

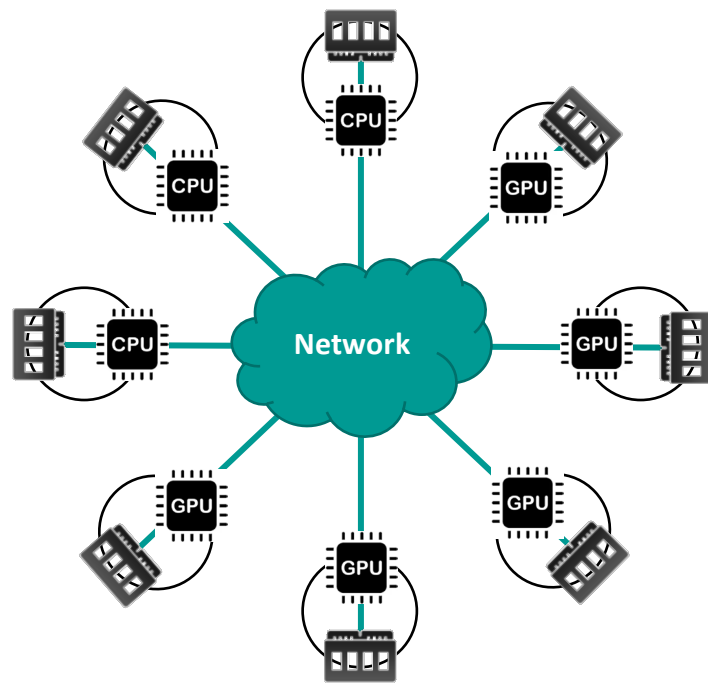
Memory Centric AI Machine

(Dynamo LLM Serving with Memory Pool)

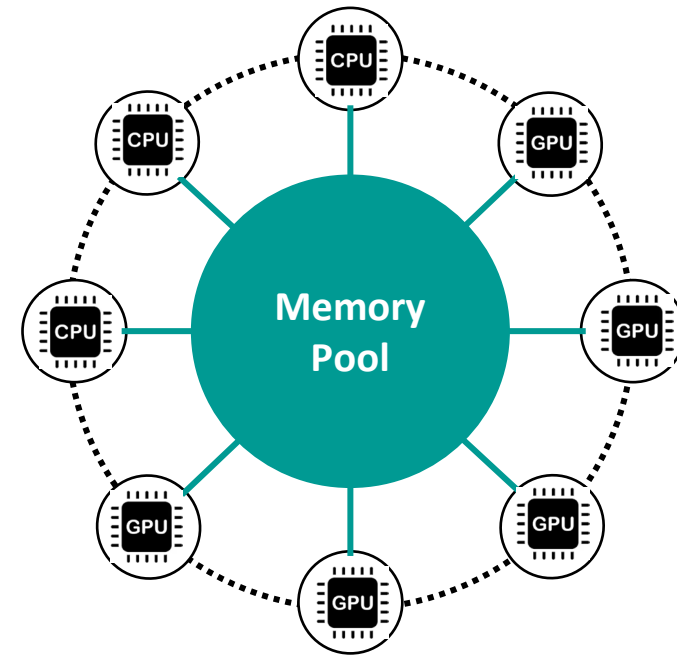
Jongryool Kim



Memory Centric Computing

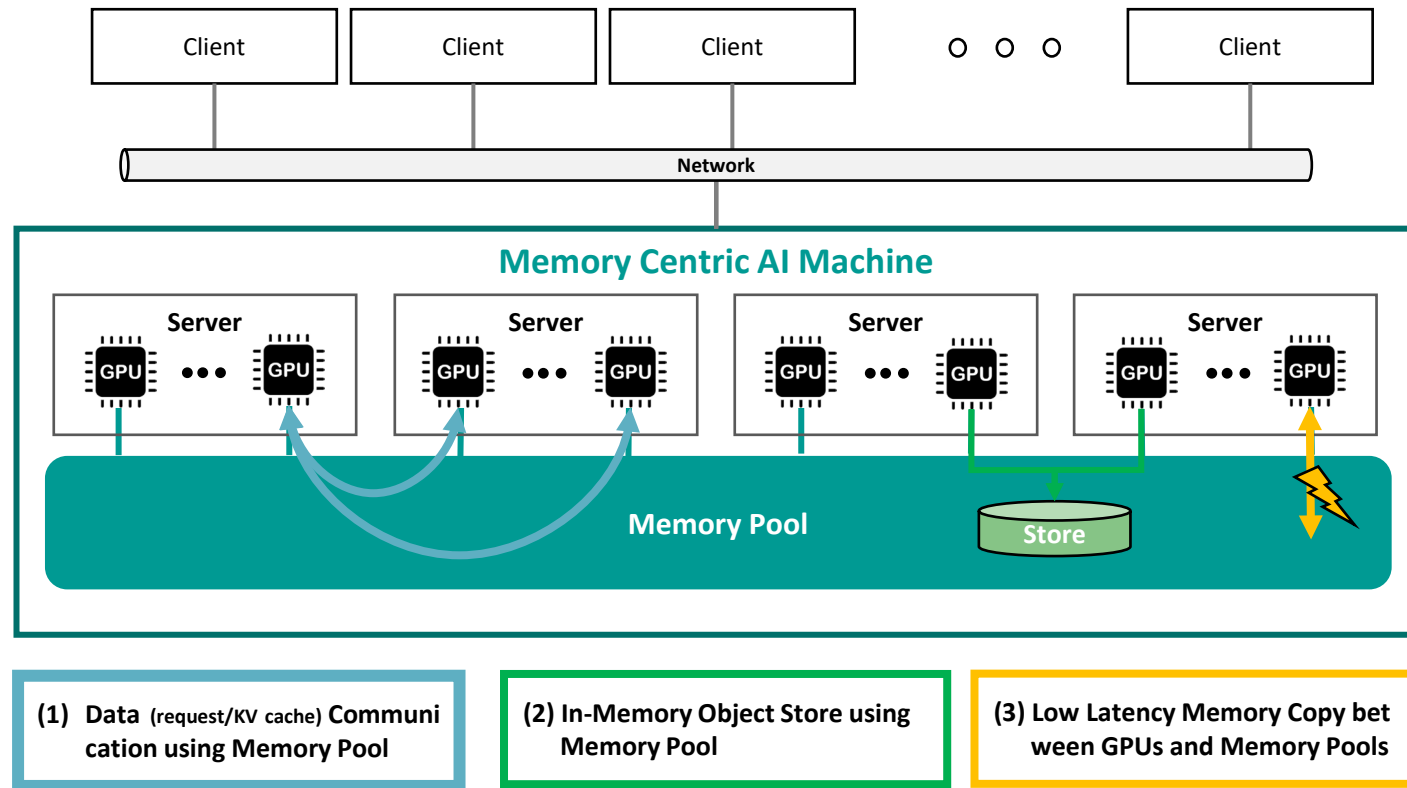


[Distributed AI System]



[Memory Centric AI Machine]

Memory Centric AI Machine



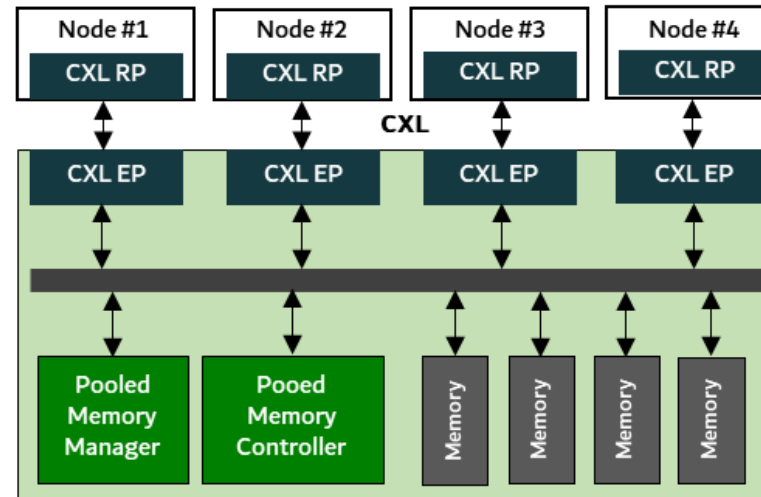
[Memory Centric AI Machine]

CXL Disaggregated Memory

- 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 defined 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]

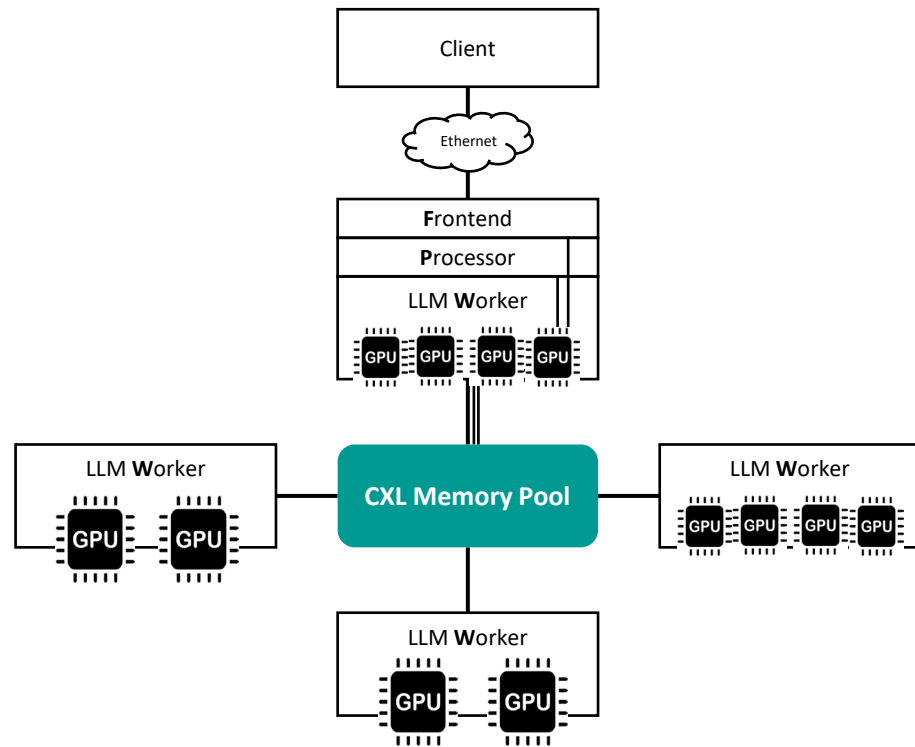


[Niagara HW/SW Research Platform]



[Rack-Scale System with Niagara]

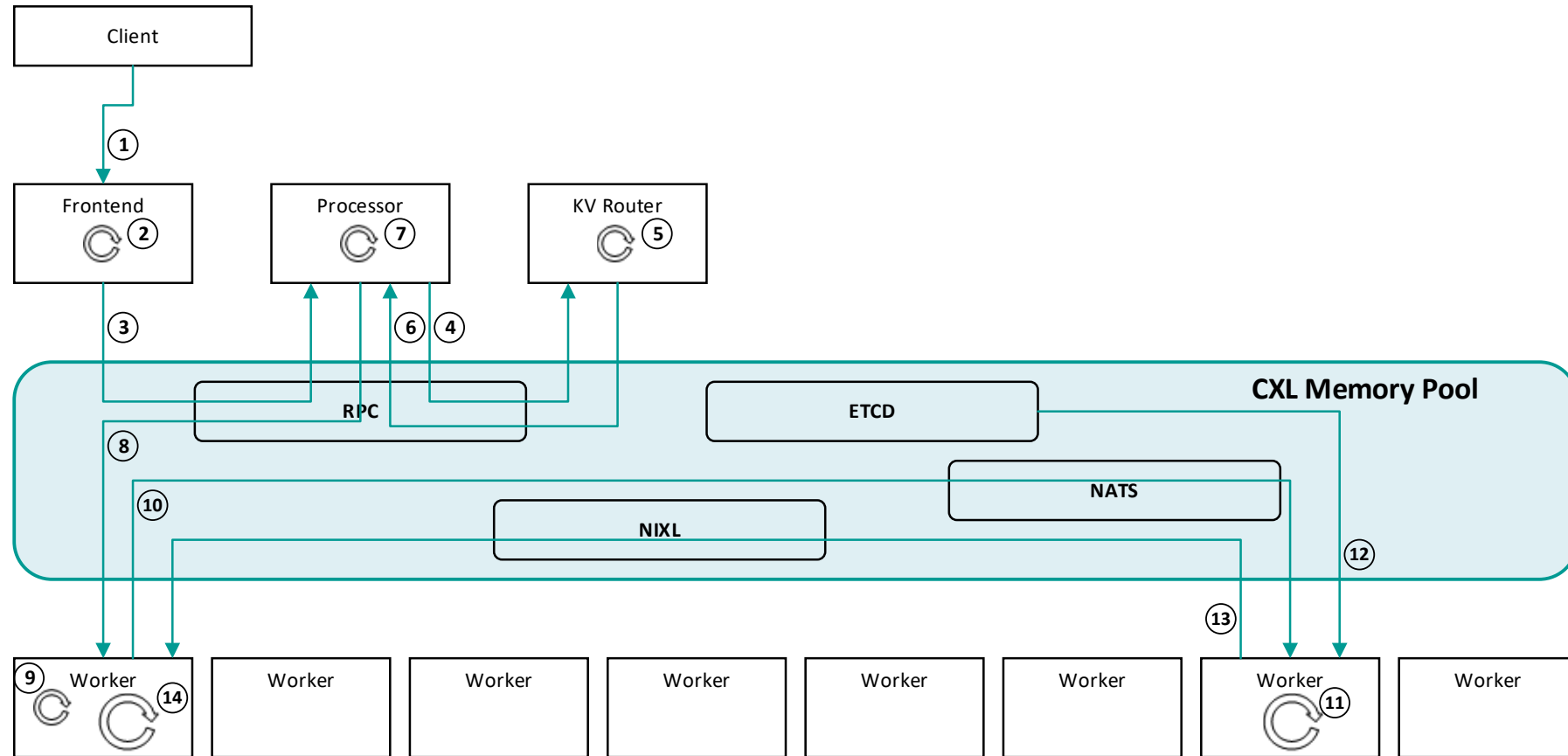
Dynamo LLM Serving with Niagara 2.0



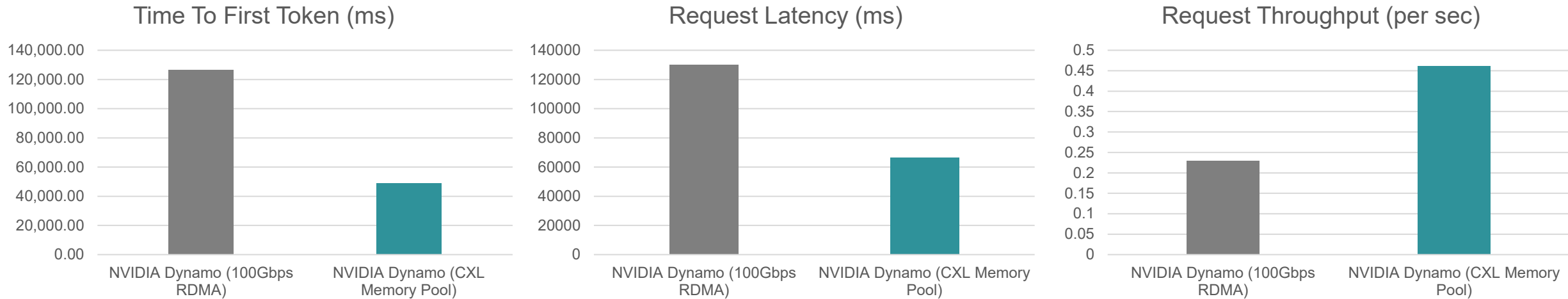
[NVIDIA Dynamo LLM Serving System with CXL Memory Pool]



LLM Serving with Dynamo and Niagara 2.0



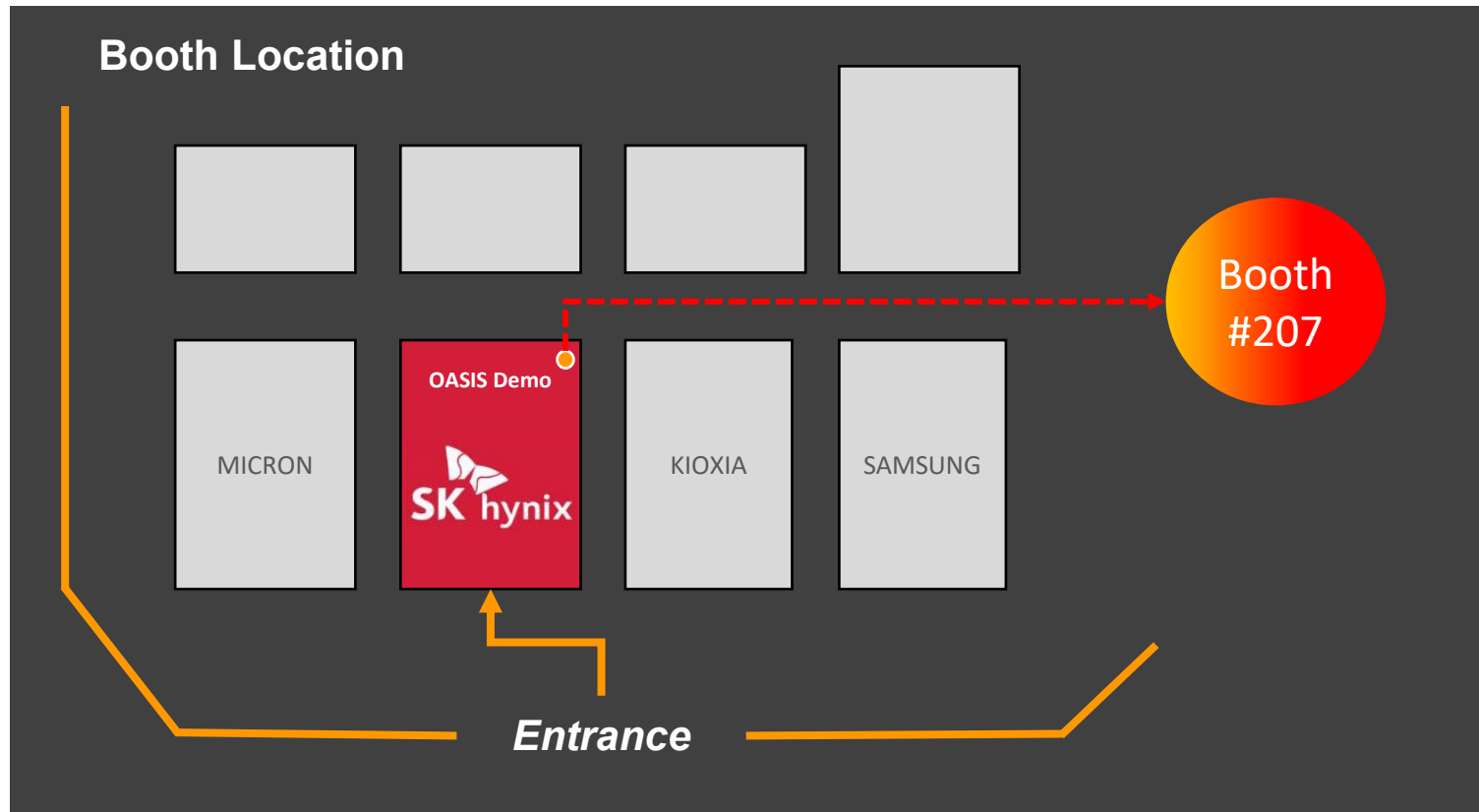
Performance Improvement



What's Next ?

- Prefix cache
- S-Lora
- GNN Training System with Memory Centric AI Machine

Learn more about SK hynix



Visit Booth #207 and Experience SK hynix products and demos