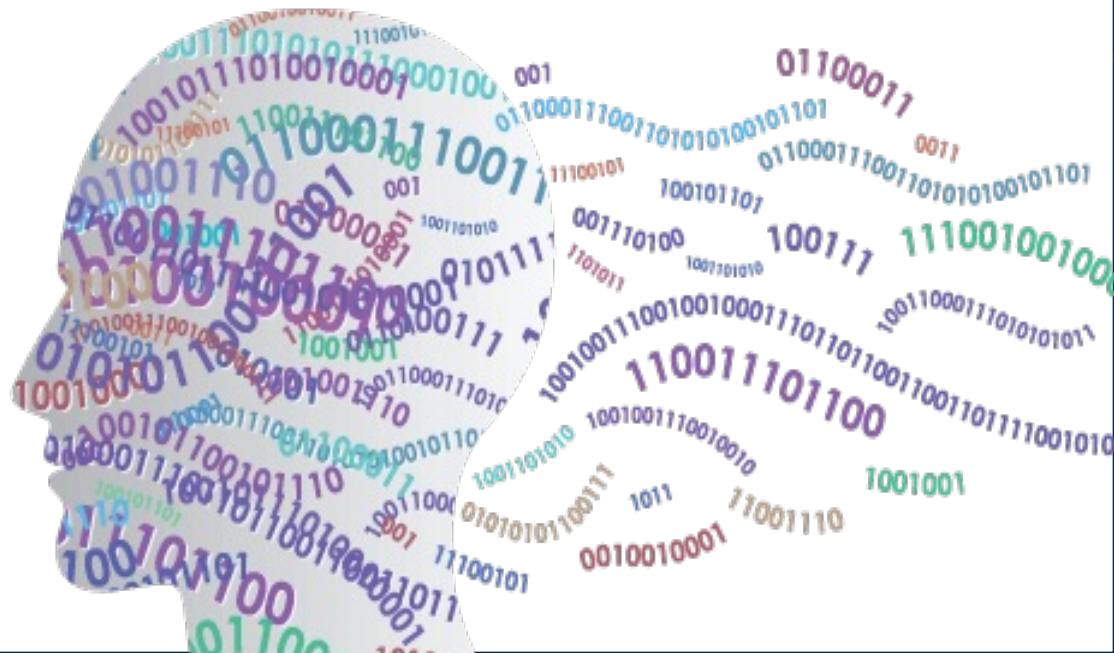
AI: Pushing Infra boundaries

Memory is a key factor

SNIA COMPUTE, MEMORY, AND STORAGE SUMMIT

Solutions, Architectures, and Community

VIRTUAL EVENT, MAY 21-22, 2024



Presented by

Manoj Wadekar, Meta

AI Systems Technologist

Meta – Community Statistics

3.98B

people use at least one of Meta services **monthly**, approximately

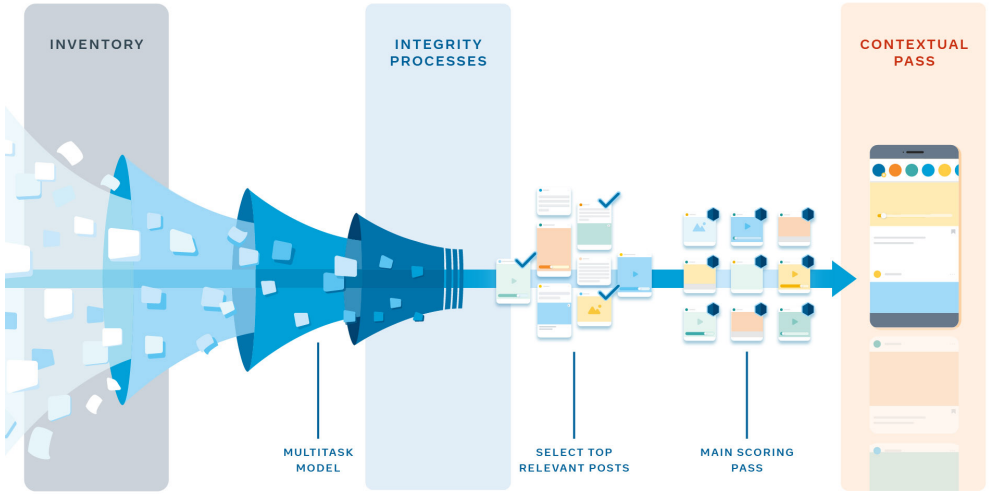
3.19B

Family **Daily** active users

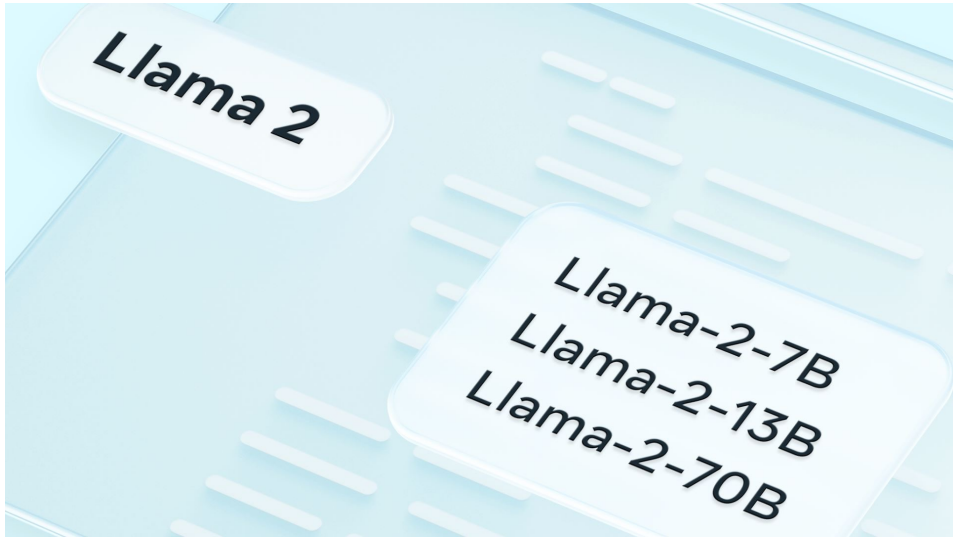
Ref: Meta 4Q'23 Results

AI Use Cases at Meta

- **Ranking and Recommendations**
  - Personalized Recommendations
  - Deep Learning Recommendation Models (DLRM)
  - Training and Inference



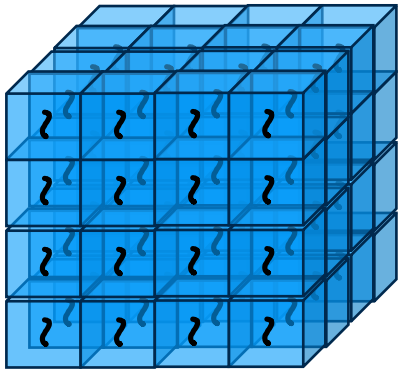
- **Generative AI: Large Language Models and more**
  - Llama2
  - Open access to LLMs for research and commercial use
  - Training and Inference (Prefill and Decode)



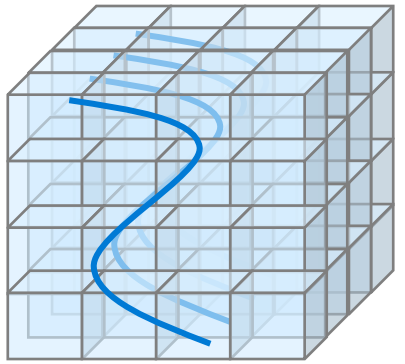
AI Challenging DC Infra

SNIA COMPUTE, MEMORY, AND STORAGE SUMMIT
Solutions, Architectures, and Community
VIRTUAL EVENT, MAY 21-22, 2024

AI needs for DC Infra

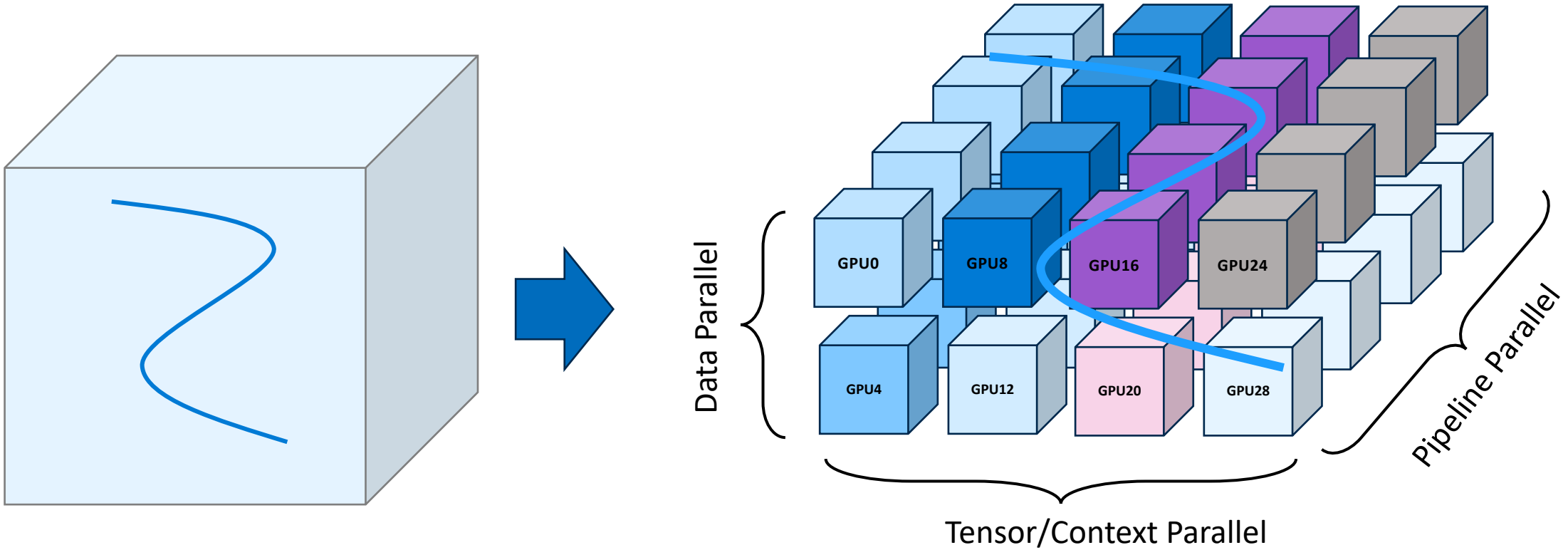


- **CPU-centric Scale-out applications**
- **Millions of small stateless applications**
- Failure handling through redundancy
- Scale performance through large number of nodes

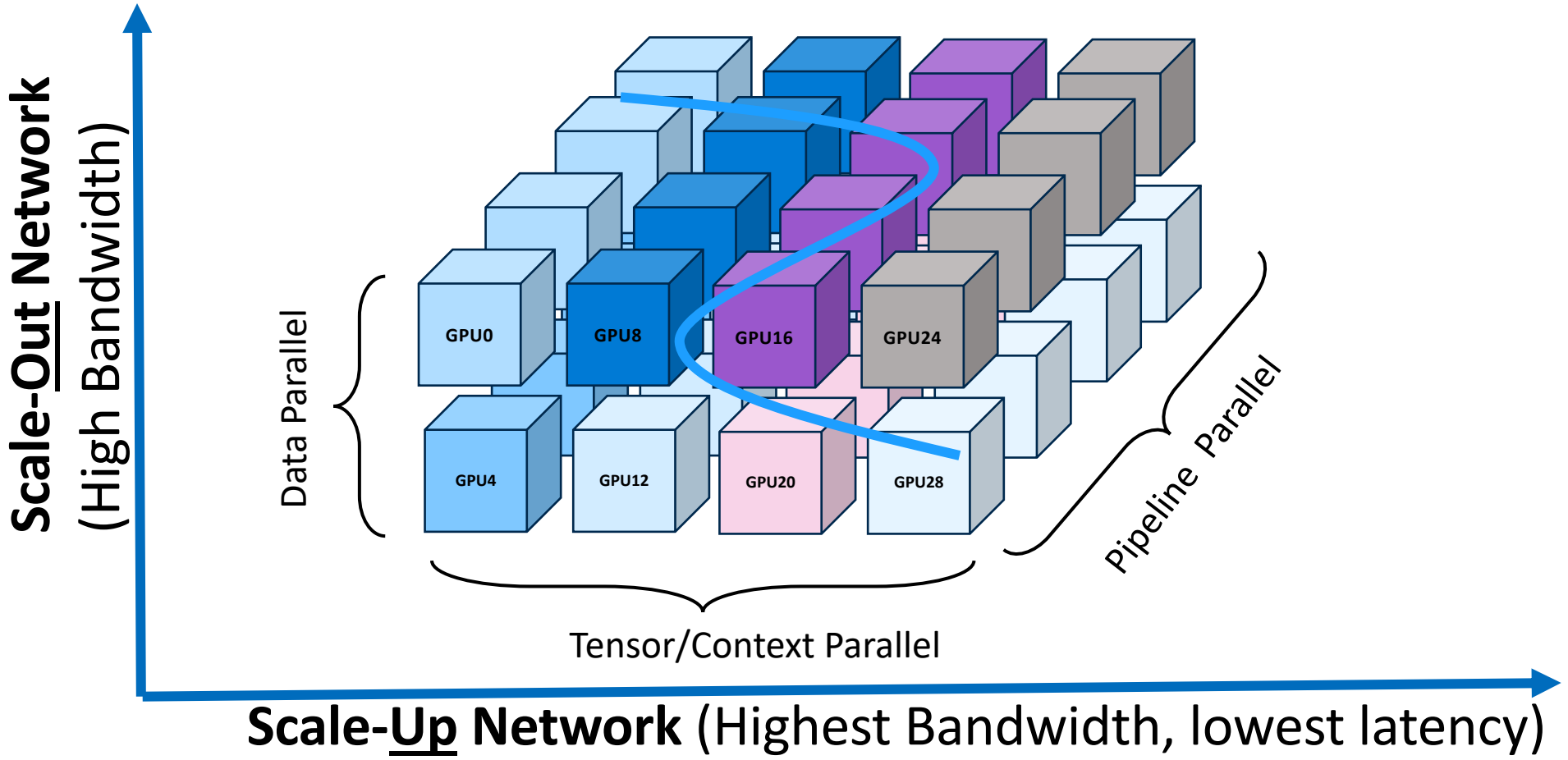


- **Accelerator-centric AI Apps**
- **AI job spread across 1000's of GPUs**
- Failure penalty of large job restart
- Performance scaling depends on all the components in the cluster (GPU/Accel, memory, network..)

AI Jobs: Scaling the performance

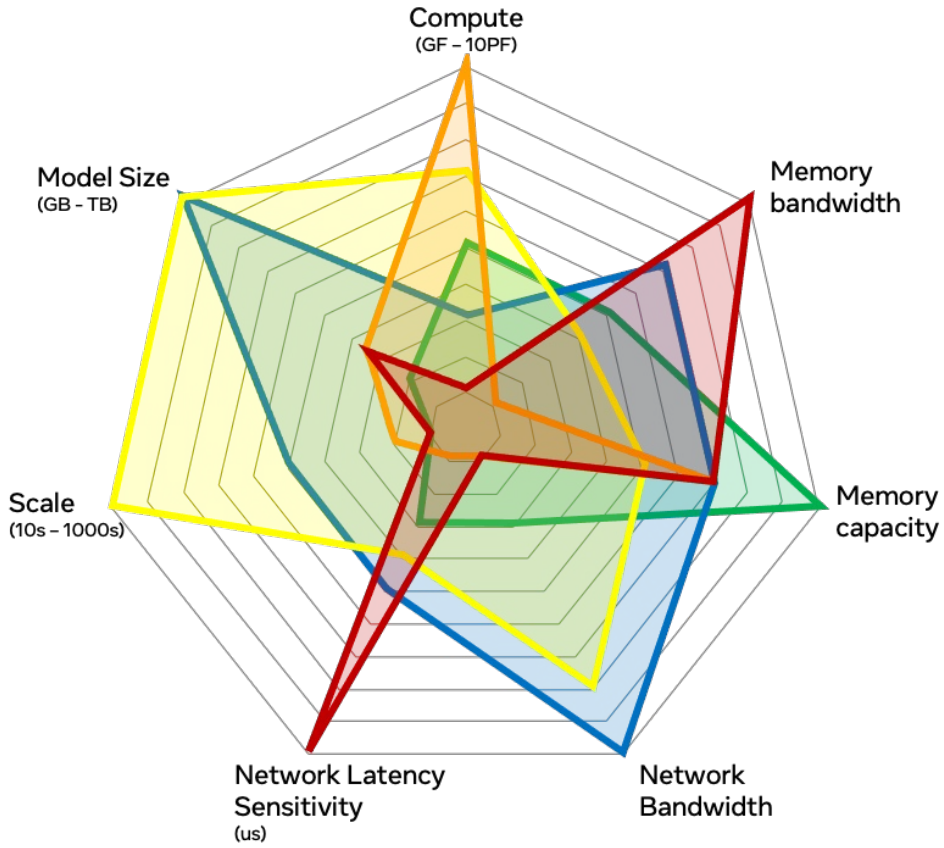


AI Jobs: Scaling the performance



Diversity of AI system requirements

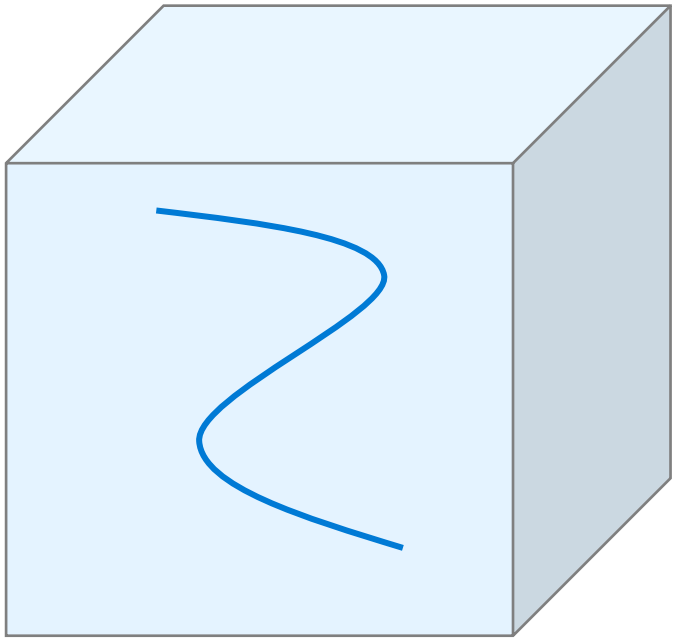
- Difficult to serve all classes of models with a single system design point
- AI use cases are pushing all the design points through software/hardware co-design
- Need for innovation in all the design points:
  - compute, network, memory, packaging, connectivity, cooling..



Memory Requirements for AI

SNIA COMPUTE, MEMORY, AND STORAGE SUMMIT
Solutions, Architectures, and Community
VIRTUAL EVENT, MAY 21-22, 2024

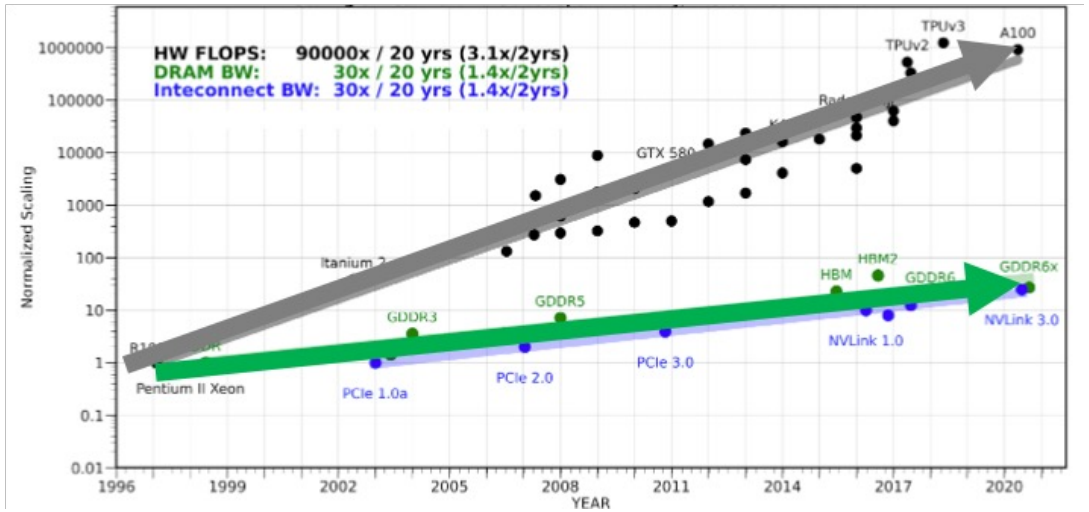
Memory Capacity and Bandwidth



- Accelerators getting larger
  - Memory needs to grow with compute
- Integrated memory innovation
  - To maintain balanced design
  - Provide high reliability
  - Lower power density

But, AI Clusters demand more memory

- Model sizes are increasing, pushing memory capacity demand
- Accelerator performance pushing, Memory bandwidth needs
- Scale-up cluster needs to operate like single accelerator
  - More memory capacity
  - More bandwidth and lower latency to memory



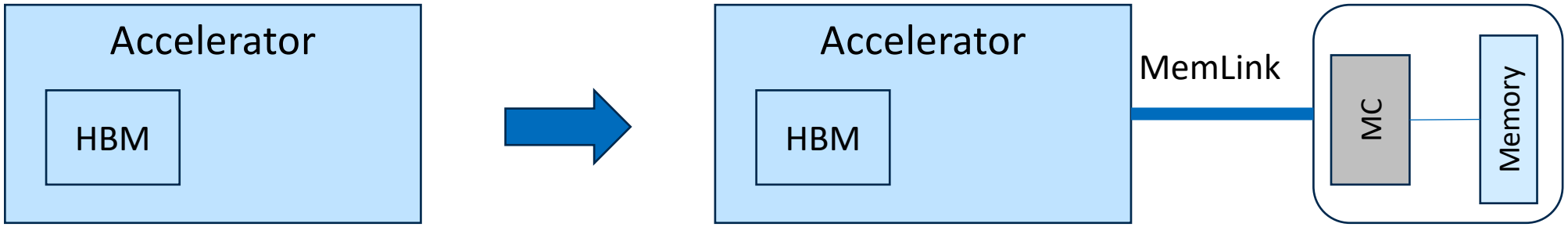
Ref: AI & Memory Wall

Key AI Use Cases for Memory

Activations	• Activations computed during FWD, needed during BWD • Can be offloaded instead of recomputed
Model Params/Embeddings	• Locality of access makes this ideal for offloading • Current layers closer in HBM and rest in higher tier memory
Gradients, Weights, Opt Params	• Offloading weights, gradients and optimizer states to higher tier memory
Other	• In-memory checkpointing to improve reliability[1]

Notes: 1 Google paper

Memory Expansion for Accelerators

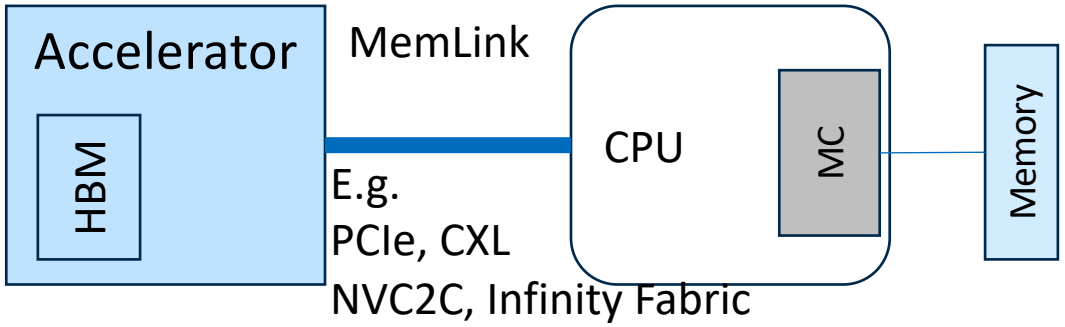


Tier1 memory (HBM) not enough

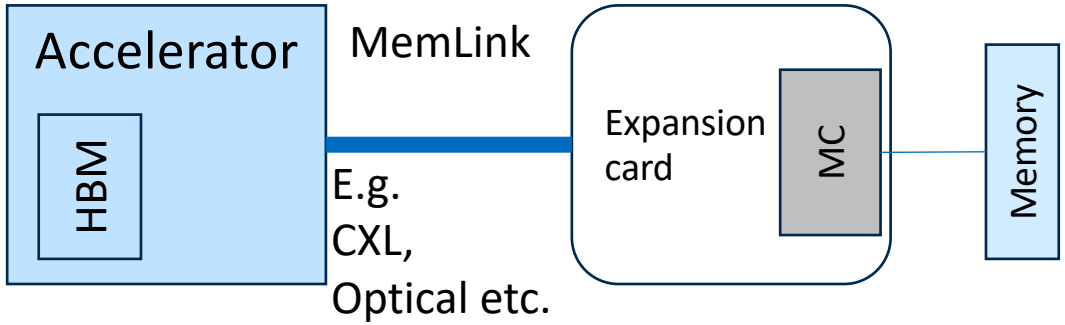
Tier2 Memory for Capacity Expansion

Tiered memory between HBM and external DRAM can provide desired solution

Memory Expansion – Node Native



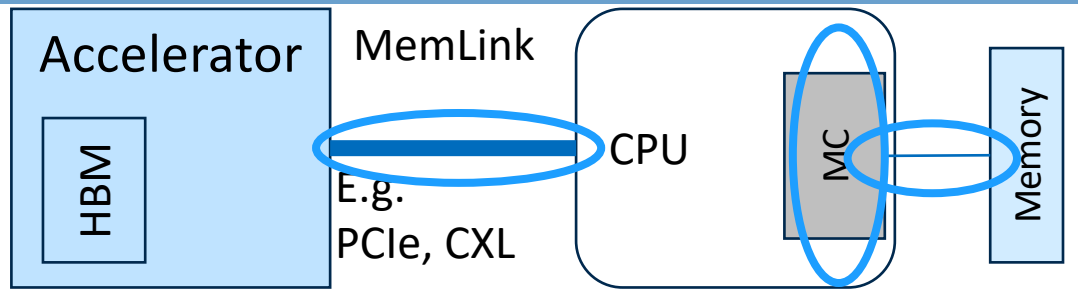
Tier2 Memory through Host CPU's Memory Controller



Tier2 Memory through Expansion card Memory Controller

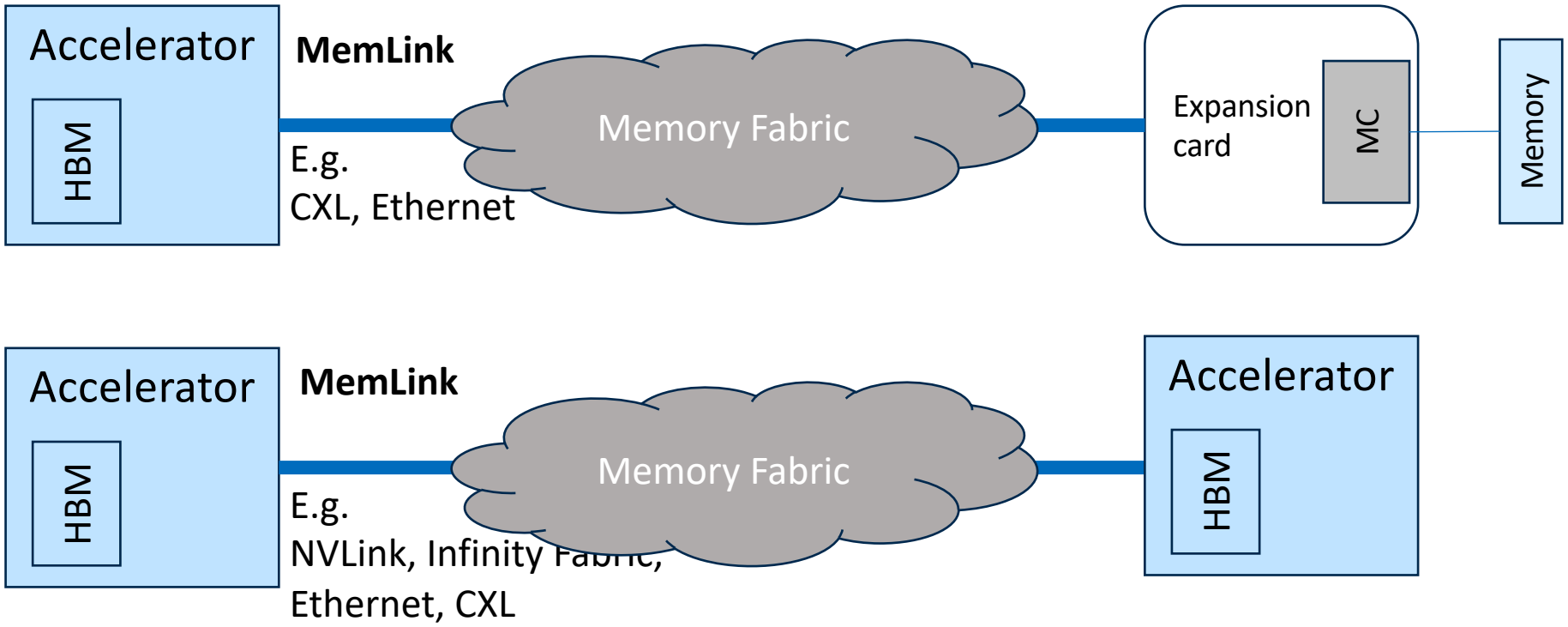
Embedding and Activation offload to reduce accelerator stranding

Node Native MemLink



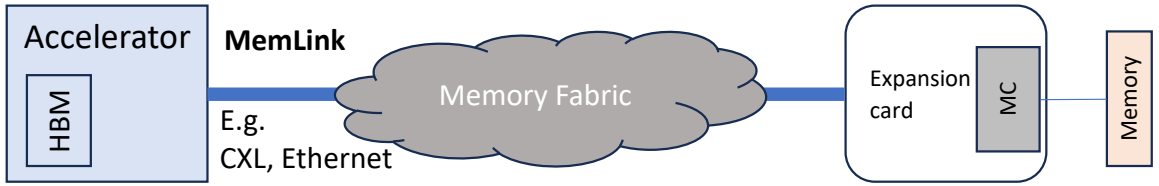
- Interconnect:
  - NV-C2C, CXL can enable higher amount of memory at high BW for accelerators
  - BUT: higher speeds are required to avoid higher number of lanes for CXL
- Memory Controller
  - High bandwidth requirements drive higher number of channels
- Memory Modules
  - Higher speed and capacity requirements to achieve BW and capacity
  - Lower power and higher reliability

Memory Expansion – Fabric Attached



Fabric Attached Memory use cases:
Embedding/Activations offload, Shared KV Cache, in-memory checkpointing

Fabric Attached MemLink



- High speed load/Store Fabrics
  - NvLink
    - Already established over multiple generations
  - CXL
    - Up and coming fabric leveraging existing technologies
    - Need to address speed challenges – significantly behind competitive solutions
  - Infinity Fabric
    - Technology from AMD allowing accelerator to accelerator (and CPU) connectivity
- Software/Hardware co-design to take advantage
  - Tiered Memory, Near-Memory-Compute, In-Memory checkpoints etc.

AI Pushing Boundaries – Call for Action!

Memory Technology

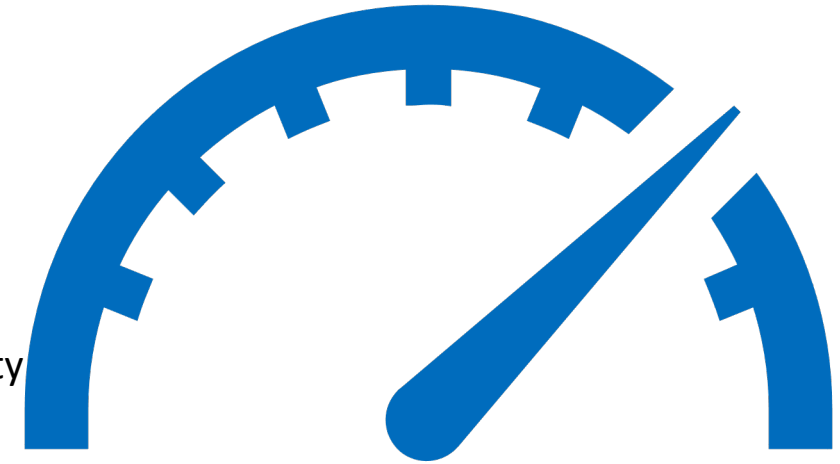
Higher performance, capacity, reliability

Systems

Architectures, SW-HW codesign

Interconnects

Speeds, Radix, Connectivity



Thank You

Please take a moment to rate this session.

Your feedback is important to us.