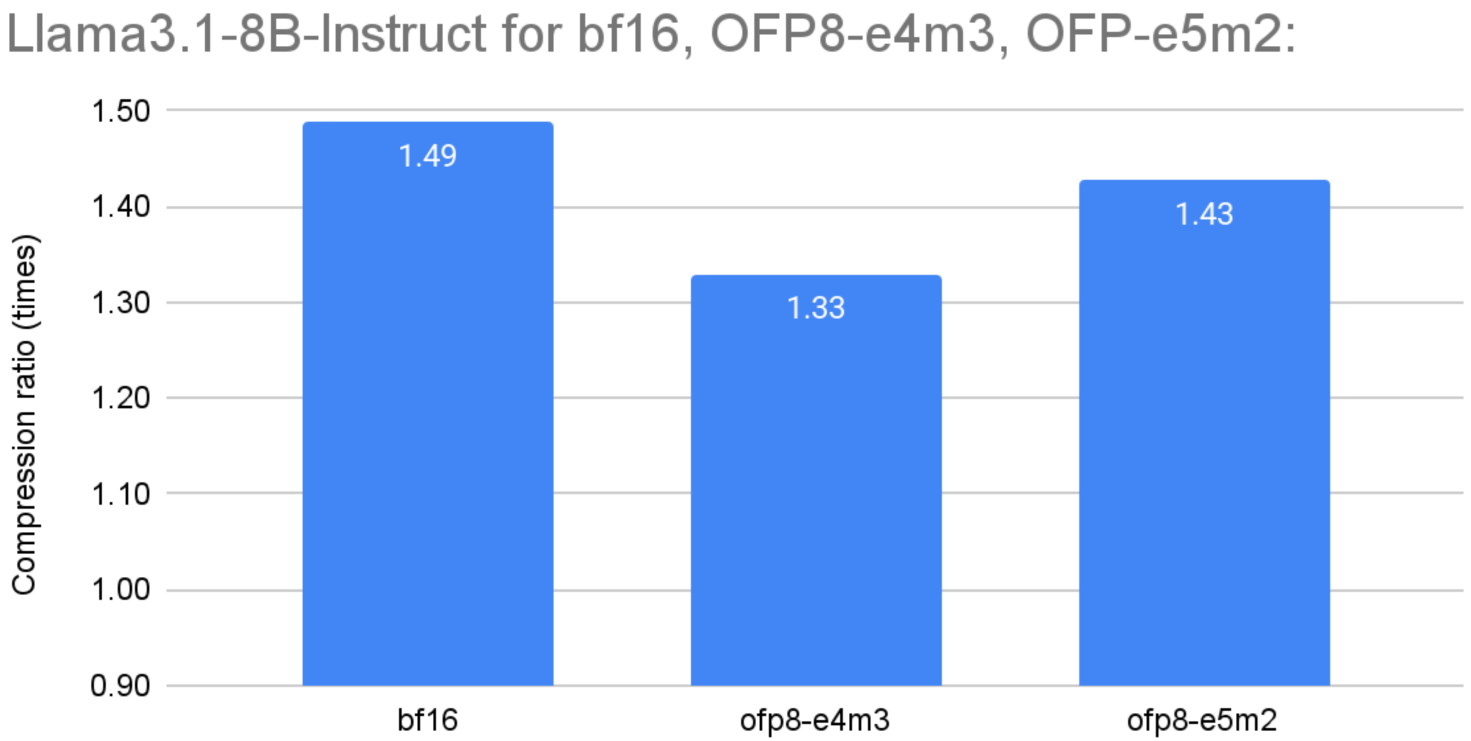
**source: https://www.zeropoint-tech.com/products/hbm-memory-expansion

Compression Performance – Proprietary Lossless cacheline compression algorithm

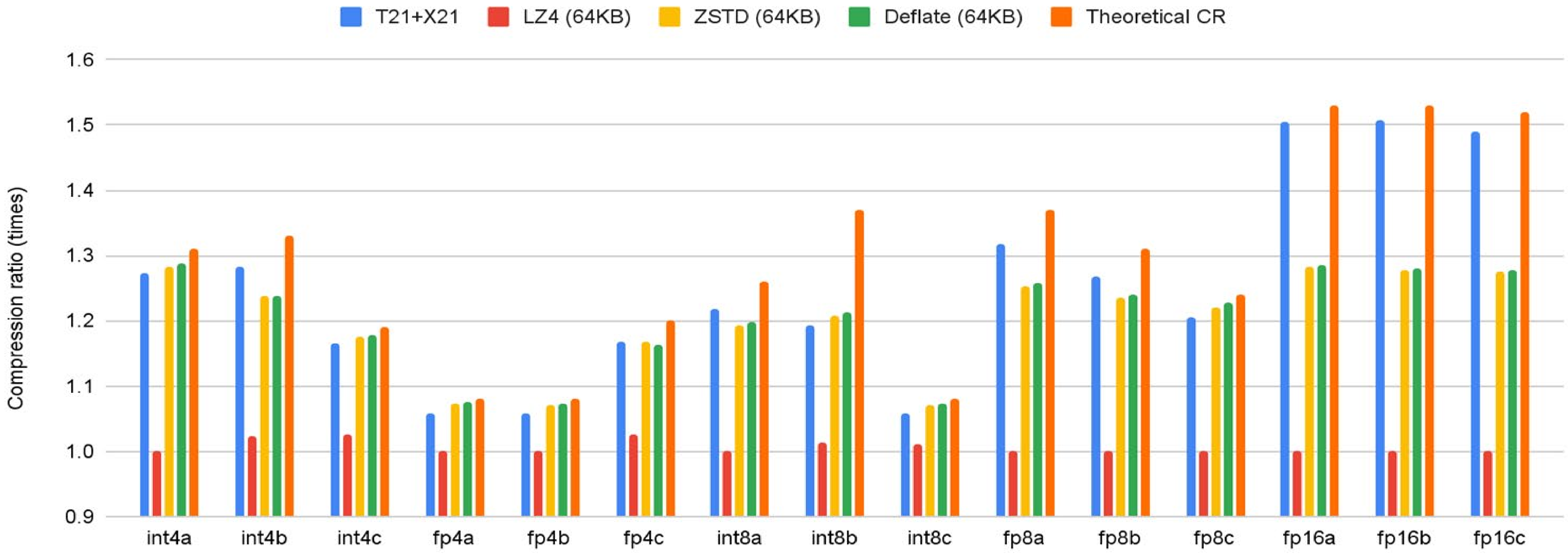
Compression ratio: results across all layers for Llama3.1-8B-Instruct

Data Formats:

- bf16
- fp8-e4m3
- fp8-e5m2



LLM Compression ratio: Proprietary Lossless compression algorithms



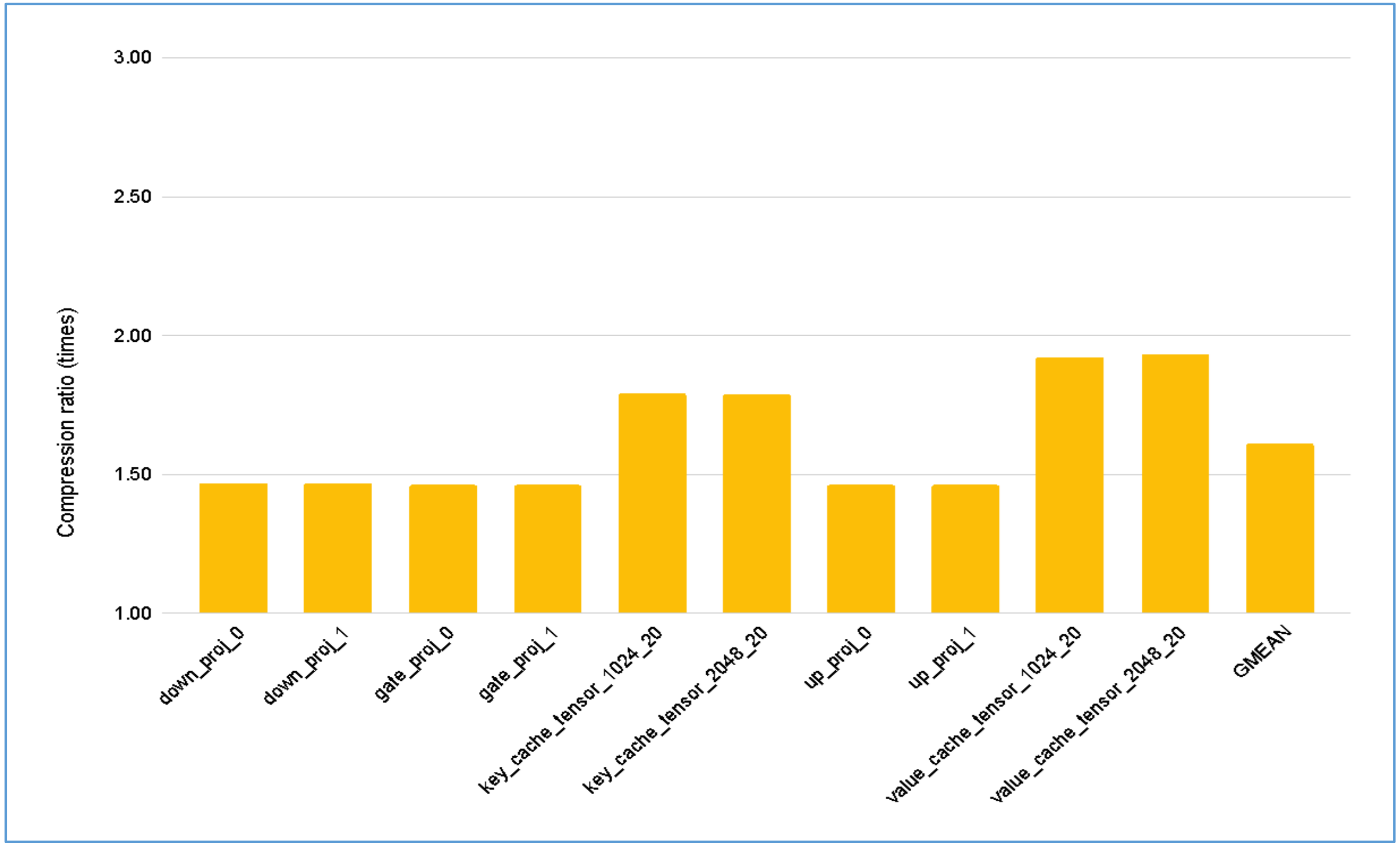
Hiigher compression ratio than state-of-art algorithms operating at **64B**
Nanosecond range (de)compression latency

Proprietary AI workload data set – Compression Performance

Propreitary X21 algorithm: compresses on **64B block** granularity

AI model data:
down_proj_(x)
gate_proj_(x),
up_proj_(x)

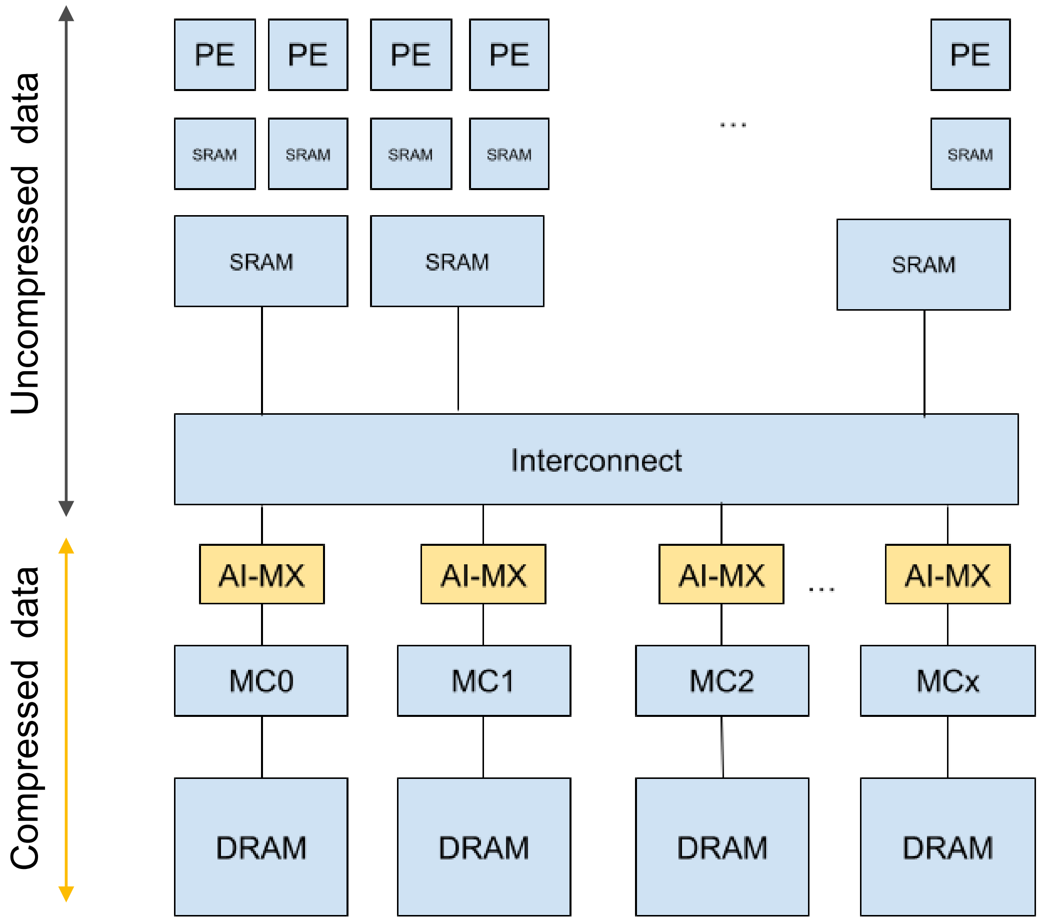
Key value cache data:
key_cache_tensor_(x),
value_cache_tensor_(x)



Compression Ratio performance:
Model data : 1.5X
KV-cache data : 2.0x

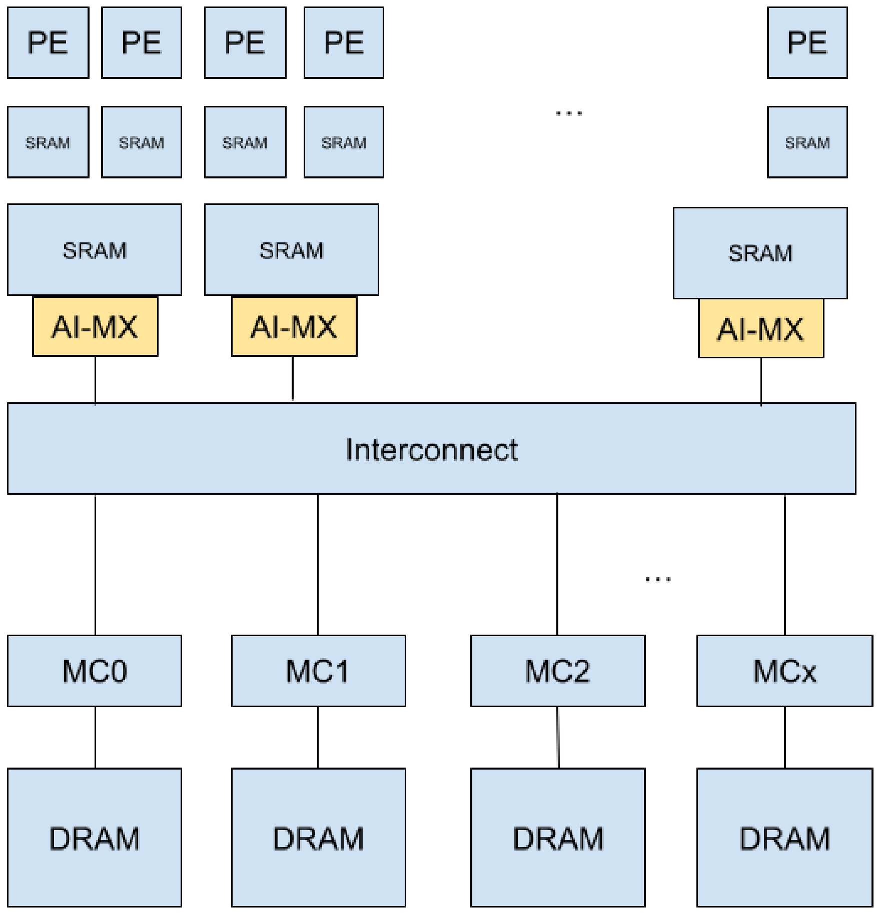
ASIC IP Block: Integration close to Memory Controller

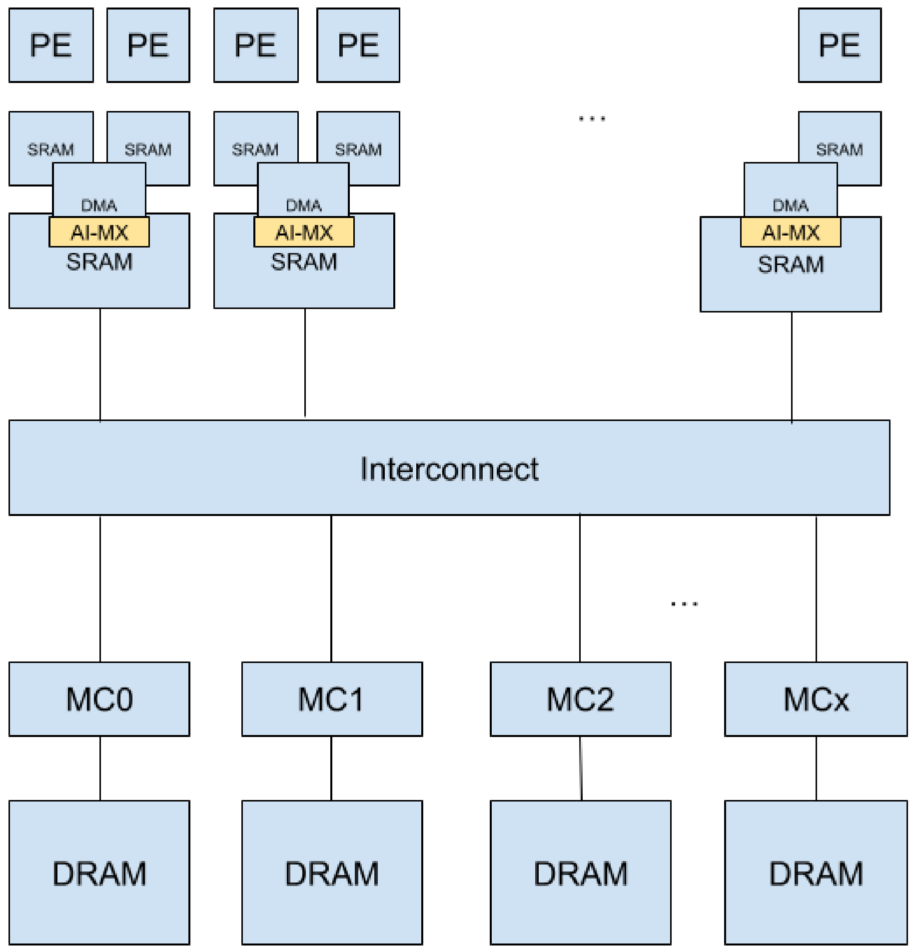
- Plug-and-play integration w/ standard interface
- One IP instance per memory channel
- Supports standard AXI5 interface specifications with 256b or 512b data bus
- Supports ECC, error logging and reporting over AXI5



Effective Bandwidth, Capacity gain

Alternate ASIC IP Integration : Closer to SRAM, DMA





Effective SRAM capacity gain

Memory Technology Agnostic IP Integration

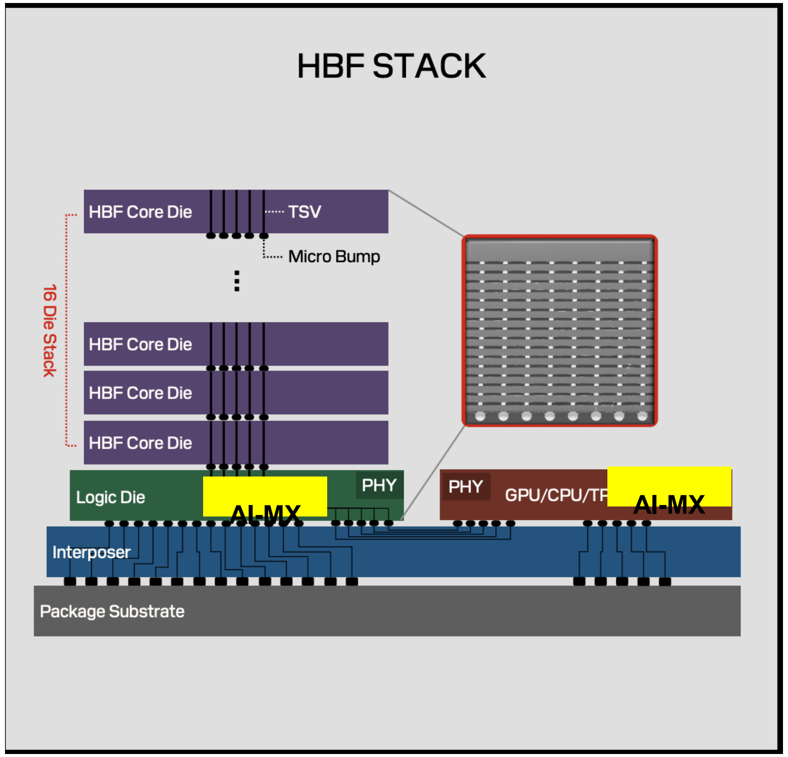
- Transparent Compression, Compaction and Address Translation
- “Drop-in” Compatible with most memory technologies, modular, scalable architecture ex: HBF

rebellions_

- **Use case**: High-performance and energy-efficient AI accelerator chips for data centers and hyperscale workloads

Source: https://www.eetimes.com/alliance-aims-to-deliver-memory-optimized-ai-for-inferencing/

Real world use cases



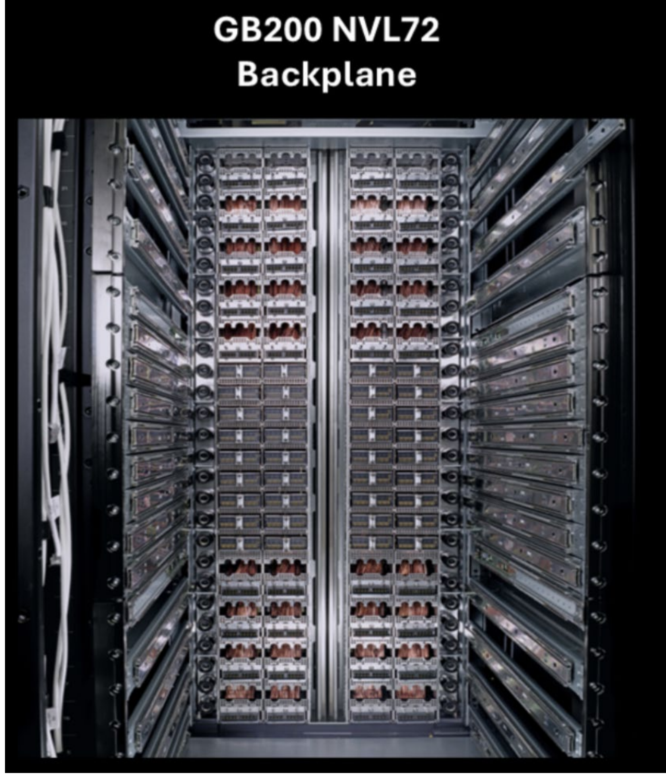
Future possibilities: HBM

Key Shifts Driving Rack-Scale AI

1. Physical Layer Innovation

- Cabled backplanes (e.g. NVL72) → dense copper, better SI
- Active midplanes (e.g. Kyber) → simplify assembly & cooling
- Next-gen connectors: NPC, CPC, CPO → high-volume manufacturable, low-cost, high-reliability

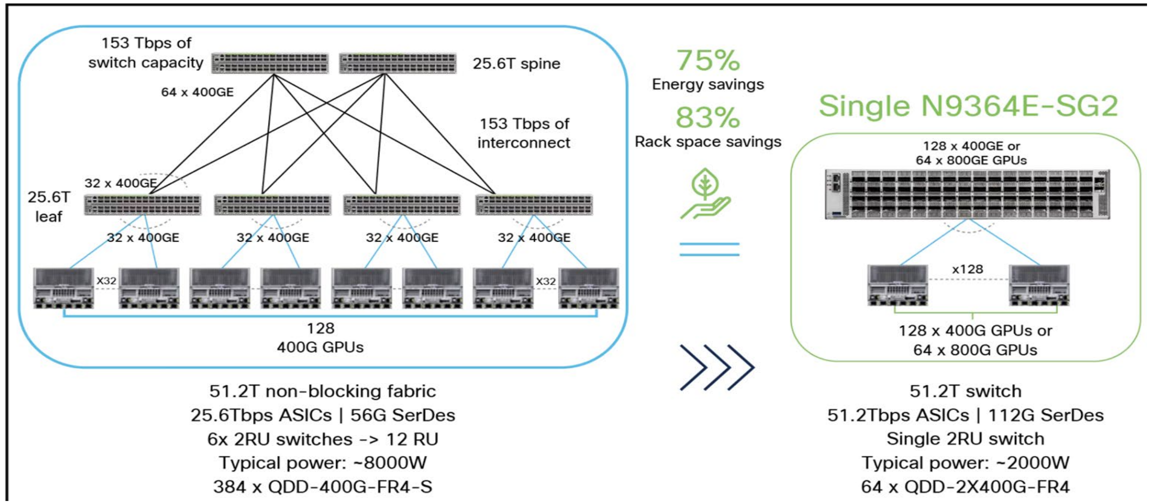
5,000 copper cables, GPU xGPU comms



8kW→2kW . Free up rack power budget

2. Topology Evolution

- Switch-based fabrics are baseline beyond 8–16 GPUs
- Advantages over direct mesh:
  - Linear scalability (avoid O(N²) link explosion)
  - In-network collective offloads (AllReduce in-switch)
  - Easier vPod creation & dynamic partitioning
  - Better cable management & SI via leaf-spine



https://www.cisco.com/c/en/us/products/collateral/networking/cloud-networking-switches/nexus-9000-switches/nexus-9000-ai-networking-wp.html

Key Shifts Driving Ecosystems: UALink – Scale-Up Fabric for the Rack

Key Specs of Ualink

- 200 Gbps per lane, up to 800 Gbps per port
- <1 µs RTT for 64B messages
- Scales to 1,024 accelerators across 4 racks
- Ethernet-based PHY for cost and commodity leverage

Why It Matters

- **Open Standard**: First memory-semantic fabric not locked to a single vendor
- **Memory Semantics**: Load/store/atomic access to peer GPU memory
  - Flattens software stack (no heavy RDMA)
  - Efficient for small transactions (<640B)
  - Treats rack as one logical memory domain

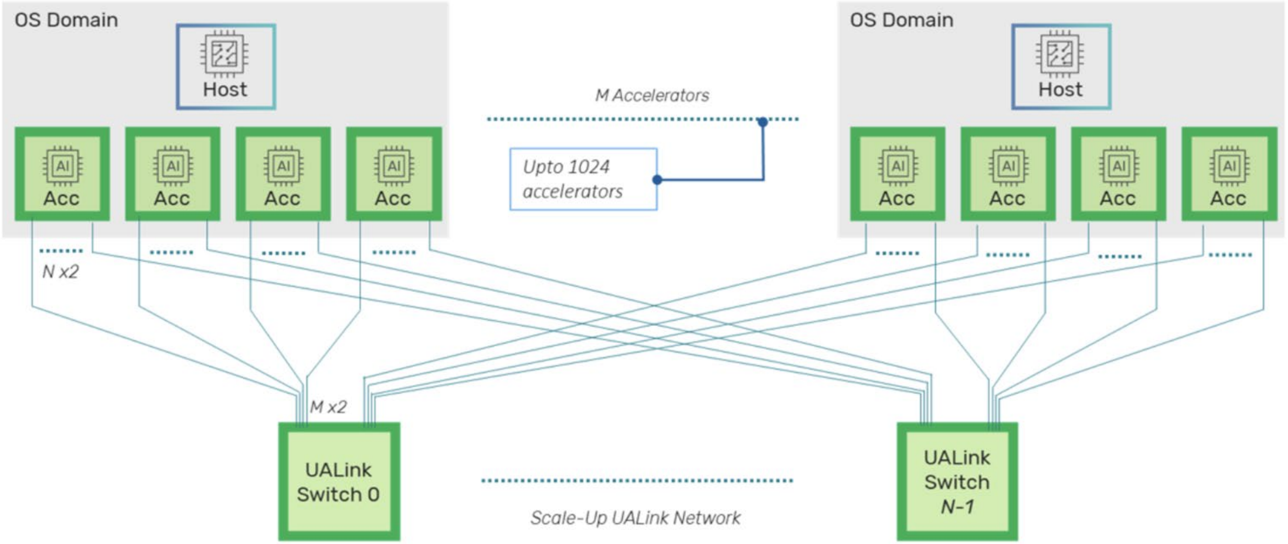


Figure 3: Scalable multi-node accelerator system with UALink high-speed interconnect

https://ualinkconsortium.org/blog/ualink-200g-1-0-specification-overview-802/

Efficient Scale up fabrics, Efficient model compression

Summary

- LLM Models & KV Cache compressible
- Inference Accelerator performance Memory Bound
- Scale up interconnects are critical for perf/TCO

Call To Action

- Cacheline granularity, lossless compression IP block democratizes access to greater effective memory capacity
- **IP Sampling Now**: Collaborate to validate use cases, data formats, new models/ KV Cache data
- Consider **open** ecosystems for **scale up interconnects**

Combine Chip level IP & Open scale up interconnects to Democratize AI access to all providers