Software Defined Memory @ UBER – Past, Present and Future

EMPOWERING OPEN.

OCP GLOBAL SUMMIT
OCTOBER 18-20, 2022
SAN JOSE, CA

Software Defined Memory at UBER

SERVER

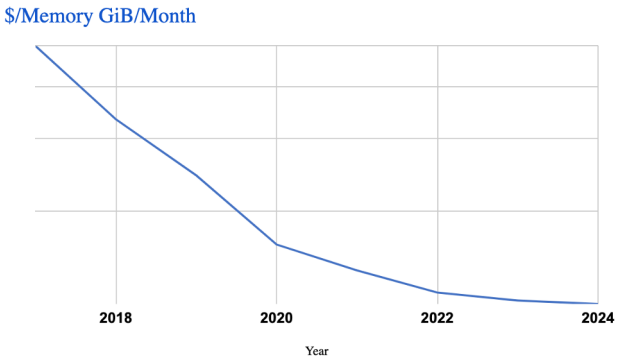
Nav Kankani, Systems Architect, Platform Architecture & Efficiency
Karim Mattar, Sr. Manager, Performance & Capacity Engineering
Lasse Vilhelmsen, Software engineer, Stateful Fleet Management

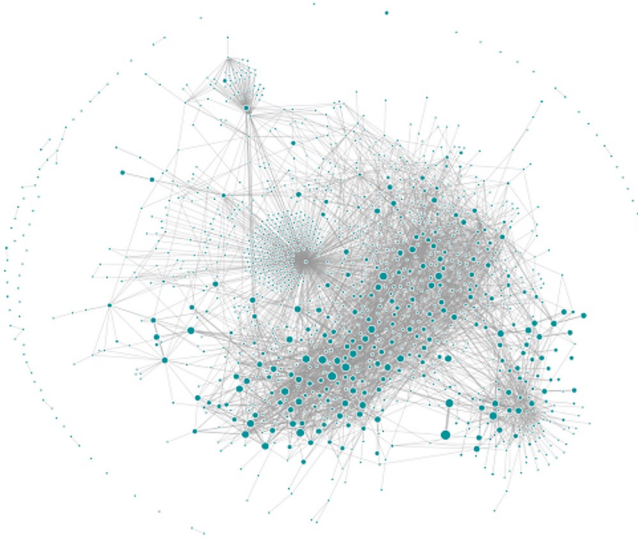
Memory Landscape

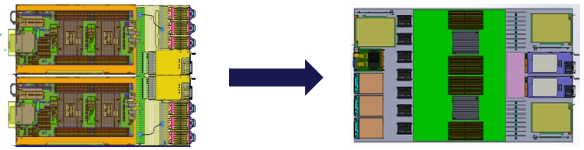
Performance ⟷ Capacity

Core Register	L1/L2/L3 Cache	Near Memory (HBM/DRAM)	Far Memory (CXL attached)	High-Performance Flash	Capacity Flash
Size	64KB-8MB	16GB-128GB	64GB-1TB	512GB-4TB	4TB-16TB
Latency	1-20ns	50ns- 200ns	200ns-1us	20us-1ms	1-5ms
Tiers	M0 (Critical)	M1 (Hot/Active)		M2 (Warm)	M3 (Inactive/Swap)
Cost/GB	1000x-100x	10x-1x	1x-0.5x	0.3x – 0.1x	0.03x

Industry & UBER Trends





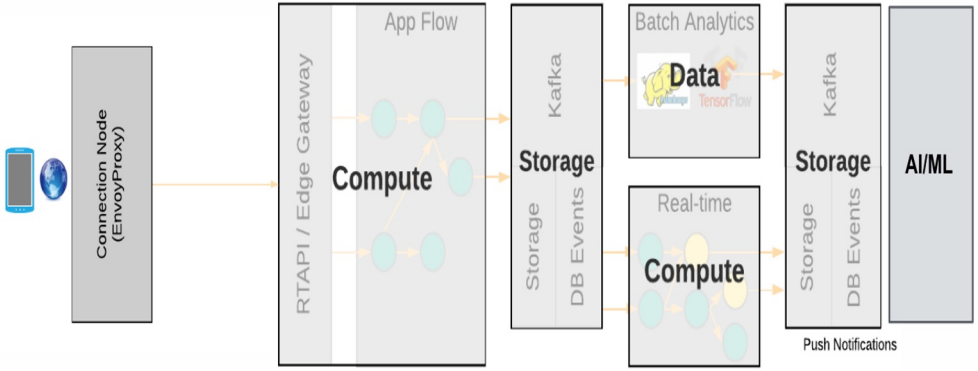


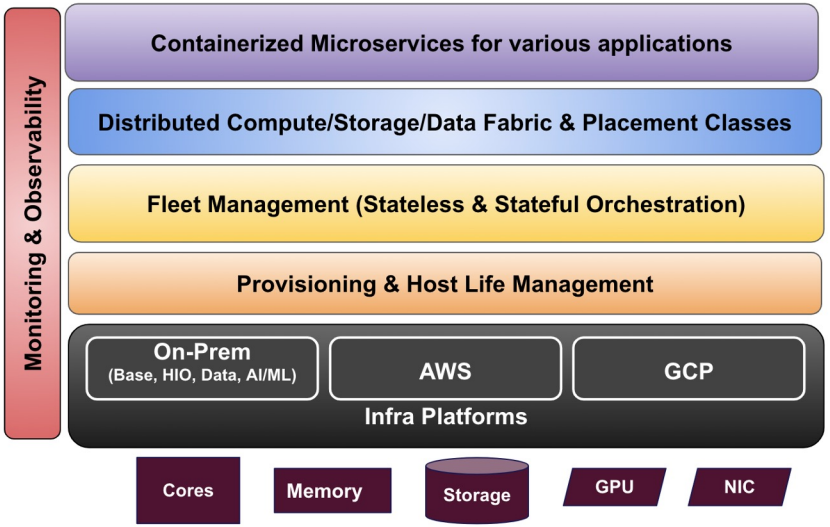
Memory cost is flattening; Highest commodity spend at UBER

Growing application demand & diversity in core:memory ratio

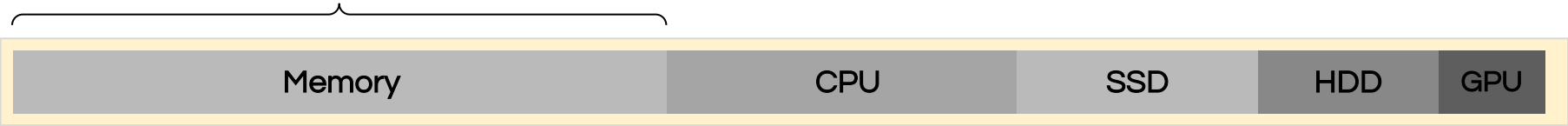
Physical limitations for memory expansion hinders scale-up strategy

UBER Infrastructure

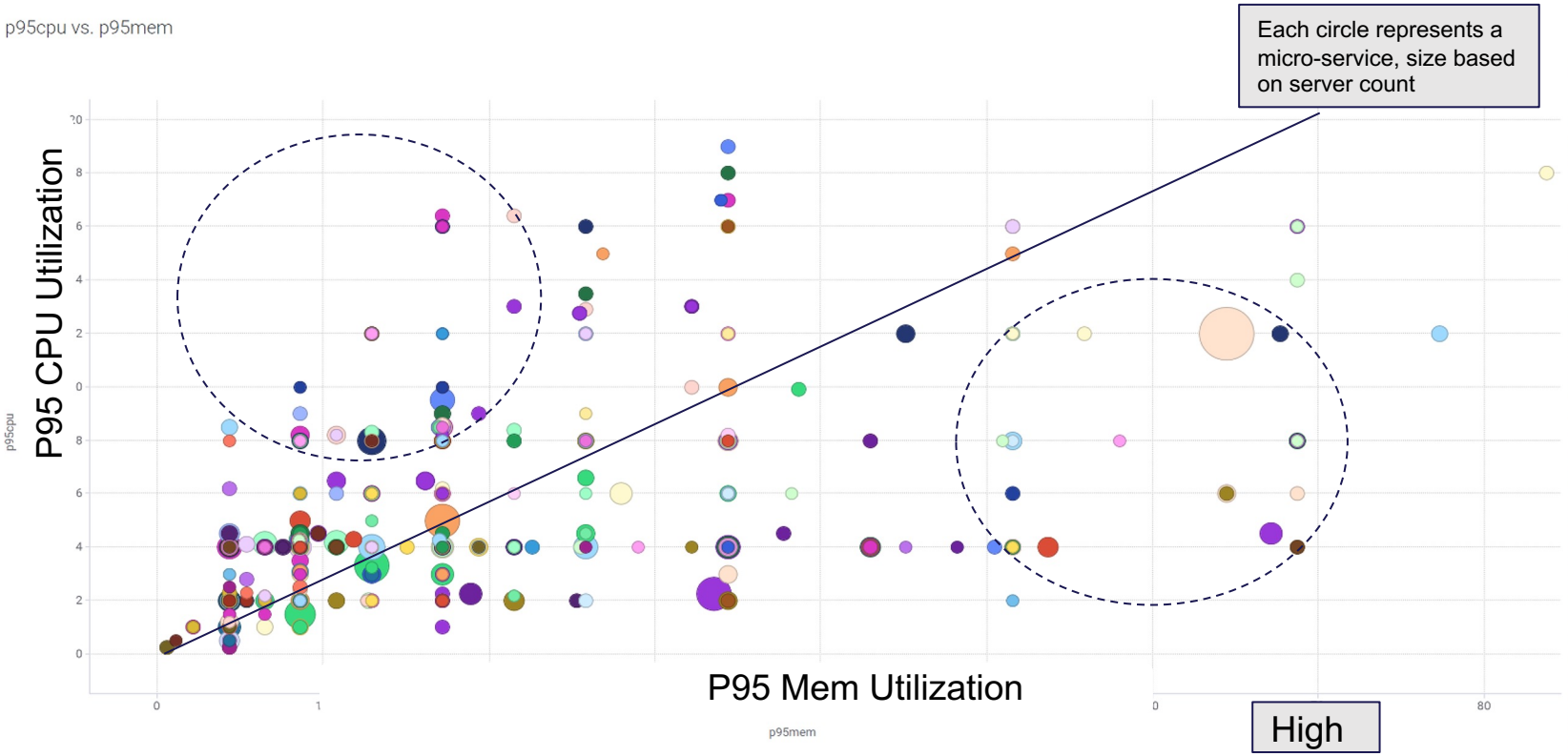




Commodity Spend dominated by Memory



Memory Usage Trends for Compute Microservices



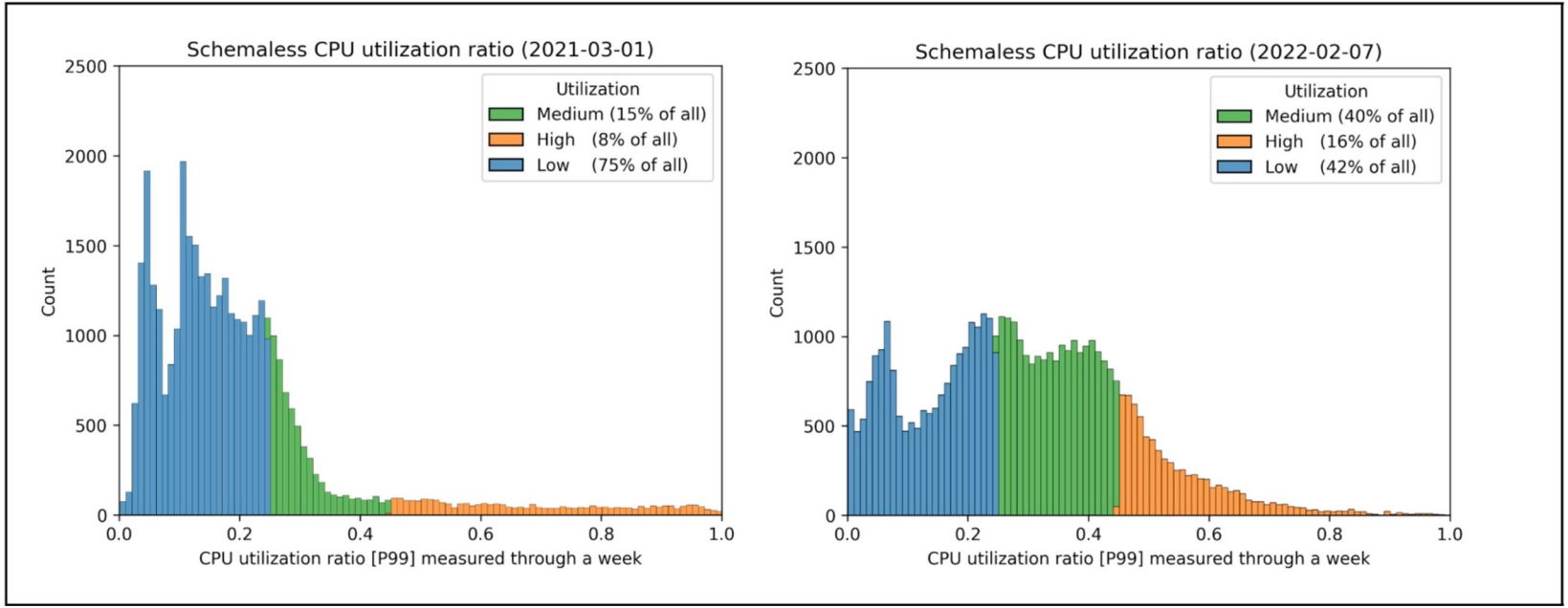
SDM Goals at UBER

- Reducing memory cost across Infrastructure
- Memory expansion to support scale up strategy
- Shared fabric level orchestration of memory

Memory Reduction Initiatives: Overview

Attributes/Timeline	Past (2018-2022)	Present (2022-2023)	Future (2023-2026)
Right-Sizing	Number of container instances were **user defined**	Centralized scaling of number of container instances on **cores**	Improved scaling based on **memory & cores**
Collocation	Technologies were run on **isolated hardware**	Technologies are **stacked** on the same hardware	Improved utilization based on resource vectors & **colocation aware of memory tiers**
Tiering	No tiering	Proof of Concept work with CXL Memory	**CXL expansion** of memory with cheaper cards
Pooling	NA	Not started	Pooling of aggregate **memory buffers**

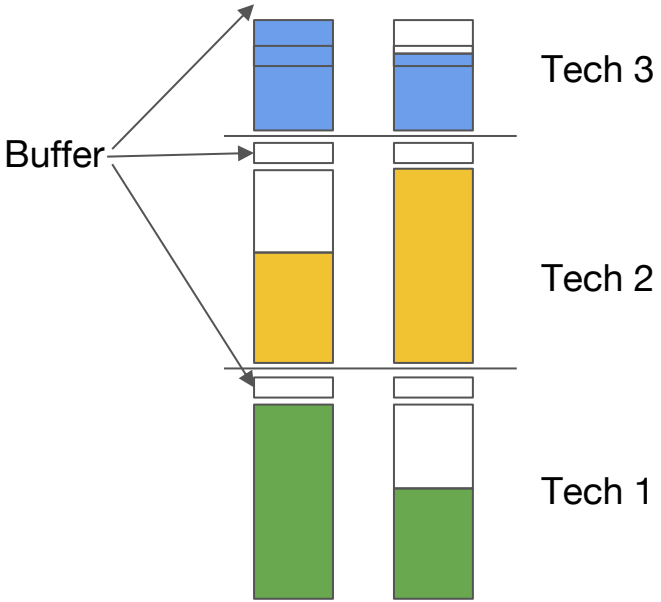
Right Sizing Allocations

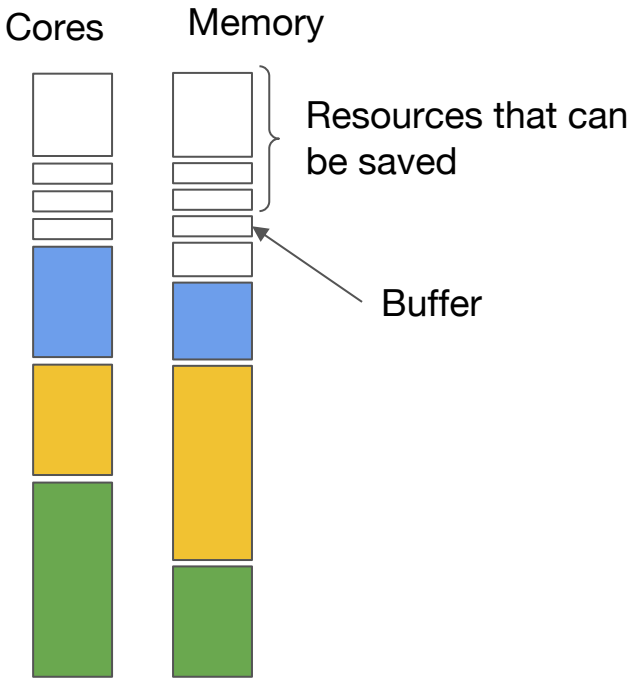


- **Right sizing allocations based on memory is important, even if it results in stranding of cores.**
- **Exploring metrics other than memory utilization for scaling memory allocations**

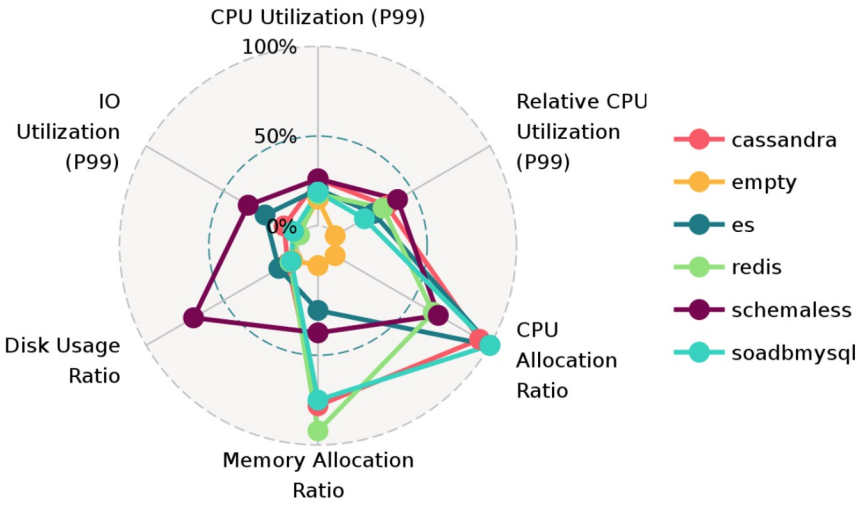
Colocation of Services

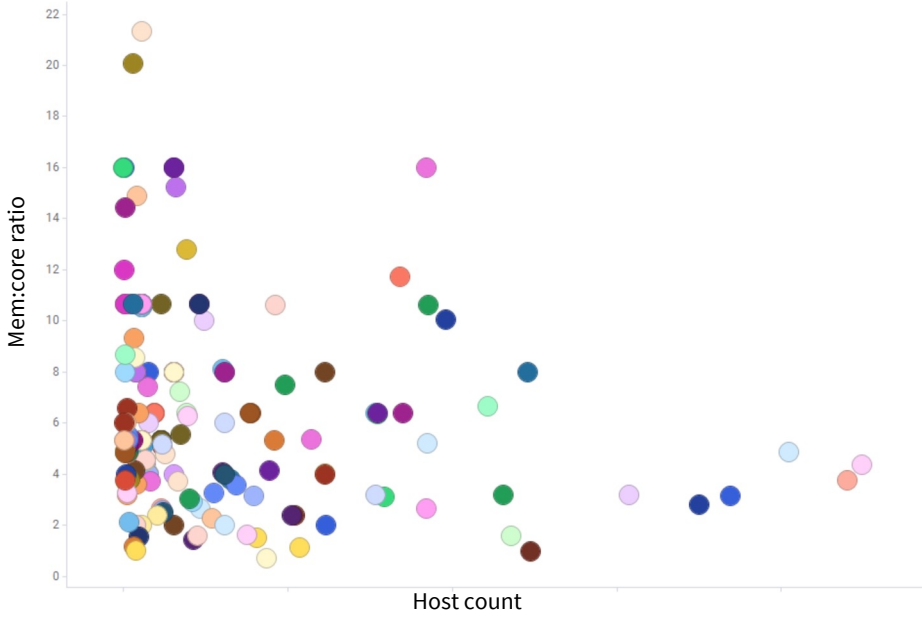
By colocating technologies, overall required host count is reduced





Memory Usage Trends for Storage Technologies

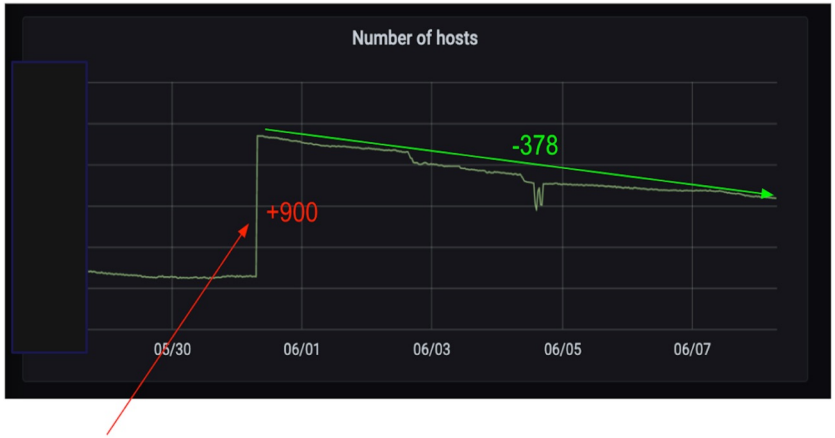




Variability of resources across storage technologies

Diversity of mem:core ratio across storage instances

Results from Co-locating Storage Technologies



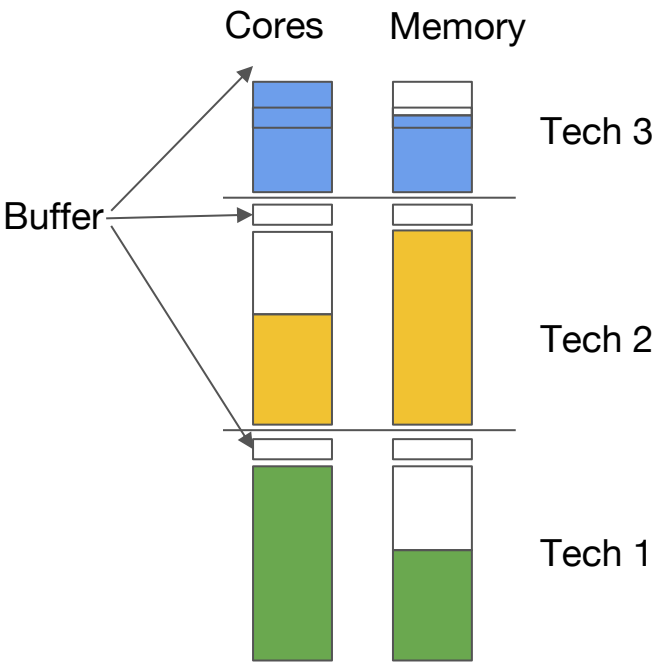
42% reduction hosts used

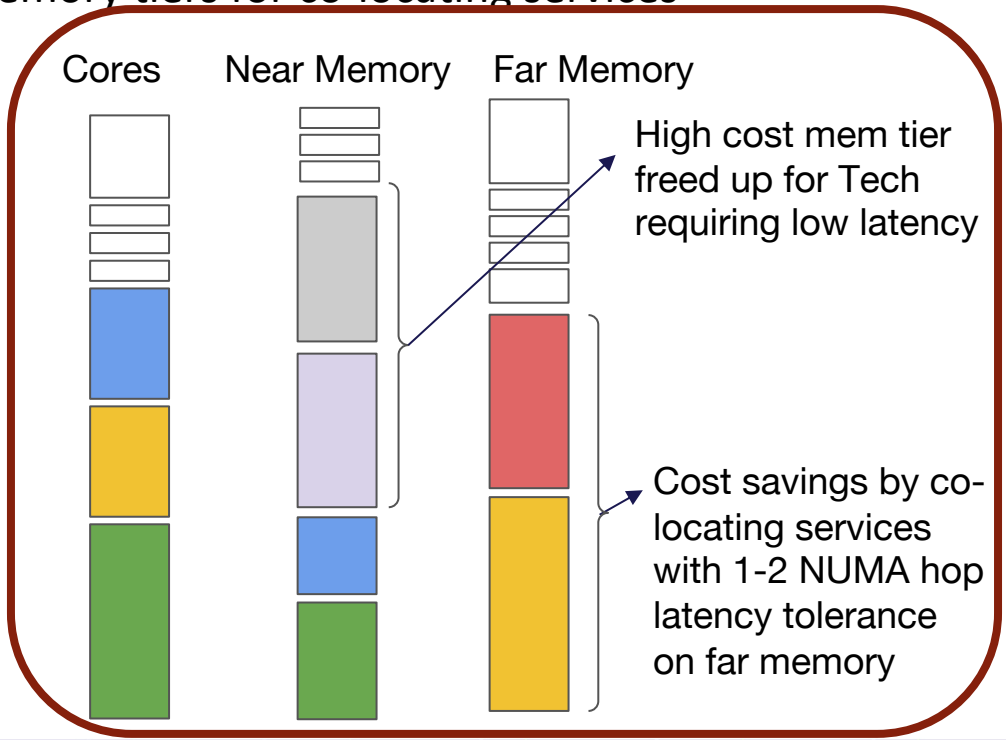
After co-locating storage technologies, memory is now the bottleneck

- Significant resource & cost savings when co-locating services based on cores.
- Colocation based on resource vectors including memory with further improve memory utilization

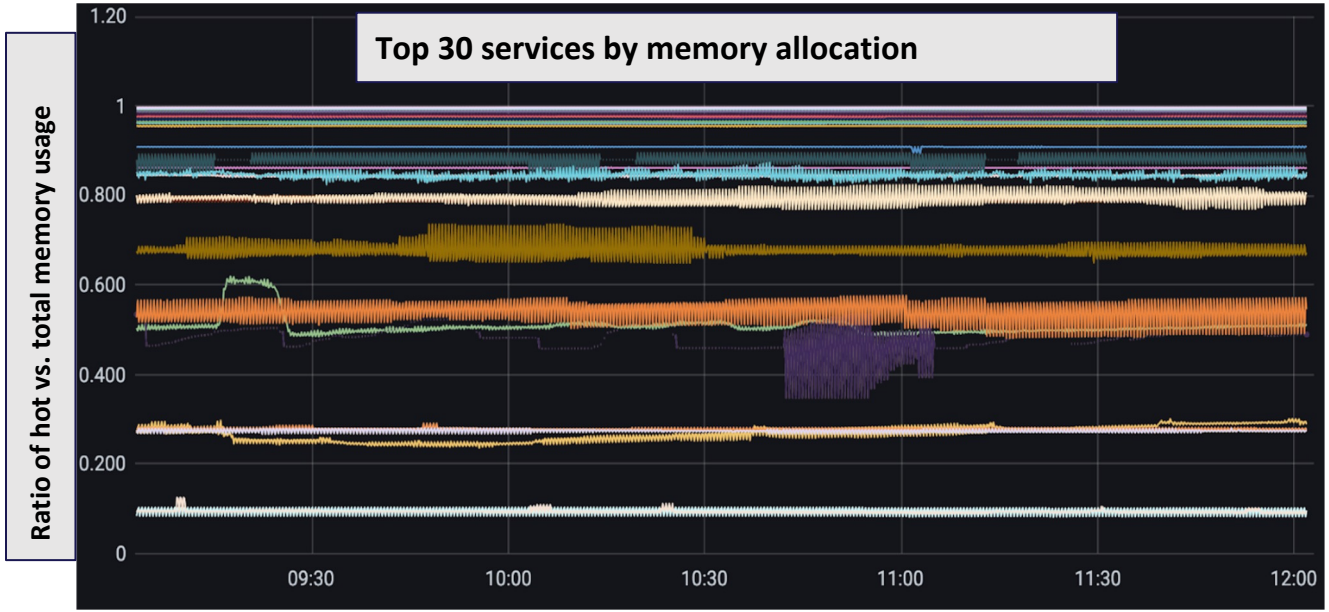
Colocation of Services aware of memory tiers

Shared Storage Fabric aware of different memory tiers for co-locating services





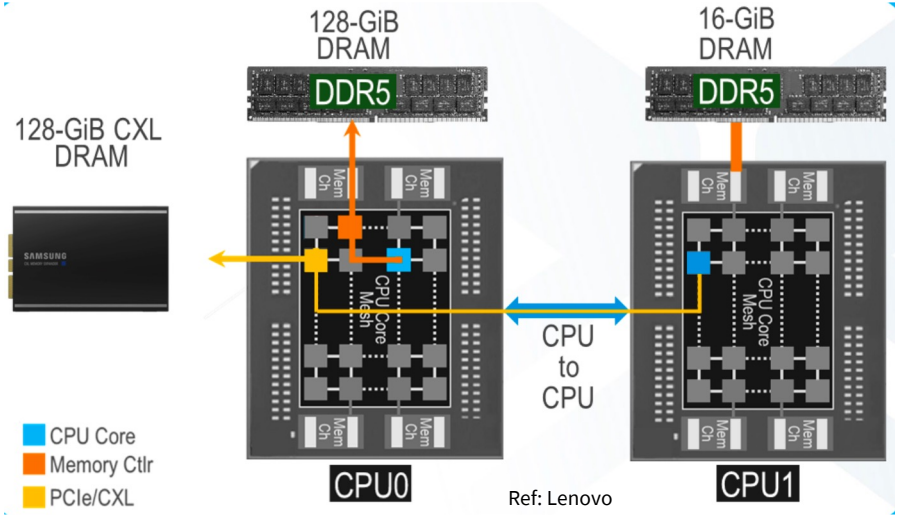
Opportunity for Memory Tiering through Hot/Cold Profiling



- 40% of services show “memory coldness” for at least 50% of the memory capacity
- Lots of opportunities for Near -> Far memory Swap opportunities in UBER infra

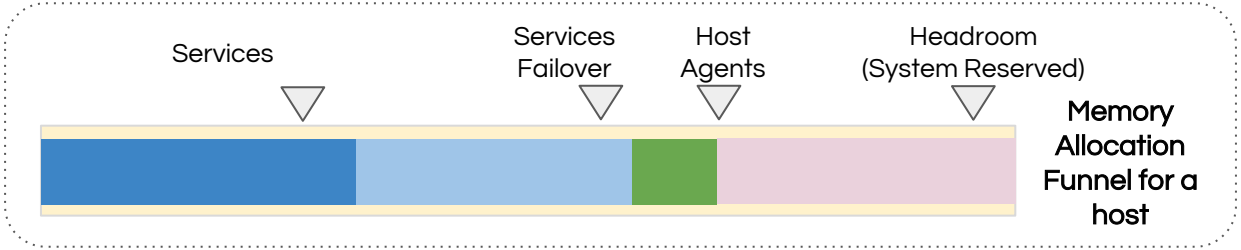
Exploring Memory Tiering with Zswap & CXL

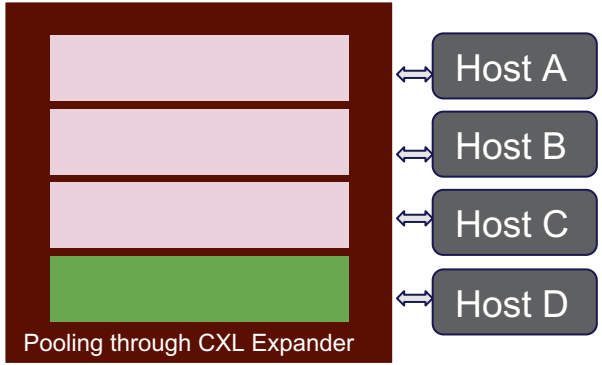
- Compress cold pages using Zswap/Swap to NVMe devices.
- Retirement of cold pages to CXL attached far-memory through Kernel managed auto-tiering and other page placement algorithms.
- Benchmarking with 1-NUMA hop delays to quantify relative application performance degradation

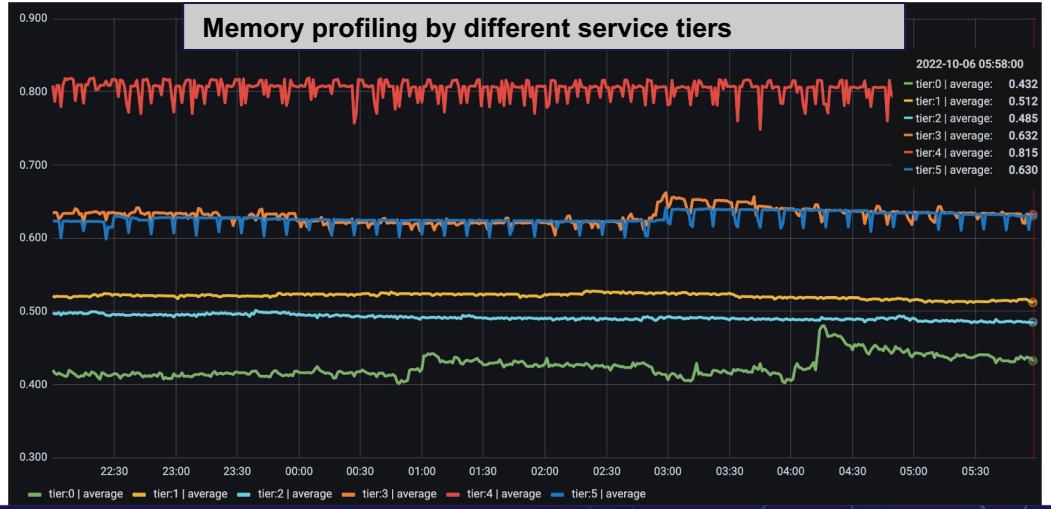


Ref: Lenovo

Unlocking further efficiency through Memory Pooling & Disaggregation







- Significant cost savings to be unlocked by pooling memory buffers!
- Initial target for memory pooling can be revocable tiers (ex: tier 3 & 5)

Few Challenges

- Failover scenario & estimation for scaling memory allocations
- Reliability impact & OOM errors due to sub-optimal kernel reclamation
- Colocation of services on heterogeneous SKU's with different Memory:Core ratios
- Swap enablement by Stateful fabric for multi-tenant services

Call to Action

- Collaboration needed across the stack to accelerate Software Defined Memory efficiency efforts
- Encourage Industry partnerships, leverage Open-source work & sharing through SDM community partners
- Standardization across CXL device vendors, common feature sets, OCP specifications for CXL memory expansion

Thank you!

EMPOWERING OPEN.

OCP GLOBAL SUMMIT

OCTOBER 18-20, 2022
SAN JOSE, CA