AI for All:
Pushing Infra Boundaries

Presenter:

Manoj Wadekar, Meta

Hyung Kim, Meta

FMS the Future of Memory and Storage

JEDEC

Meta AI is used for diverse cases

Large language models (LLMs)

Text-to-image generation

AI-enabled creation tools

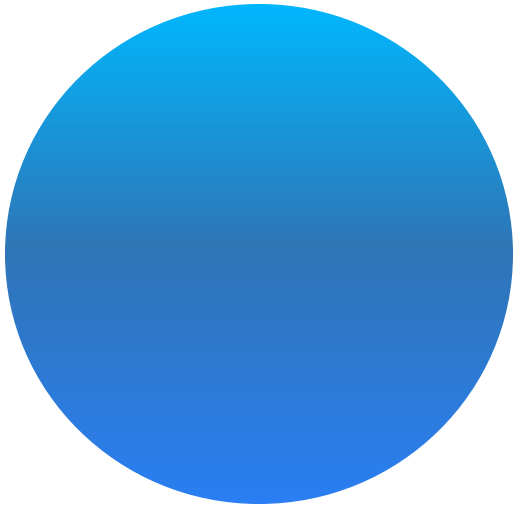
+173 %

Conversation topic growth on Instagram

Source: Meta for Business. 'Culture Rising: 2023 Trends Report.' 2023.

GenAI runs on Large Languages Models

Llama-2 65B
Circa: 2023



Total Compute (PF/s)	400
Memory Capacity (TB)	10
Training Scale (GPUs)	4k

…towards Multi-Modality

Llama-2
Circa: 2023

Text 1x Tokens

Llama-3
Circa: 2024

Text 7-8x Tokens

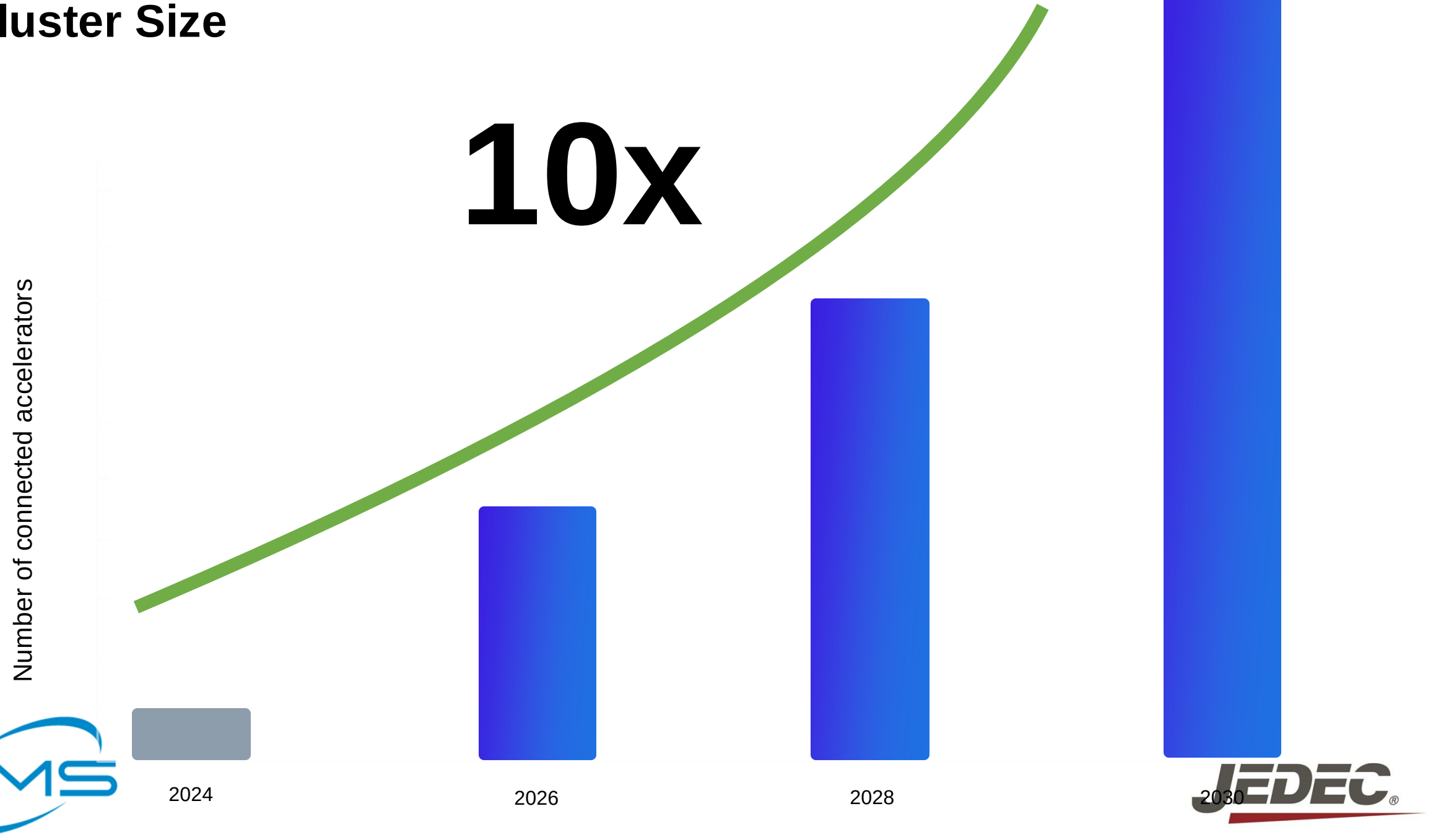
Llama-Next
Circa: 202x

Videos

Images

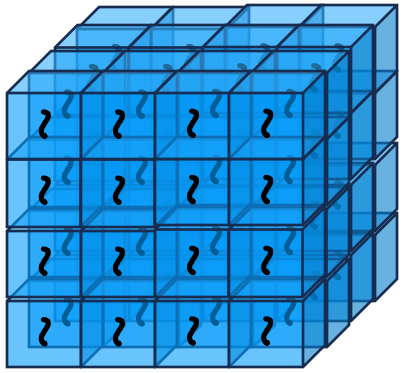
Audio

AI Cluster Size

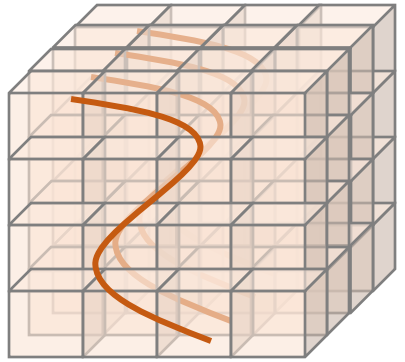


AI Challenging DC Infra

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



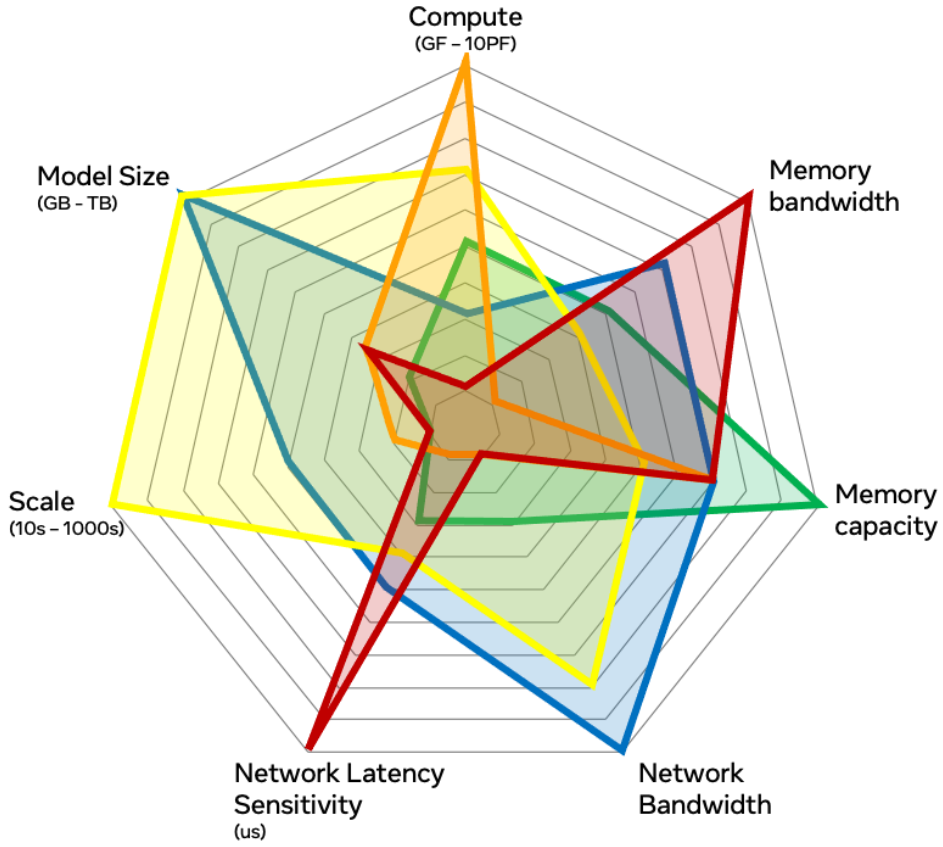
- **Accelerator-centric AI Apps**
- **AI job spread across 1000's of GPUs**
- Failure penalty of large job restart
- Performance scaling depends on all the components in the cluster (GPU/Accel, memory, network..)

FMS

JEDEC

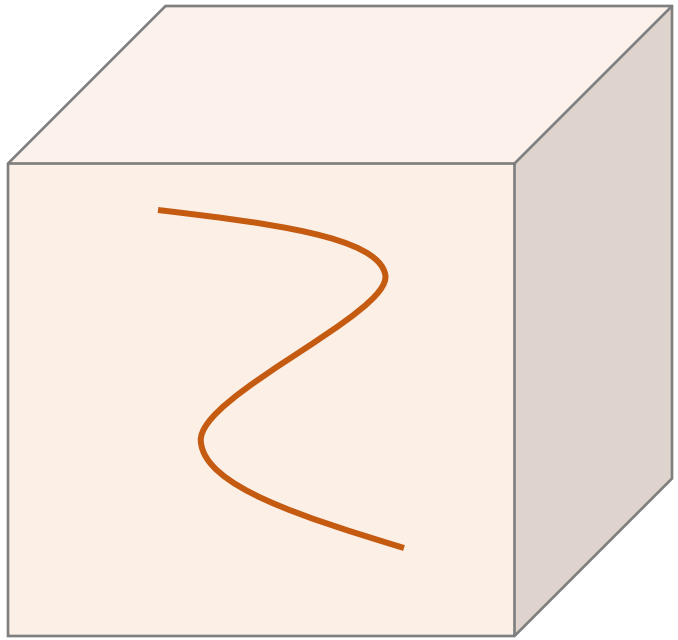
Diversity of AI system requirements

- 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..



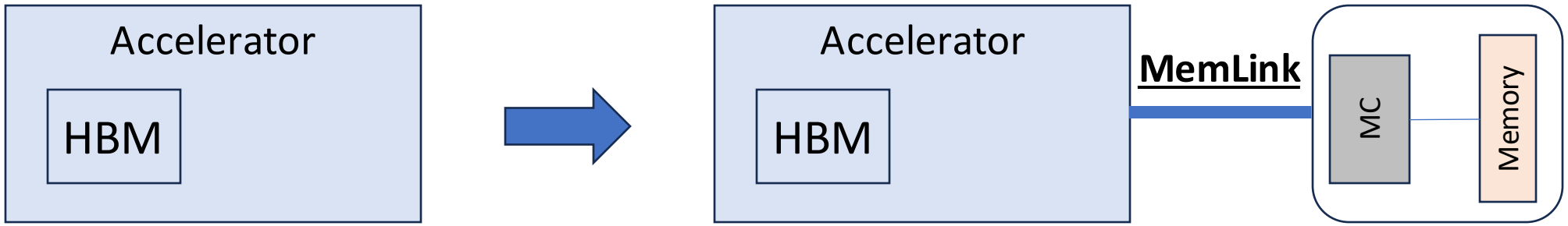
Memory Requirements for AI

Memory Capacity and Bandwidth



- Accelerators and Models getting larger
  - Memory needs to grow with compute
  - Model sizes are increasing, pushing memory capacity demand
- Integrated memory innovation
  - To maintain balanced design
  - Provide high reliability
  - Lower power density
- Tiered Memory for accelerators

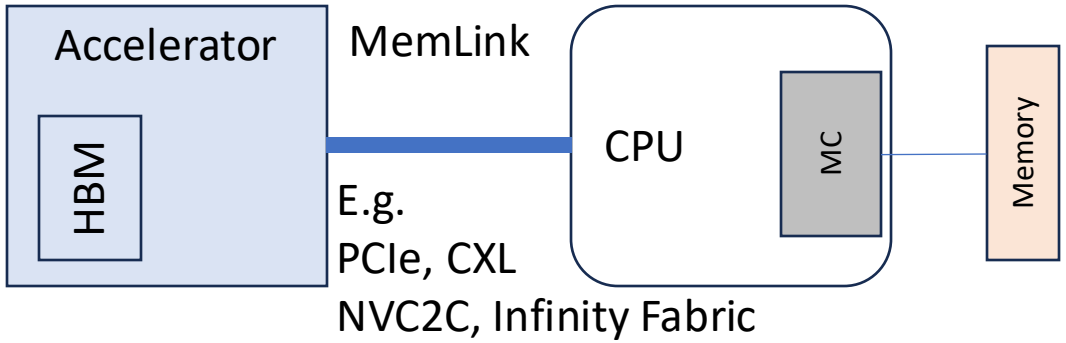
Memory Expansion for Accelerators

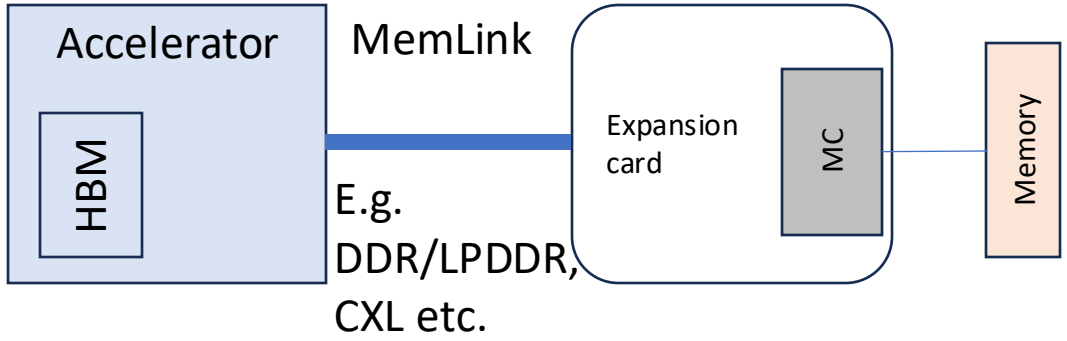


Tier1 memory (HBM) not enough Tier2 Memory for Capacity Expansion

Tiered memory between HBM and external DRAM can provide desired solution

Memory Expansion – Node Native



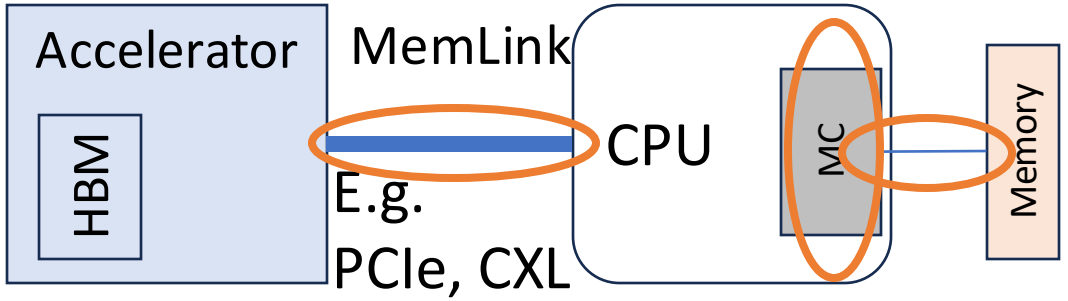


Tier2 Memory through Host CPU's Memory Controller

Tier2 Memory through Expansion card Memory Controller

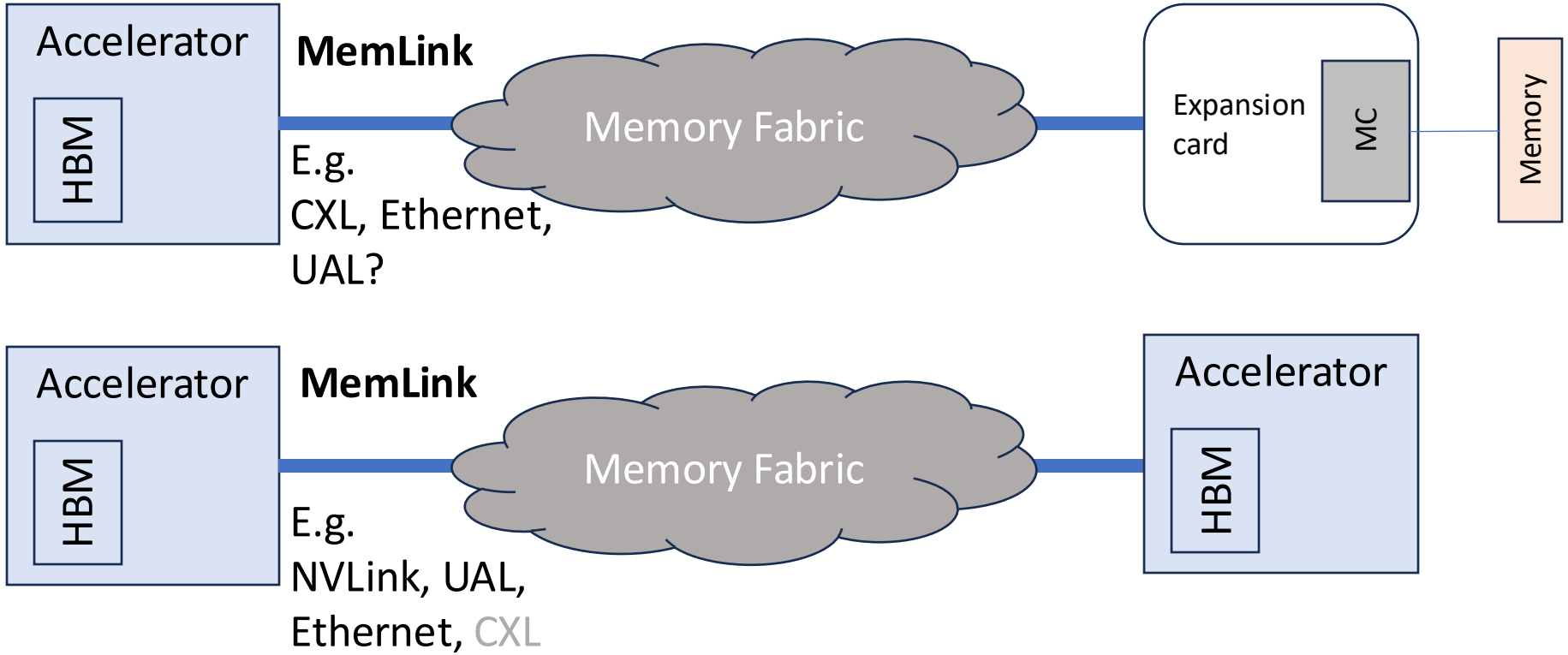
Embedding and Activation offload to reduce accelerator stranding

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

AI Pushing Boundaries – Call for Action!

Memory Technology

Higher performance, capacity, reliability

Systems

Architectures, SW-HW codesign

Interconnects

Speeds, Radix, Connectivity

Thank You