Memory pooling and emerging architectures for efficient memory utilization using CXL™

EMPOWERING OPEN.

OCP GLOBAL SUMMIT

OCTOBER 18-20, 2022
SAN JOSE, CA

Memory Pooling and Emerging Architectures for Efficient Memory Utilization using CXL™

Ahmad Danesh
Sr. Dir, Product Management
Astera Labs

Dan Ernst
Principal Architect
Microsoft

Siamak Tavallaei
Chief Systems Architect
Google

OPEN PLATINUM™

Outline

- Abstract
- The State of Cloud Memory
- How Memory Pooling Addresses Stranded Memory
- Architectures and Tradeoffs
- Memory Pooling Demo
- Management Integration
- Call to Action

Abstract

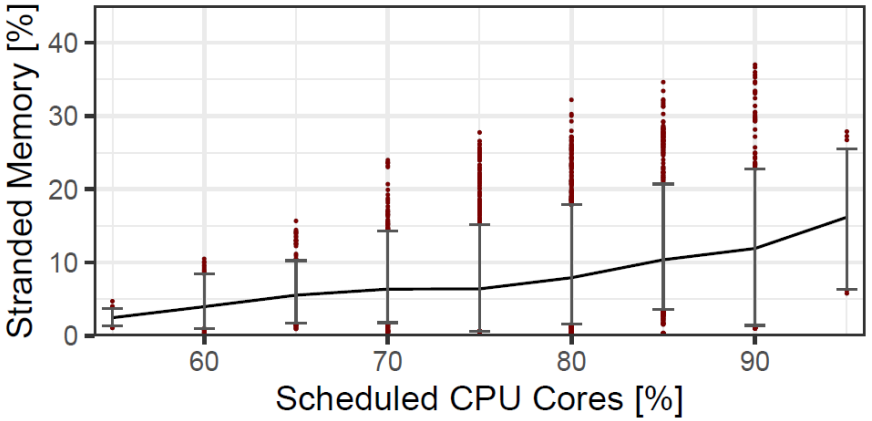
The prevalence of compute intensive workloads continues to increase throughout the data center and with it the need for increased memory bandwidth and capacity. While these workloads require more memory, workloads running on another server have unutilized (stranded) memory. With rising DRAM costs and DRAM accounting for a large percentage of server costs, increasing memory requirements and stranded memory will drive higher costs into data centers. In this presentation, we will describe how memory pooling with Compute Express Link™ will help address these challenges to efficiently utilize the deployed DRAM in a data center. We will compare memory pooling architectures enabled by the CXL™ 2.0 specification and new use-cases enabled with the recently released CXL 3.0 specification. We will briefly explore the ongoing collaborations within ecosystem and share insights for community feedback and continued collaboration to develop new OCP specifications and form-factors for memory pooling.

The State of Cloud Memory

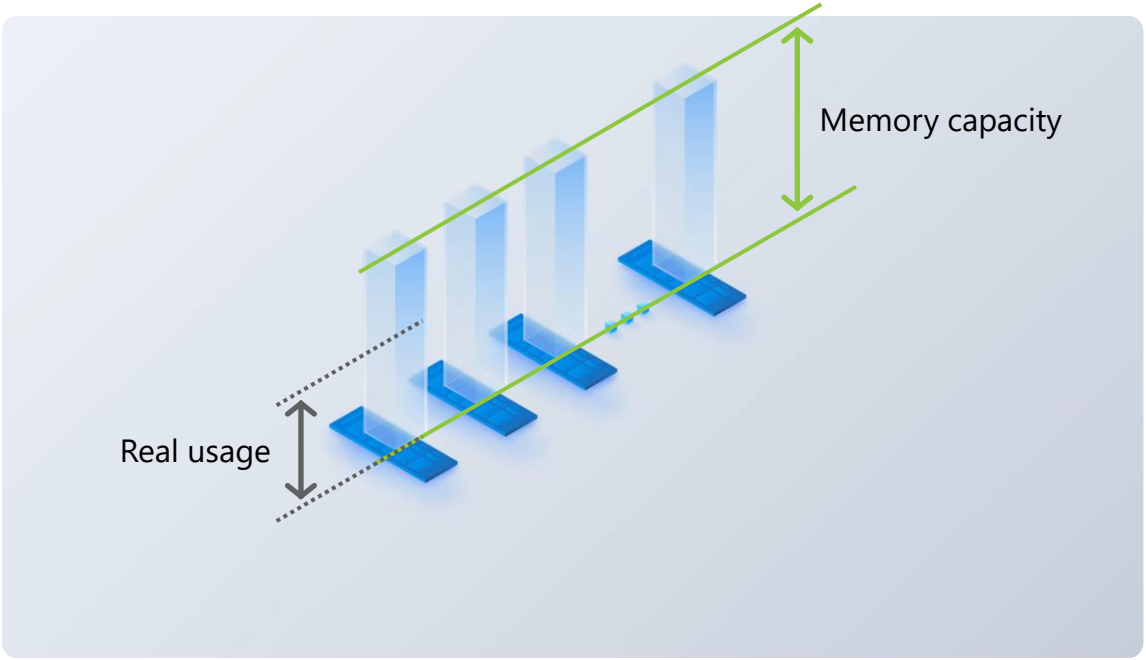
- Memory is becoming the largest portion of server costs for Cloud Datacenters
  - DRAM can account for up to 50% of server costs

- Often that memory is poorly utilized
  - Up to 25% is *stranded*
    (no free CPU cores but unsold memory)
  - Customers also overprovision memory

Median VM: **45%** untouched memory

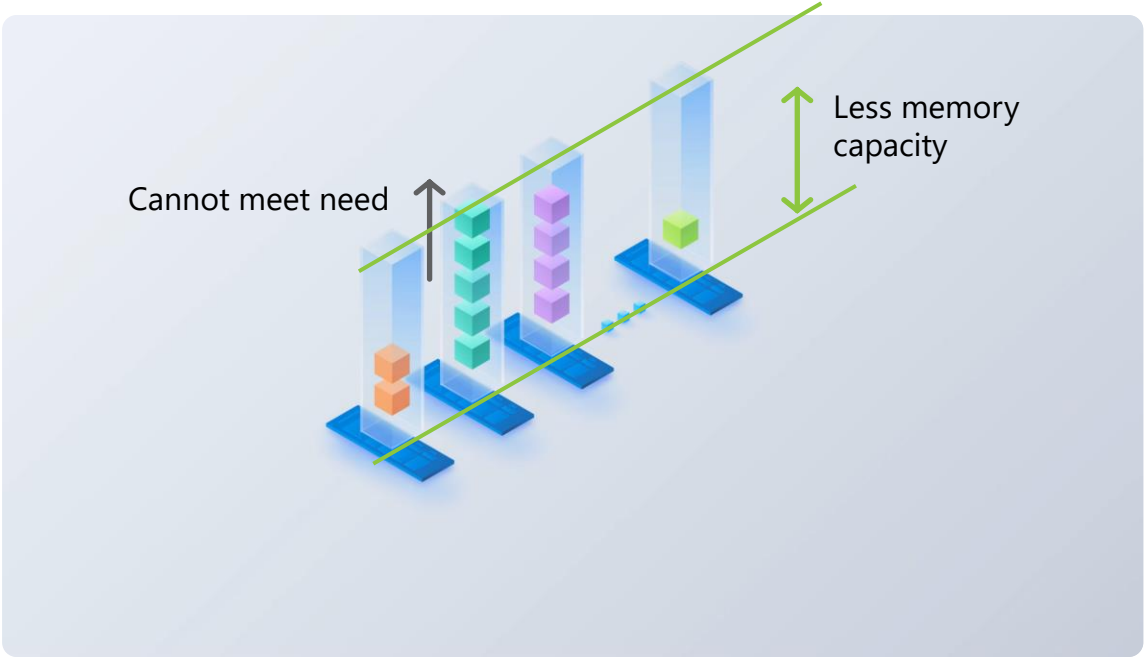


Stranded Memory in Today's Servers



Expensive memory is often under utilized

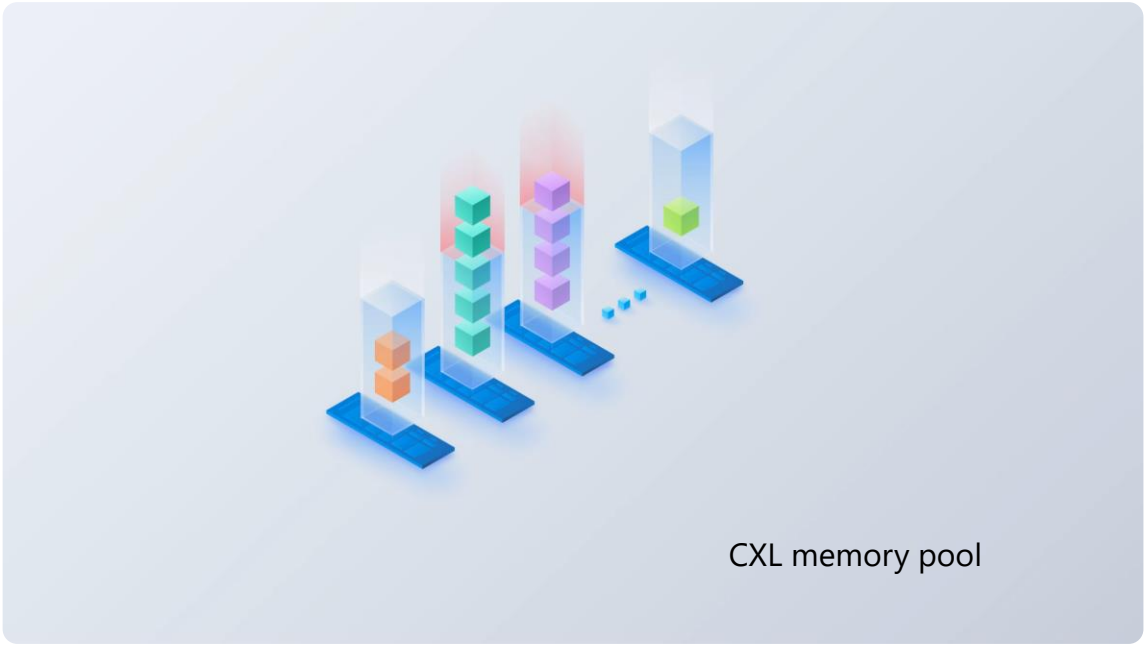
High Capacity for some Workloads



Expensive memory is often under utilized

Cannot just deploy less memory, as sometimes full amount is needed

Memory Pooling



Expensive memory is often under utilized

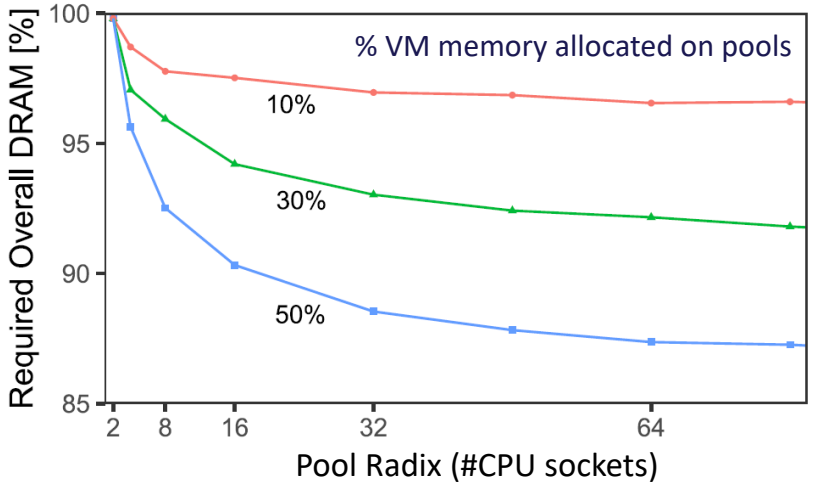
Cannot just deploy less memory, as sometimes full amount is needed

---

Use memory pooling to reduce total memory while still enabling high use

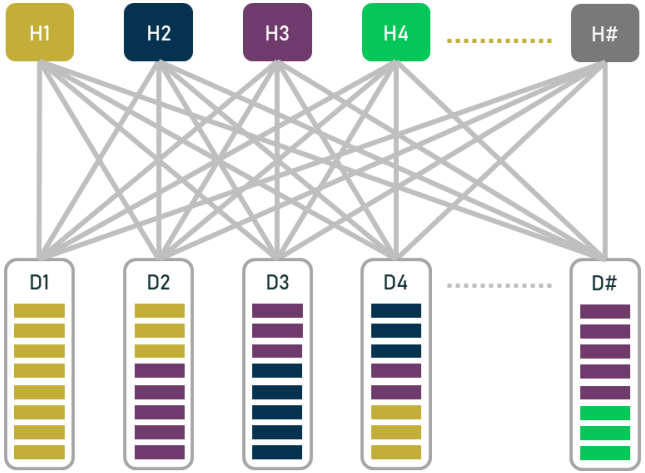
Memory Pooling Can Help

- **Simple idea:**
  - Share stranded memory with nearby servers that have free cores
- **Analysis results:**
  - Even small pool sizes are useful
- **Challenge**
  - Managing performance of workloads in hybrid memory
  - Small pool sizes can keep latency/ cost/power overheads low
- **Exploration of some solutions**
  - https://arxiv.org/abs/2203.00241 (updated 10/17/22)



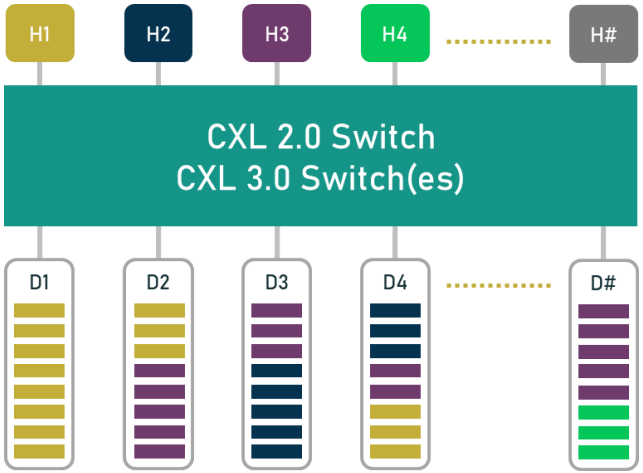
CXL Memory Pooling Architectures

Direct-Attached MH-SLD (CXL 1.1/2.0/3.0)



BENEFITS	CHALLENGES
• Lower latency	• Scale, cabling

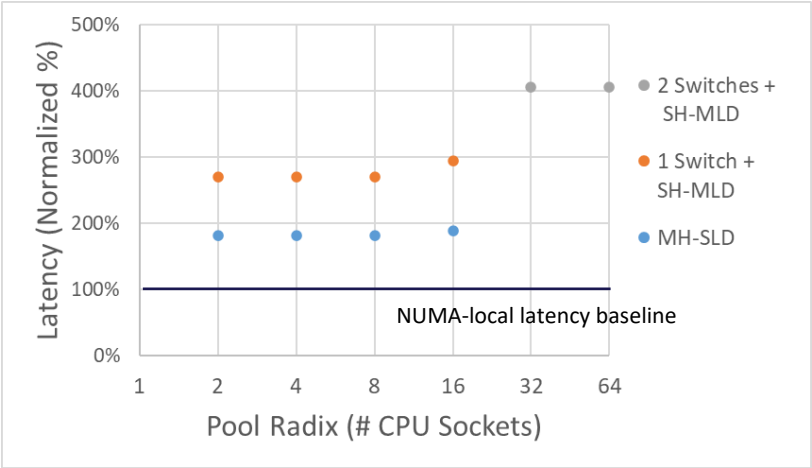
Switch-Attached SH-MLD (CXL 2.0/3.0)



BENEFITS	CHALLENGES
• Blast radius	• Higher cost/power

A Tradeoff in Design

- **Challenge**
  - Minimize latency to provide consistent performance for broad range of cloud workloads for general purpose servers
- **Insights**
  - The lower the latency, the broader the range of workloads that can be serviced with the pool and the higher the TCO savings
- **Analysis results**
  - By removing a layer of switching, multi-headed devices (MH-SLD) provide lower latency for an equivalent pool radix

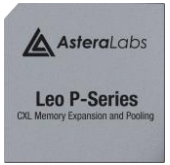


*Latency is estimated and not based on vendor data

Memory Pooling with MH-SLDs is Here Today

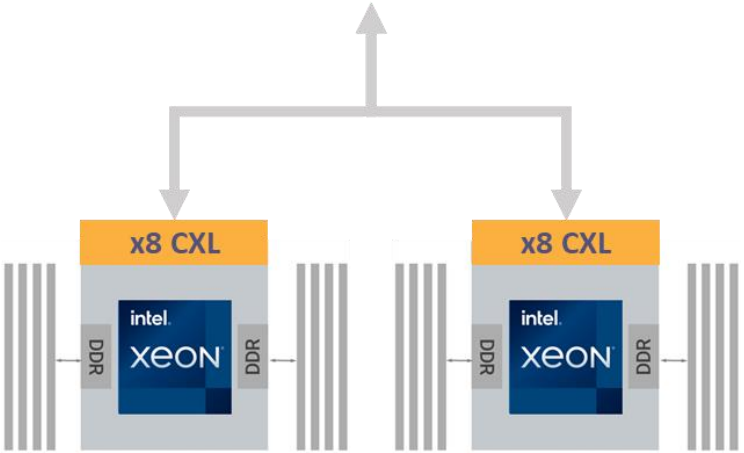
AsteraLabs

Leo Memory Connectivity Platform

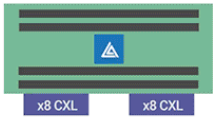


intel

4th Gen Intel® Xeon® Processor

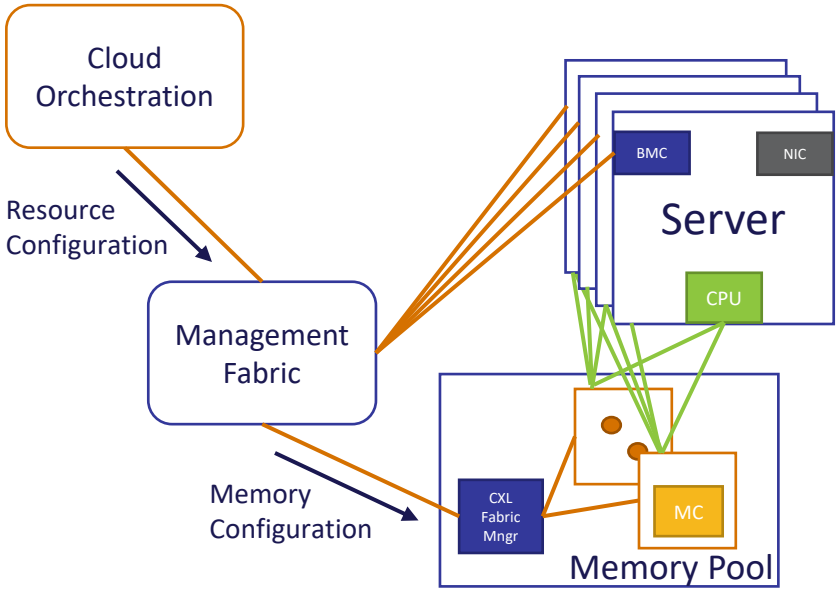


Astera Labs – CXL Memory Pooling Demo



See the Live Demo at Astera Labs Booth B11

Management Integration



Call to Action – Participate!

- CXL memory pooling is here today
- OCP will play a role in delivering infrastructure to support future pooling systems
  - Hardware solutions for modular systems, connectivity, and management
  - Infrastructure software elements (CXL Fabric Manager)
- Join the efforts with the software-defined memory workgroup

Open Discussion

EMPOWERING OPEN.

OCP GLOBAL SUMMIT

OCTOBER 18-20, 2022
SAN JOSE, CA