Exploring CXL Memory Disaggregation: Use Cases and System Benefits

MEMORY FABRIC FORUM

OCP GLOBAL SUMMIT

OCT 15-17, 2024
SAN JOSE, CA

Exploring CXL Memory Disaggregation: Use Cases and System Benefits

Jungmin Choi

Memory System Architect

SK hynix

Agenda

□ Motivation

- Growing Memory Bandwidth and Capacity Gap
- Challenges in Today's Datacenter

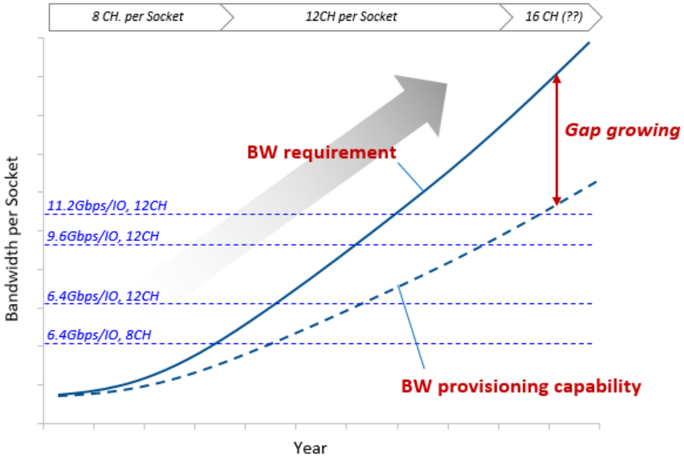
□ Solution

- Niagara: CXL Disaggregated Memory Prototype
- Use Cases of CXL Disaggregated Memory

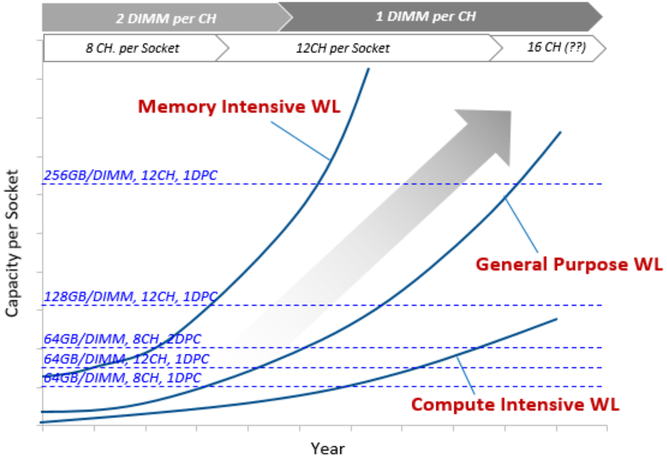
□ Future Work

Growing Memory Bandwidth and Capacity Gap

- Increase in core counts requires continued increase in memory bandwidth & capacity
- The gap between such requirements and platform provisioning capability is growing
- CXL creates new opportunities beyond physical limitations, and efficient memory disaggregation is possible



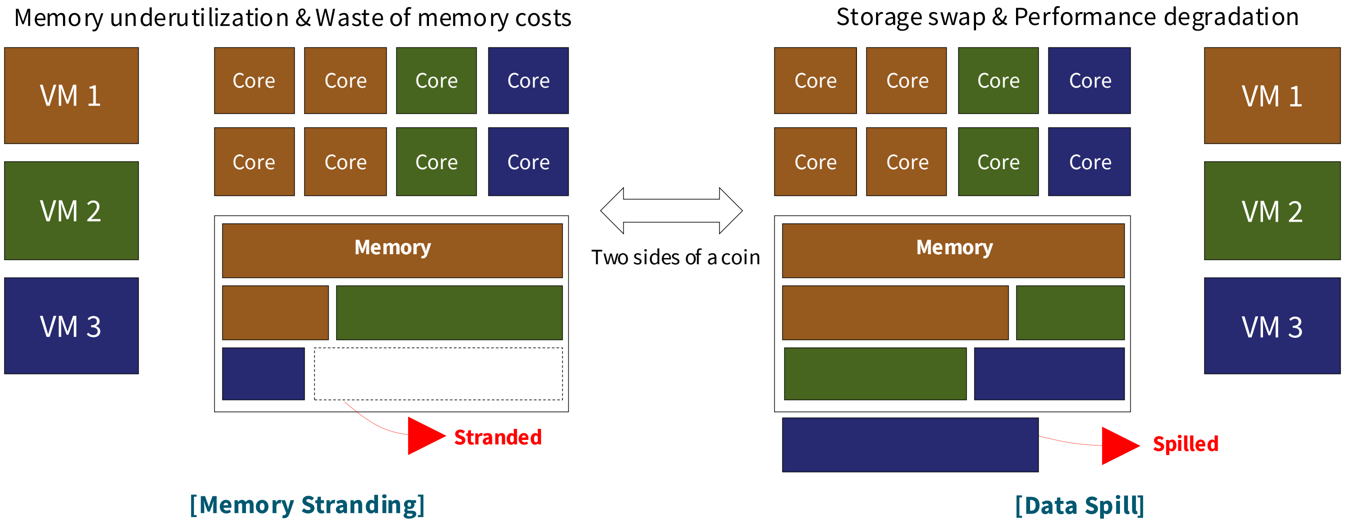
[Memory Bandwidth Requirement]



[Memory Capacity Requirement]

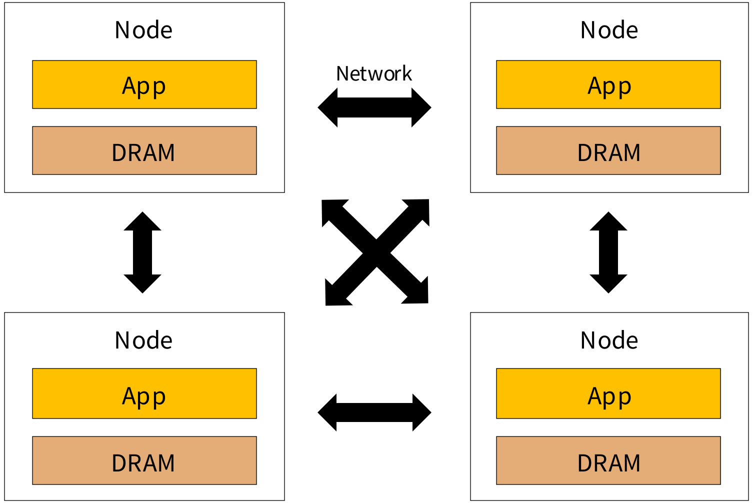
Challenges in Today's Datacenter

- Challenge 1 : Memory stranding & data spill
  - The memory utilization of each node in a compute cluster varies time to time
  - Unused memory in each node can never be utilized by other nodes, which causes memory stranding and data spill



Challenges in Today’s Datacenter

- Challenge 2 : Data transfer overhead & data duplication
  - In a distributed computing system, there is a network-based data transfer overhead between remote nodes
  - Duplication of shared data between nodes increases local memory pressure

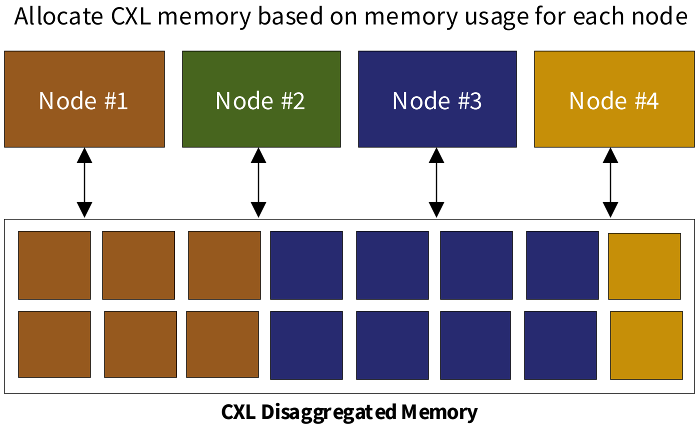


Serialization & Deserialization

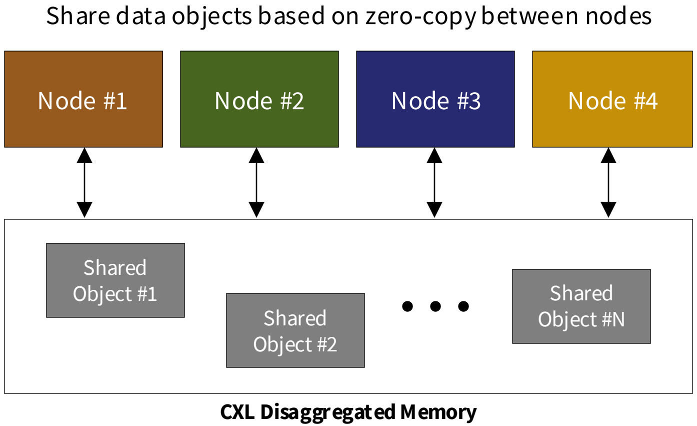
Data Duplication

CXL Disaggregated Memory System

- CXL disaggregated memory system can support memory pooling & sharing
  - Memory pooling : Mitigate memory stranding and data spill by sharing memory resources between nodes
  - Memory sharing : Remove data transfer overhead and data duplication by sharing data between nodes



[Memory Pooling]



[Memory Sharing]

Solution to Overcome the CXL Drawback

- A main drawback of CXL-based memory is the additional latency
- Hotness tracking can be used to address the additional latency of CXL memory
- Current implementation limitations [1]
  - PEBS (Processor Event-Based Sampling) can only track 49% of main memory traffic
  - Running the tracking algorithm consumes the CPU cycles
- Hotness tracking inside CXL memory
  - Hotness tracking is a technique used to monitor frequently accessed regions on a remote memory
  - It provides information to the application or OS
  - The application or OS is responsible for page migration

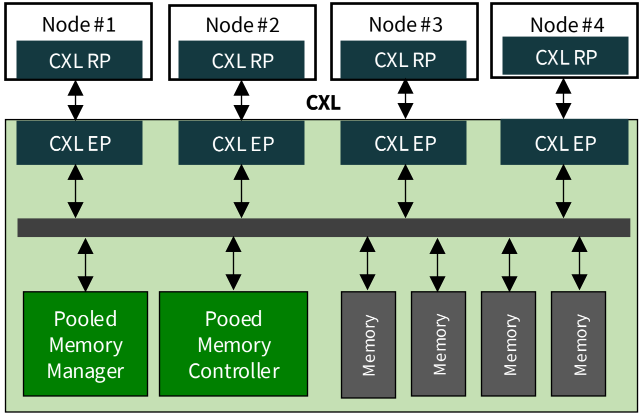
[1] OCP CMS Hotness Tracking Requirements White Paper

CXL Disaggregated Memory Research Platform

- 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 specification 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]

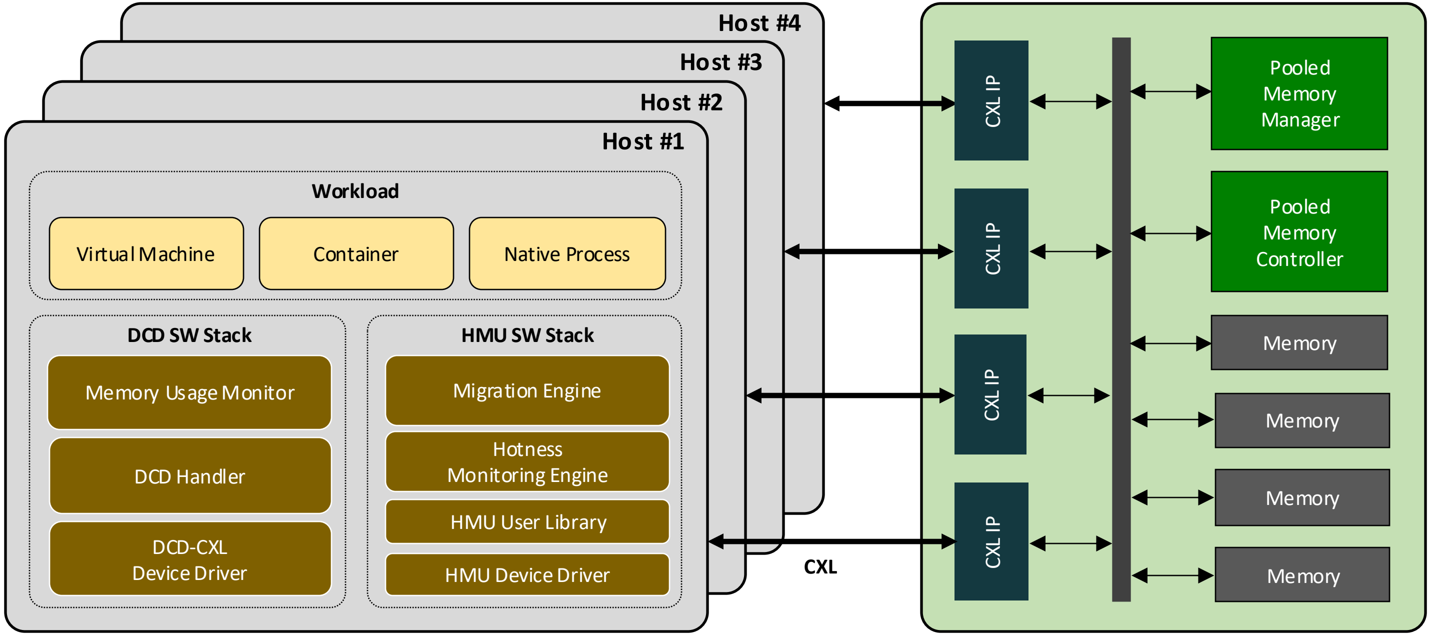


[Niagara HW/SW Research Platform]

[Rack-Scale System with Niagara]

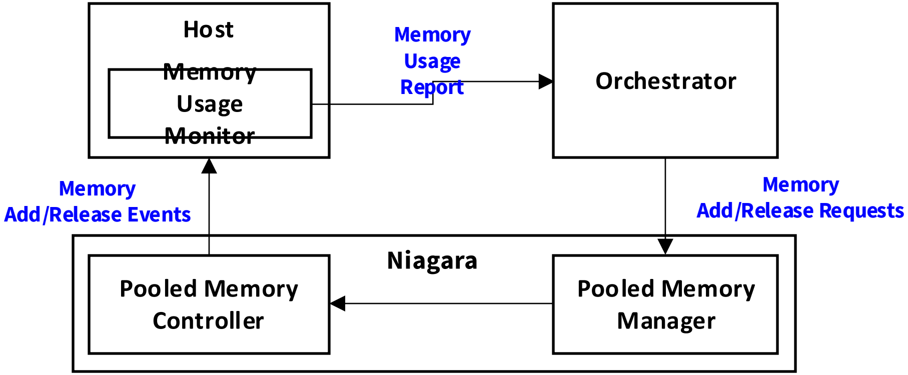
Niagara DCD & HMU Architecture

- Niagara provides HW/SW integrated solution for DCD and HMU



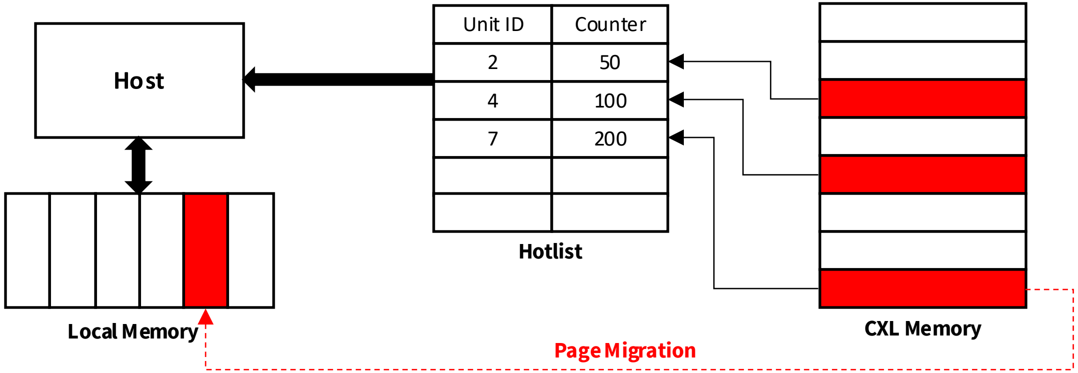
DCD Mechanism

- Memory Usage Monitor traces memory usage of workloads on the host server and reports it to Orchestrator
- Based on the memory usage statistics, Orchestrator issues memory add/release requests to PMM (Pooled Memory Manager)
- PMC (Pooled Memory Controller) sends events to the host after dynamically memory allocation and deallocation



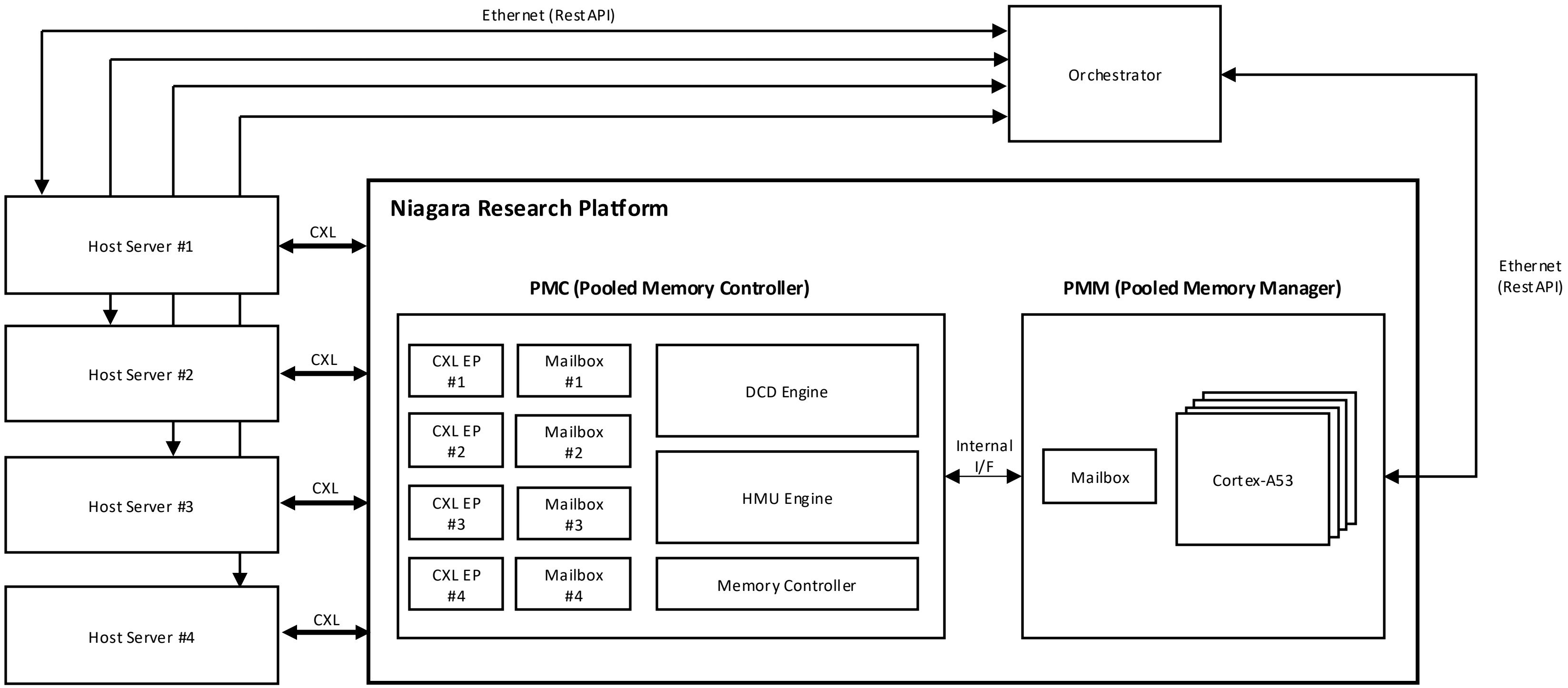
HMU Mechanism

- Host can configure the HMU to obtain optimal hotness information
  - Tracking address range, Unit size (hotness counting granularity), Epoch length, Hotness threshold, Etc.
- Host can read the hotlist of CXL memory region
  - If the hotness counter exceeds threshold, it is registered in the hotlist
  - Host can decide to page migration using the hotlist



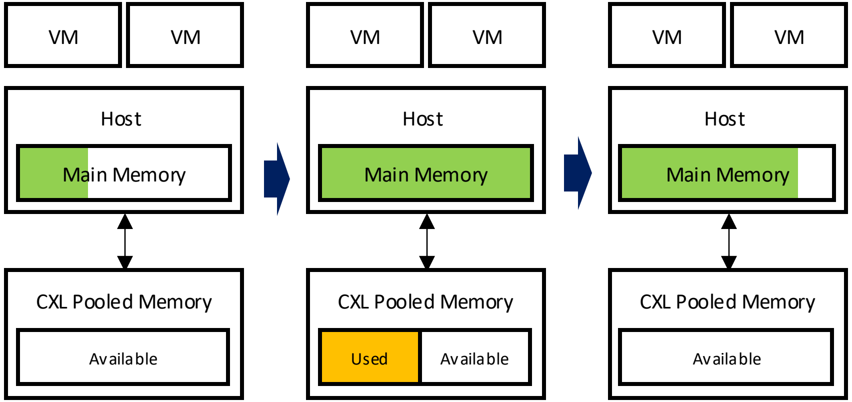
Unit ID	Counter
2	50
4	100
7	200

DCD/HMU-Enabled Infrastructure



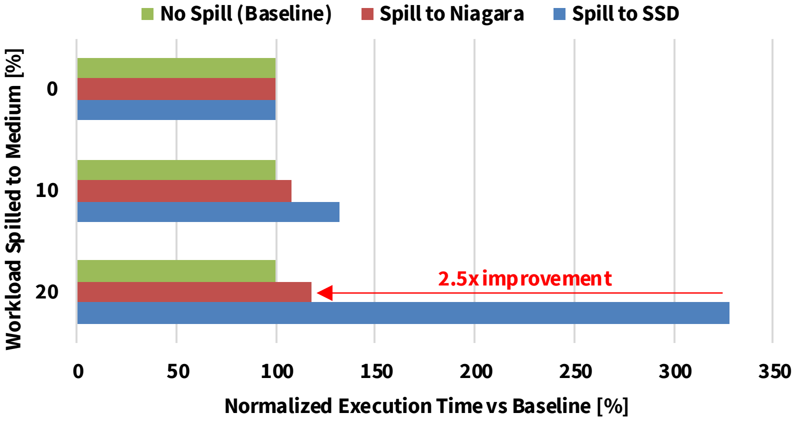
Use Case and System Benefit - DCD

- Memory Pooling (Collaborate with MemVerge)
  - Dynamically allocate/deallocate disaggregated memory resources for each node without RESET
  - Improve memory utilization and performance of a system equipped with CXL disaggregated memory



[Memory Pooling without Workload Interruption]

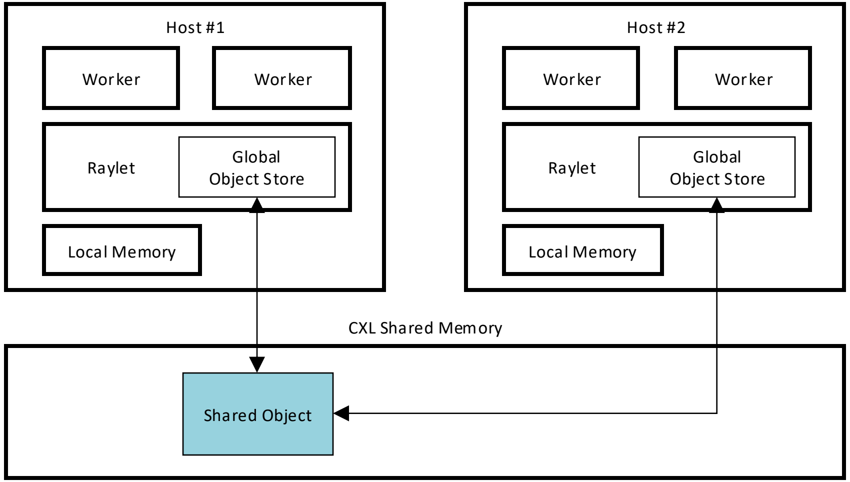
Spill to Niagara outperforms NVMe by up to 2.5x
(even though Niagara is an FPGA-based prototype)



[Execution Time of CloudSuite In-Memory Analytics Benchmark]

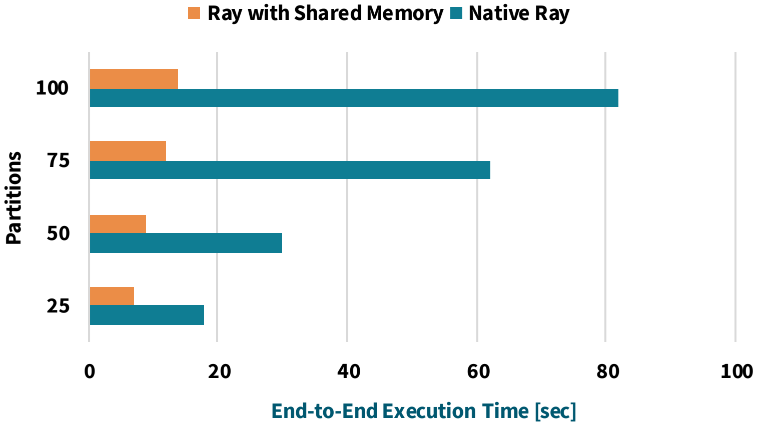
Use Case and System Benefit - DCD

- Memory Sharing (Collaborate with MemVerge)
  - No more object serialization and transfer over network for remote object access
  - No more duplicate object copies on different nodes → zero copy



[*Ray-based AI/ML System using CXL Shared Memory]

Ray with Niagara outperforms native Ray by up to 5.9x

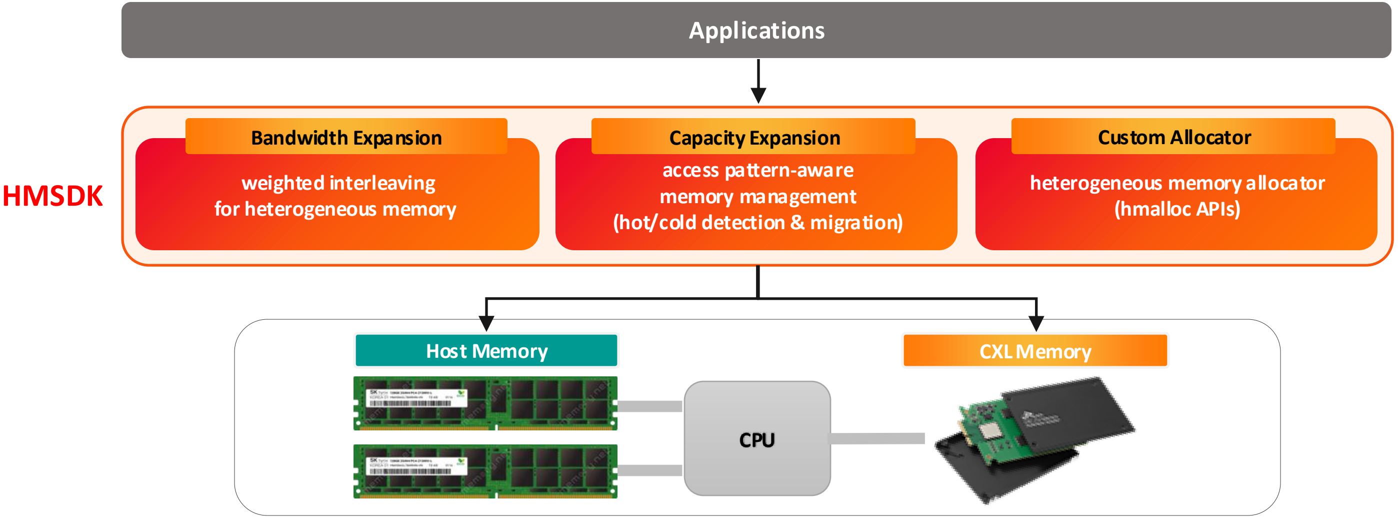


[Execution Time of Ray Shuffle Benchmark]

*Ray is an open source based distributed computing framework for AI/ML

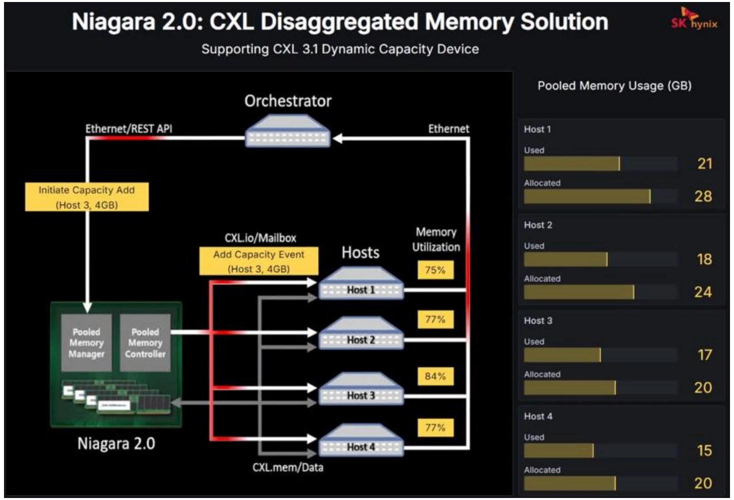
Use Case and System Benefit - HMU

- HMSDK (Heterogeneous Memory Software Development Kit)
  - HMU can reduce the profiling overhead of DAMON(Data Access MONitor), allowing HMSDK to monitor memory with finer granularity for enhanced accuracy
  - Better page migration decision with HMU can lead to performance improvement


Applications
HMSDK
Bandwidth Expansion
weighted interleaving for heterogeneous memory
Capacity Expansion
access pattern-aware memory management (hot/cold detection & migration)
Custom Allocator
heterogeneous memory allocator (hmalloc APIs)
Host Memory
CPU
CXL Memory

Demo in SK hynix Booth #B14

- Demonstrate the dynamic memory allocation/deallocation and hotness tracking of CXL disaggregated memory based on requests from CXL host servers



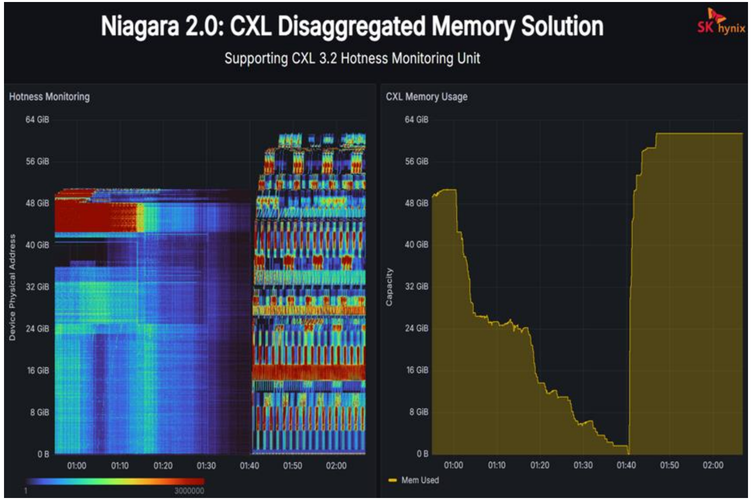
Niagara 2.0: CXL Disaggregated Memory Solution

Supporting CXL 3.1 Dynamic Capacity Device

Pooled Memory Usage (GB)

	Used	Allocated
Host 1	21	28
Host 2	18	24
Host 3	17	20
Host 4	15	20

[DCD Demo]



Niagara 2.0: CXL Disaggregated Memory Solution

Supporting CXL 3.2 Hotness Monitoring Unit

[HMU Demo]

Future Work and Call to Action

- Evaluation of system performance improvement based on page migration using Hotlist
- Research on disaggregated memory system architecture for AI Applications
  - System benefit for AI applications such as LLM (Large Language Model) and DLRM (Deep Learning Recommendation Model)
- Research on value-added function for efficient use of disaggregated memory
  - Near data processing
  - Fault tolerant disaggregated memory system

- Join OCP CMS (Composable Memory Systems) community and contribute to specifications
  - OCP CMS: https://www.opencompute.org/wiki/Server/CMS
- Get involved in open collaboration to enable CXL HW/SW ecosystem

Thank you!