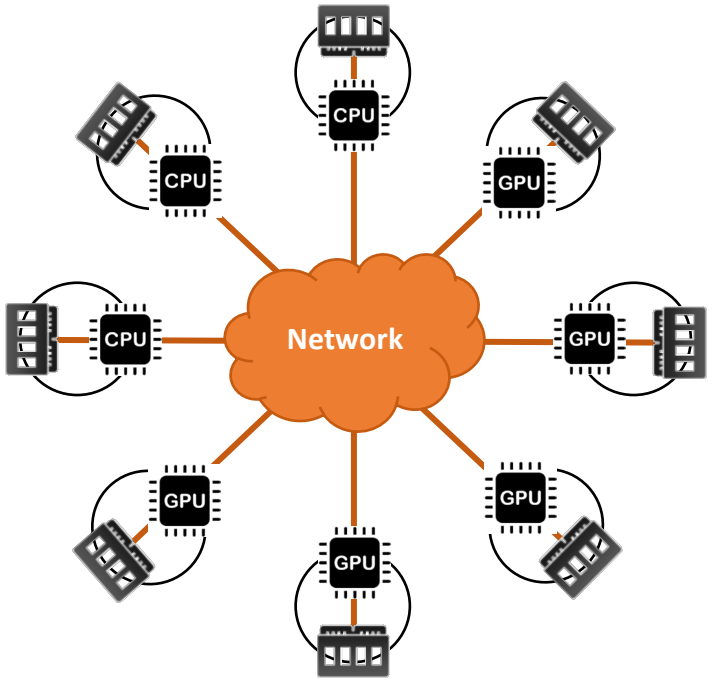
Memory Centric AI Machine

(LLM Serving using Memory Pool)

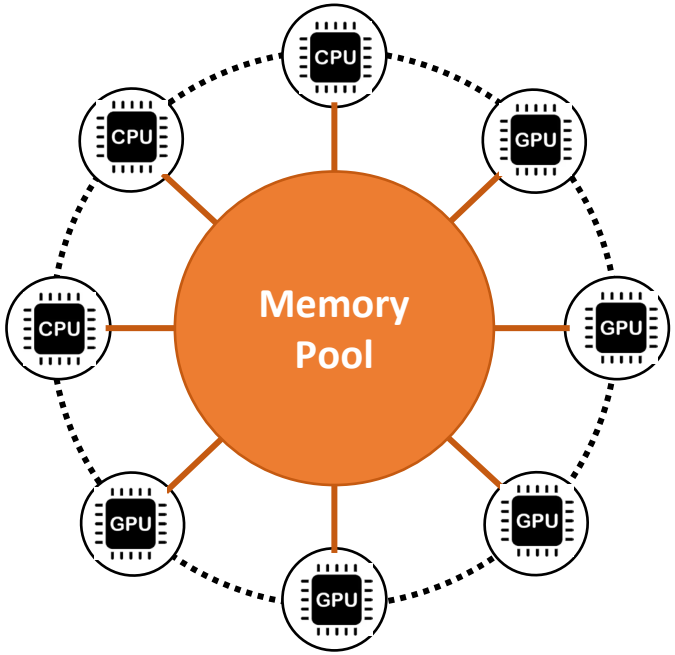
Jongryool Kim

SK hynix

Memory Connects ALL



[Distributed AI System]



[Memory Centric AI Machine]

Benefit of Communication using Memory Pool

Low latency data communication

Reducing memory usage and the number of memory copies

cMPI: Using CXL Memory Sharing for MPI One-Sided and Two-Sided Inter-Node Communications (@ SC 2025)

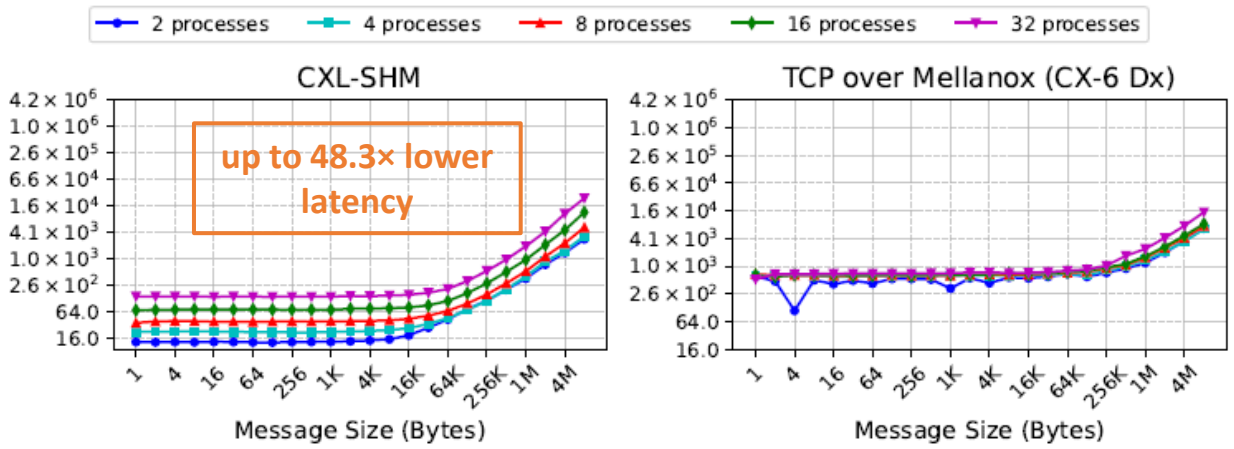
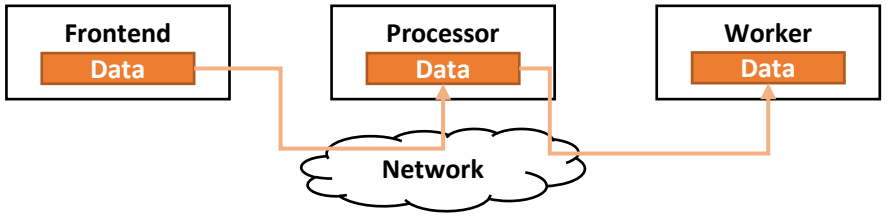
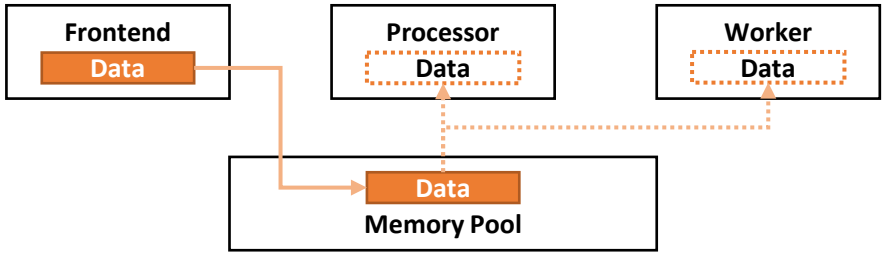


Figure 6: Latency of one-sided MPI communication.



[Data Communication-by-value using Network]

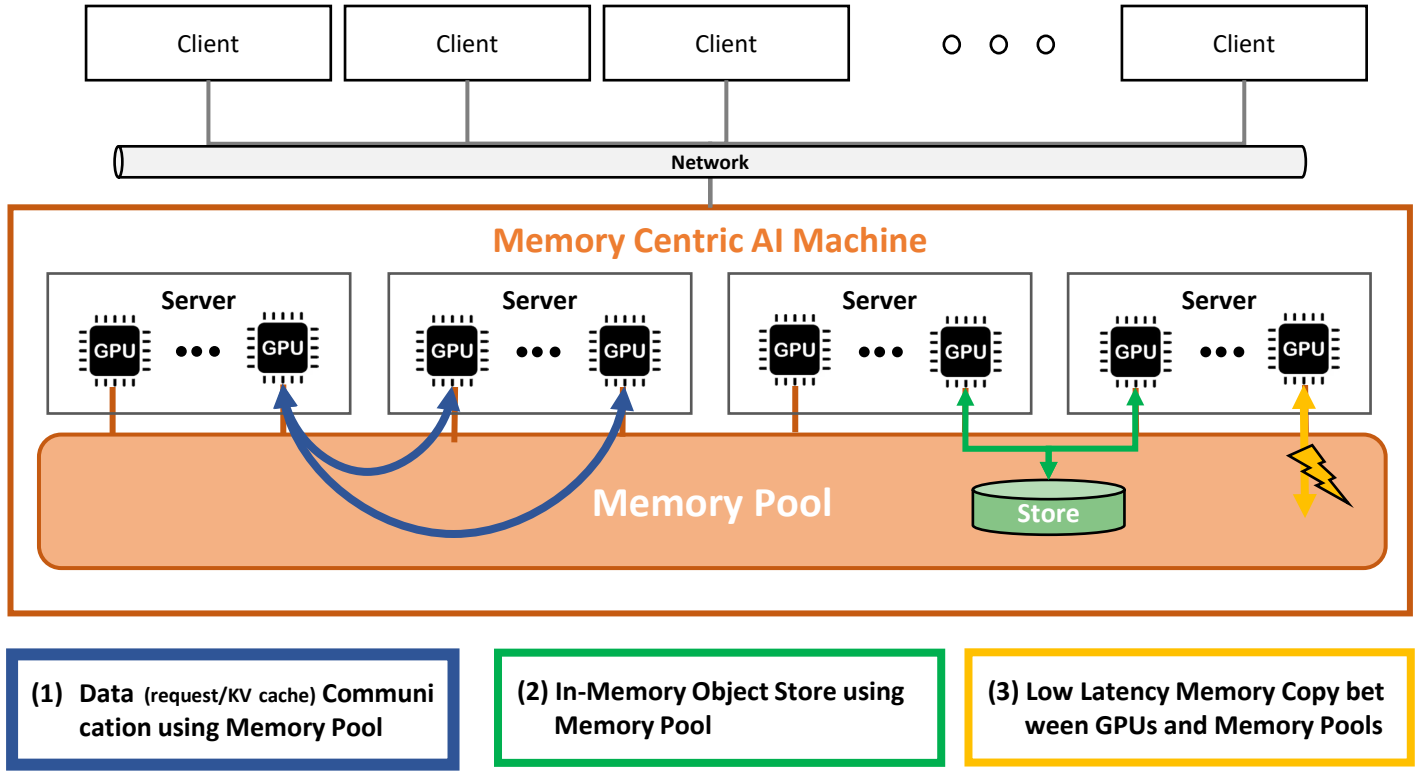


[Data Communication-by-reference using Memory Pool]

Memory Centric AI Machine

Memory Centric AI Machine Platform

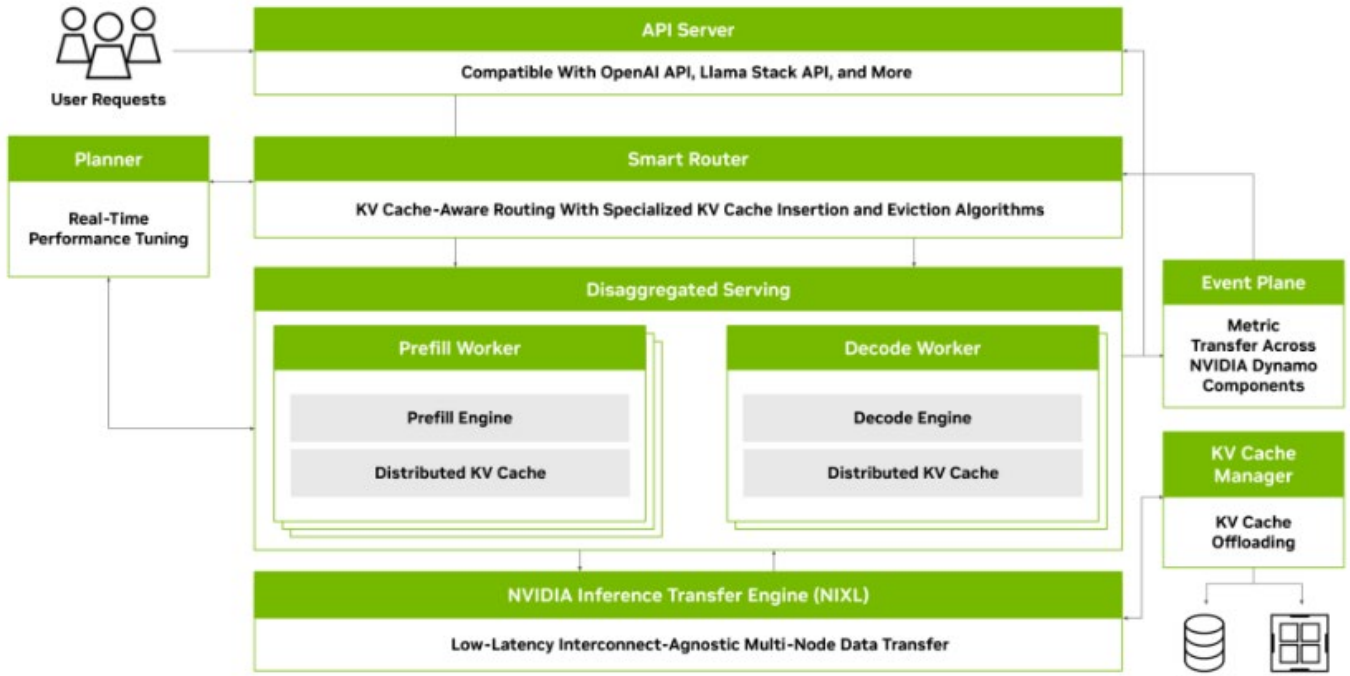
- Used for building and scaling distributed systems, especially those involving AI and HPC



Dynamo LLM Serving

Low latency distributed inference framework for scaling reasoning AI models

- NVIDIA announced Dynamo at GTC 2025
- Open Source, High throughput and Low latency, Scalability, Distributed Inference, Modular Design



https://developer.nvidia.com/blog/introducing-nvidia-dynamo-a-low-latency-distributed-inference-framework-for-scaling-reasoning-ai-models/

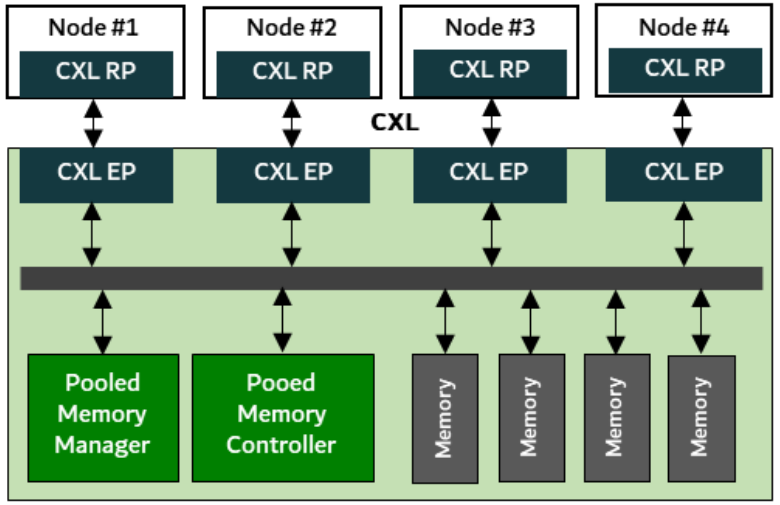
CXL Pooled Memory : Niagara

Built a Niagara HW/SW research platform, an FPGA-based CXL disaggregated memory prototype

- 2U memory appliance which can connect up to 8 CXL host servers (without CXL switch)
- Supports up to 4 channels of DDR4-DIMM (1TB)
- Supports DCD (Dynamic Capacity Device) and HMU (Hotness Monitoring Unit) feature in CXL spec. 3.x

CXL Interface	CXL 2.0, Gen4x8
	Up to 8-port
Memory	4CH DDR4 DIMM
	Up to 1 TB
Functionality	Dynamic Capacity Device
	Hotness Monitoring Unit

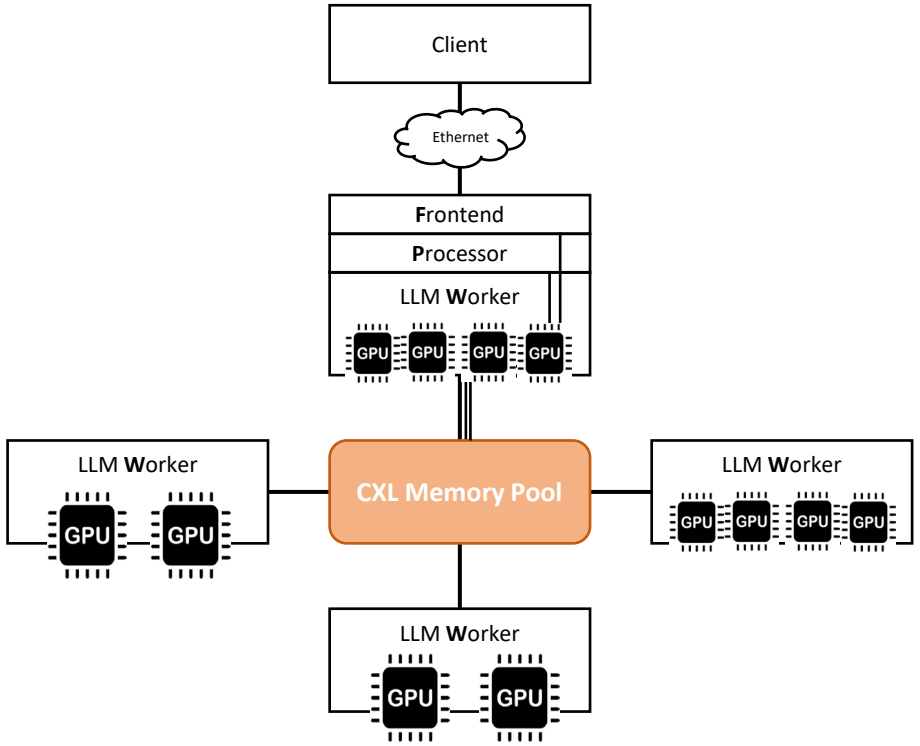
[Niagara Specification]



[Niagara HW/SW Research Platform]

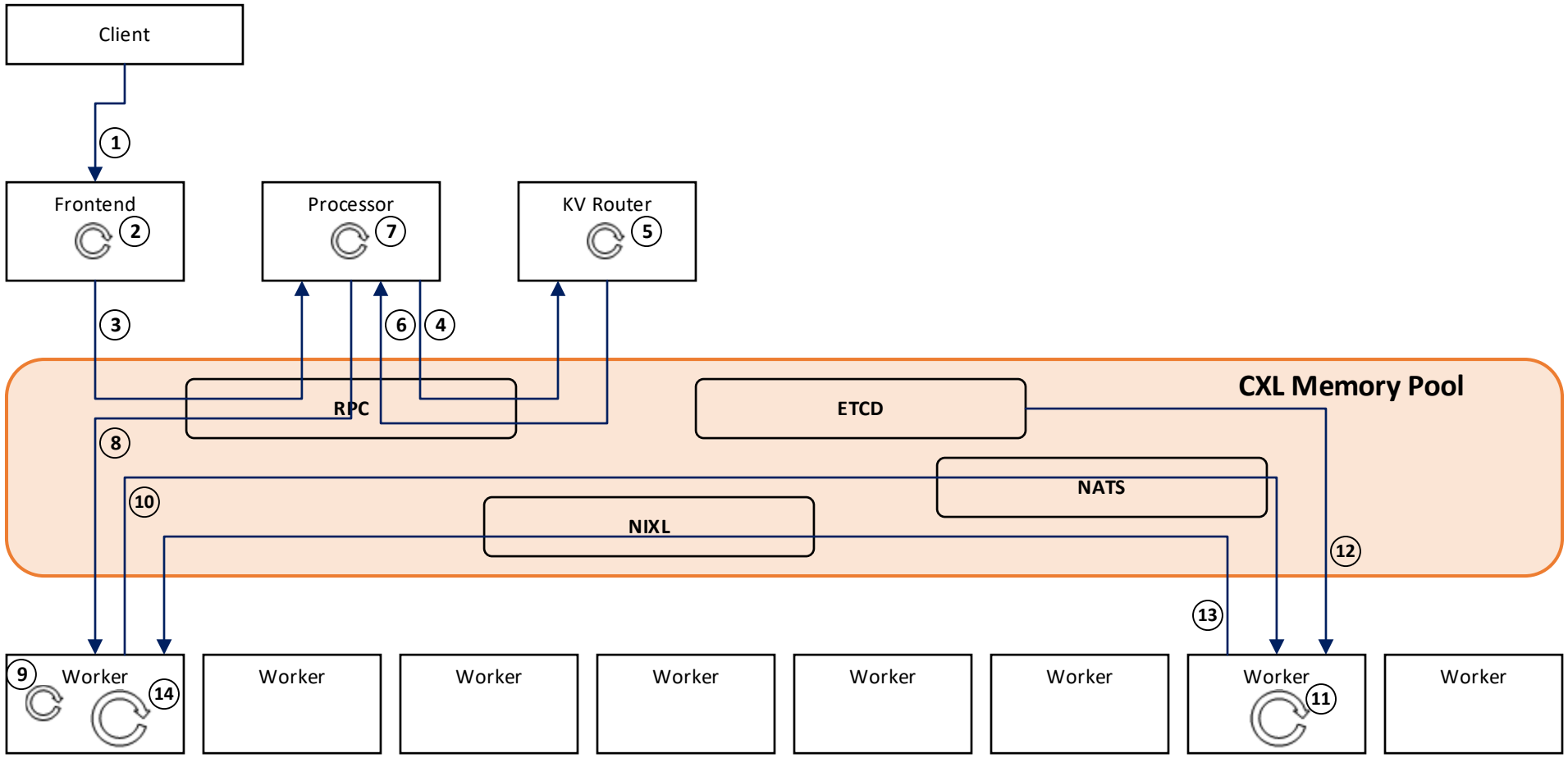
[Rack-Scale System with Niagara]

Dynamo LLM Serving with CXL Pooled Memory



[Dynamo LLM Serving System with CXL Memory Pool]

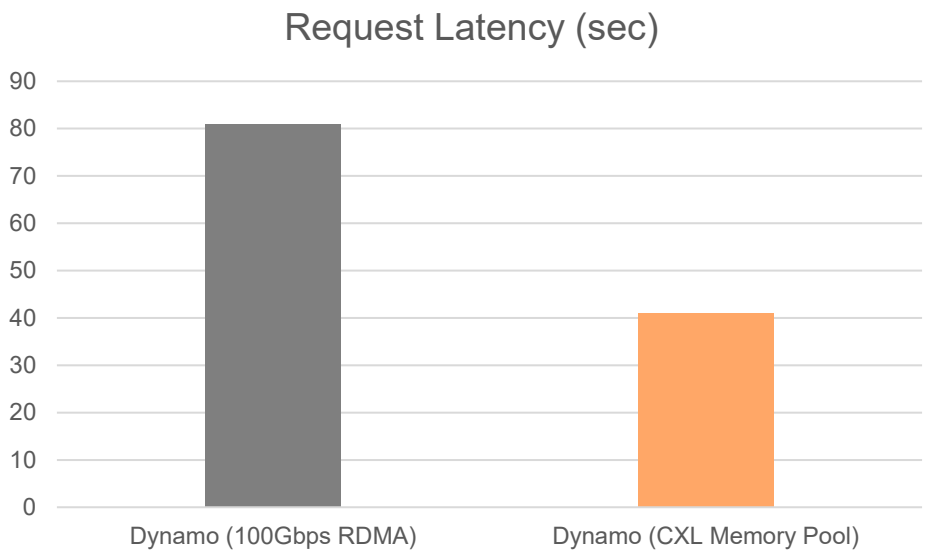
Dynamo LLM Serving with CXL Pooled Memory

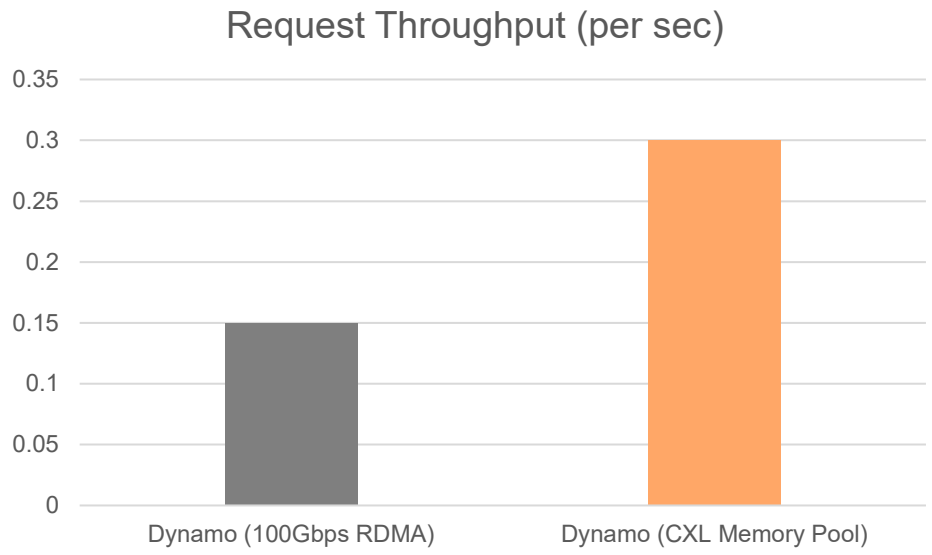


LLM Serving Performance

LLM Serving using CXL Pooled Memory can improve the LLM serving performance

- DeepSeek-R1-Distill-Llama-8B, 15000 input tokens, 150 output tokens

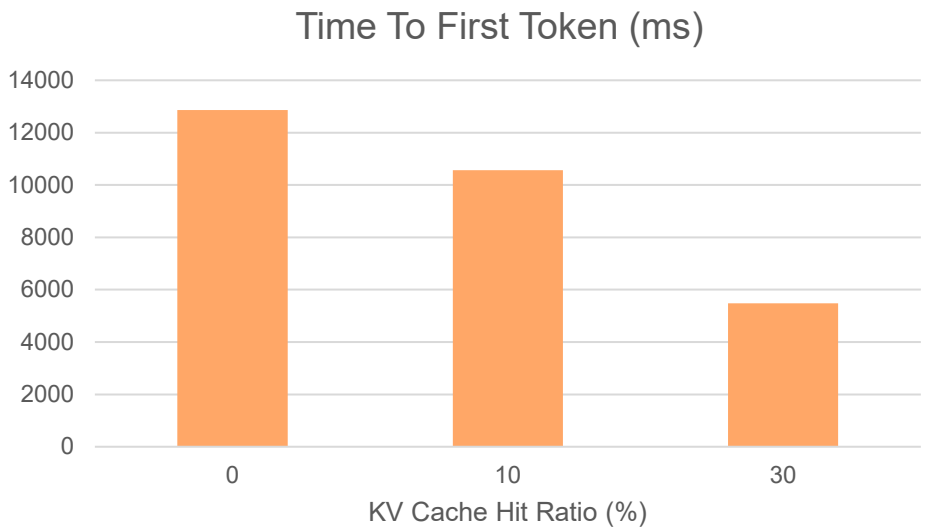


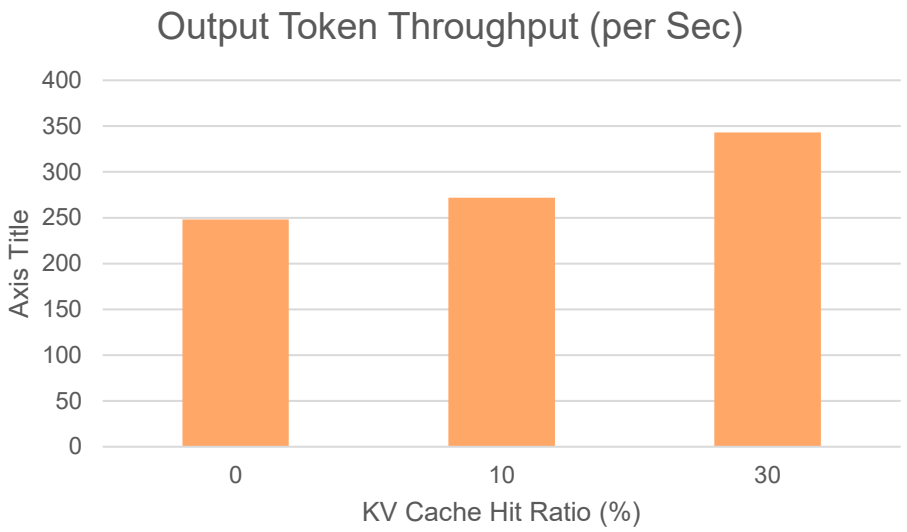


Memory Storing and Reusing

Communication using memory requires storing data to memory and loading data from memory

- Reusing data from memory → Context / Prefix Cache





What's Next ?

Memory Storing and Reusing

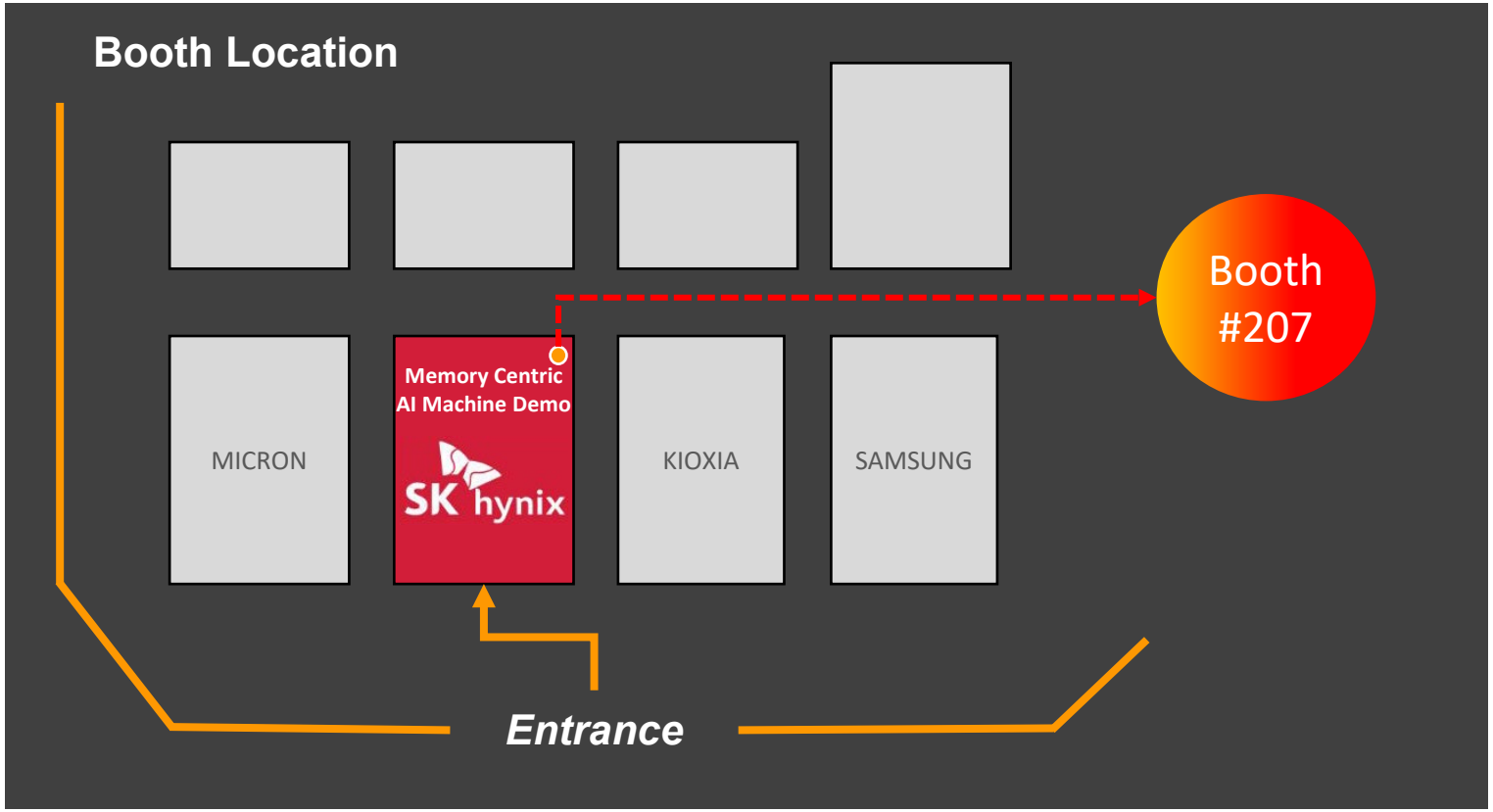
- Prefix / Context cache
- Multiple LLM models loading - S-Lora

New systems using Memory Centric AI Machine

- Fine-tuning LLM Training System
- GNN Training System
- GPU based Quantum Simulation

Scalable Memory Centric AI Machine (to be revealed soon)

Learn more about SK hynix



Visit Booth #207 and Experience SK hynix products and demos

Booth #207

Meet the future of memory.
Just steps from the entrance.

Innovation starts here,
Literally.

SK hynix