Flash Memory Summit

Memory Tiering with CXL Ready Systems and Devices

Presenter:

Ravi Kiran Gummaluri

Micron Technology

Agenda

- Memory demand and scaling challenges
- CXL memory expansion
- SW + HW heterogenous Interleave
- Memory capacity and bandwidth solutions
- Workload Performance Analysis on CXL ready systems
- Conclusions and Next steps

Memory Demand and Scaling challenges

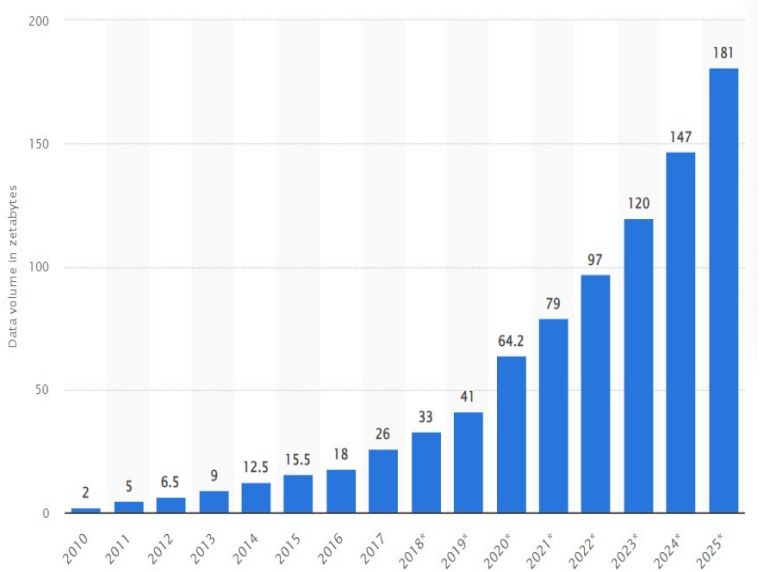


Figure 1: Growing memory usage

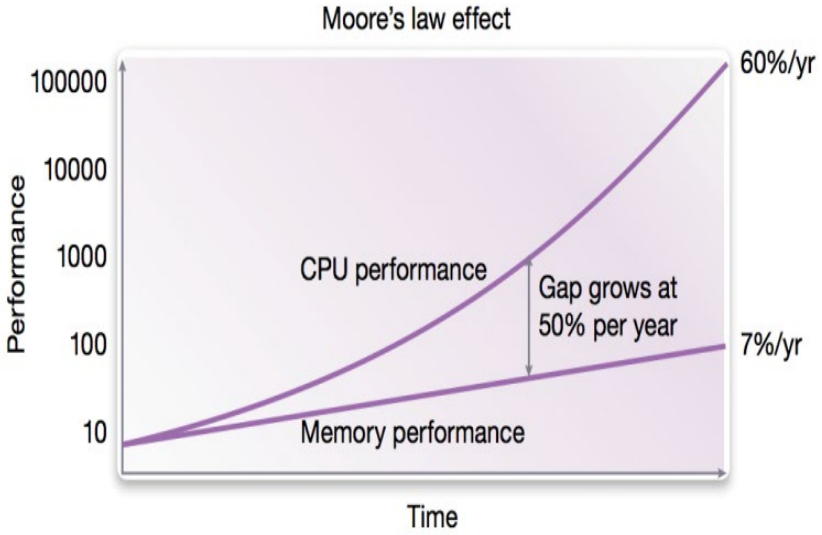


Figure 2: Memory wall

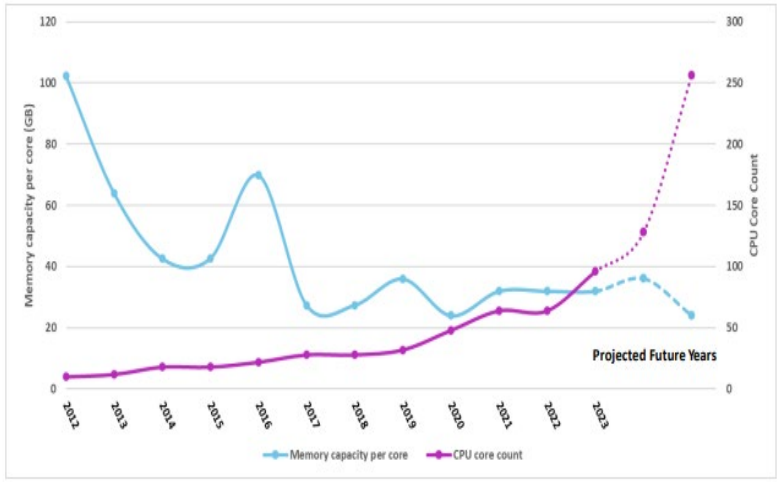


Figure 3: Memory capacity Vs CPU cores

- Growing demand for Memory need in data center applications . (~26 % yoy)
- DRAM is not scaling -> Memory Capacity is doubling every four years.
- Memory Latency -> is only improving 1.1 times every two years.
- Processor speed -> has been doubling every two years.
- Increased TCO for Data Centers -> Memory is ~ 50% of the overall server cost .
- How do we solve increased Memory Bandwidth , Capacity requirements and reduce TCO ?

Figure 1 : Source: https://www.statista.com/statistics/871513/worldwide-data-created/
Figure 3 : Source: Based on capacity and core counts from publicly available AMD and Intel datasheets, and public statements.

CXL Memory expansion

- **CXL Memory Expansion**
  - Cache-line granular access semantics.
  - CXL-Memory appears to a system as a CPU-less NUMA node. (Not dependent on CPU Arch)
  - Hot Pluggable memory
  - Works with various form factors E1.S, E3.S , Add on Card etc
  - Interoperable with various memory types (DDR4, DDR5, LPDDR5, NVM
- **CXL Memory Capacity Expansion**
  - CXL Direct attached Memory Tiering
    1. Application Transparent
       - OS Managed
       - User Space Library
       - 2LM mode
    2. Application Managed
       - Application Aware (ex: libnuma)
       - Modified (ex : libmemkind)
  - CXL Switch / Fabric attached Memory Tiering
    - Another Memory tier added to system with higher latencies.
- **CXL Memory Bandwidth Expansion**
  - CXL Heterogenous interleave solutions
    1. Hardware based Interleave
    2. Software and HW heterogenous interleave.
    3. Software based NUMA interleave.

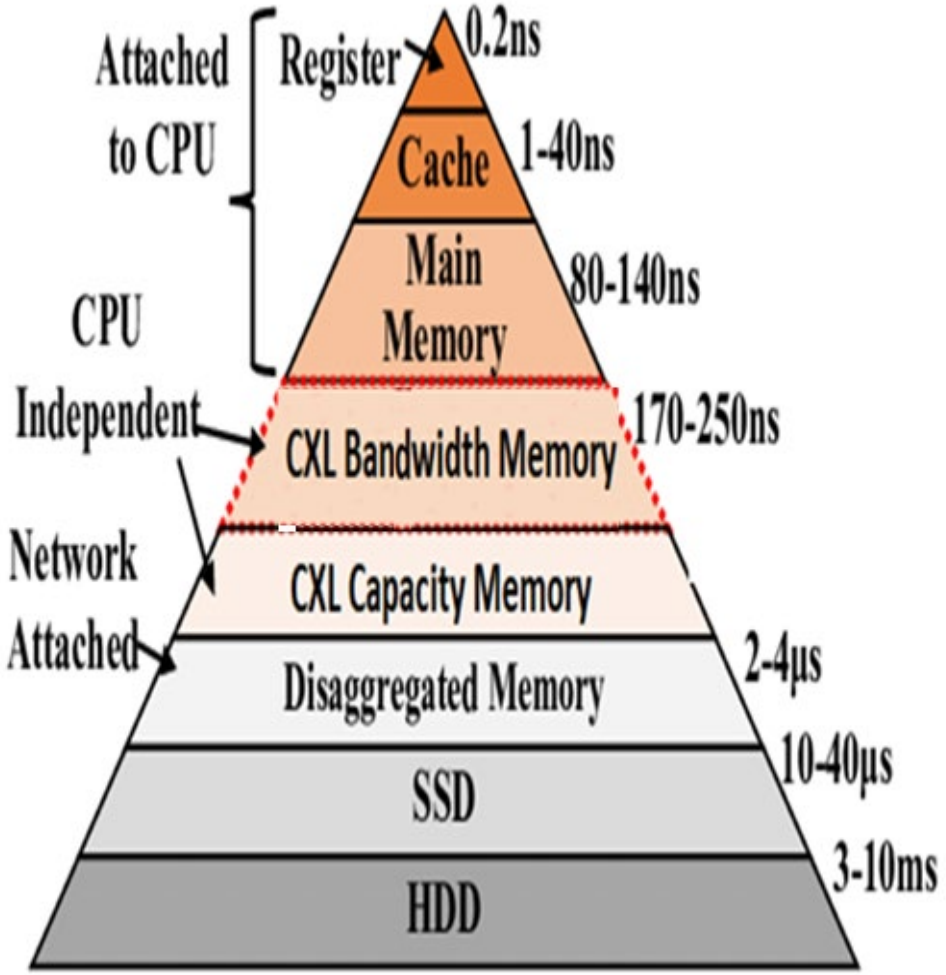
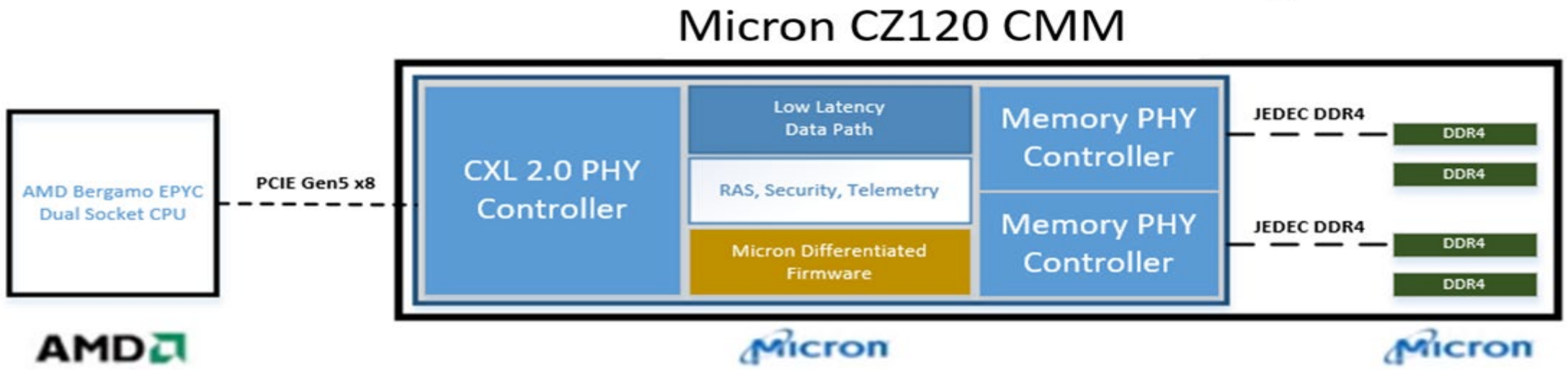


Figure 4 : Memory Hierarchy

SW + HW Heterogenous interleave configuration

HW Configuration :

AMD Bergamo	128 Cores/Socket	12 Mem Channels/Socket DDR5 – 4800MTps	460 GBps Total System Bandwidth	768 GB Total System Capacity/Socket	135% System Capacity Increase
MICRON CZ120 CMM	4 CXL Cards/Socket PCIE Gen 5.0 x8	160 GBps 2R:1W CMM Bandwidth	1024 GB Total CMM Capacity/Socket		33% System Bandwidth Increase

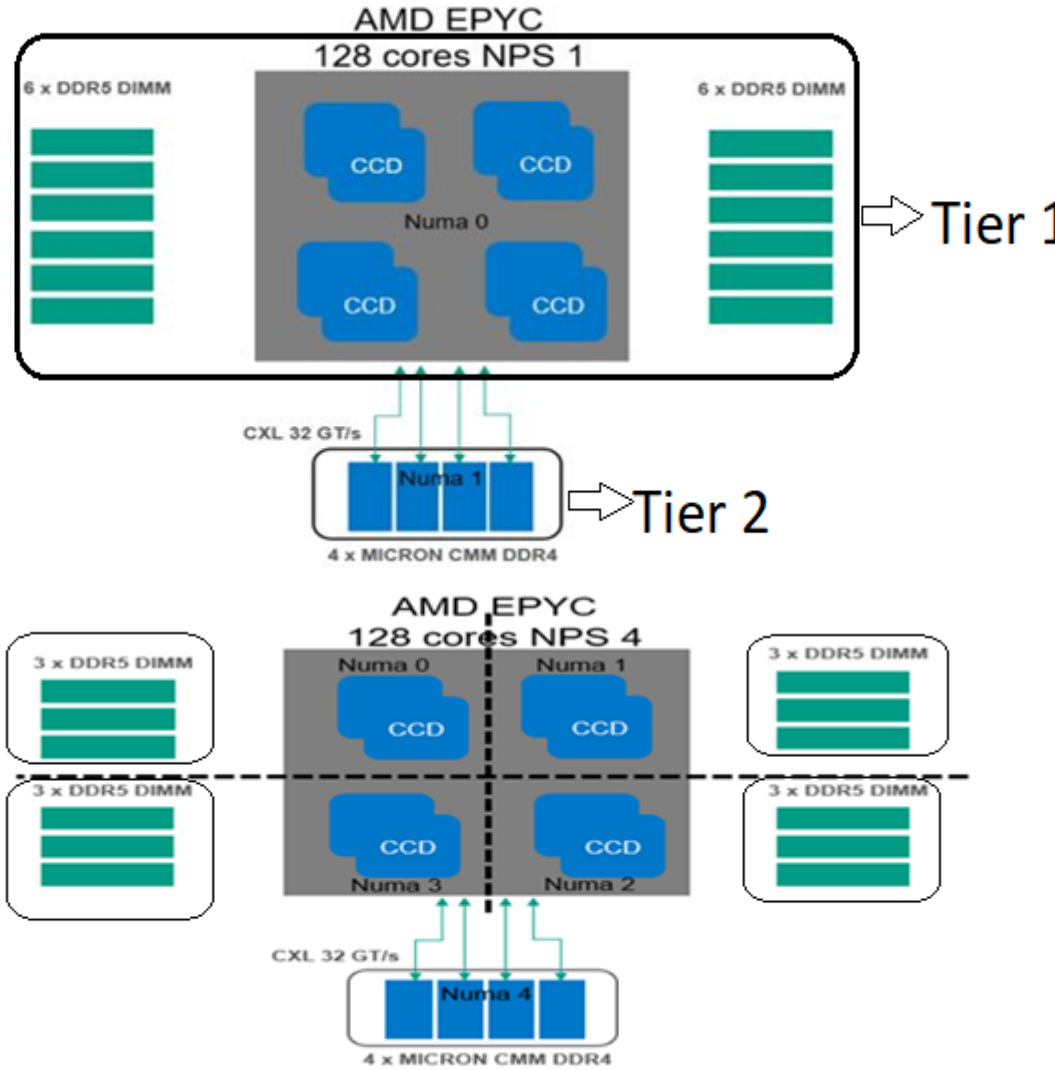


SW Configuration :

numactl [**--interleave nodes**] [**--preferred node**] [**--membind nodes**] [**--cpunodebind nodes**] [**--physcpubind cpus**] [**--localalloc**] [--] command {arguments}

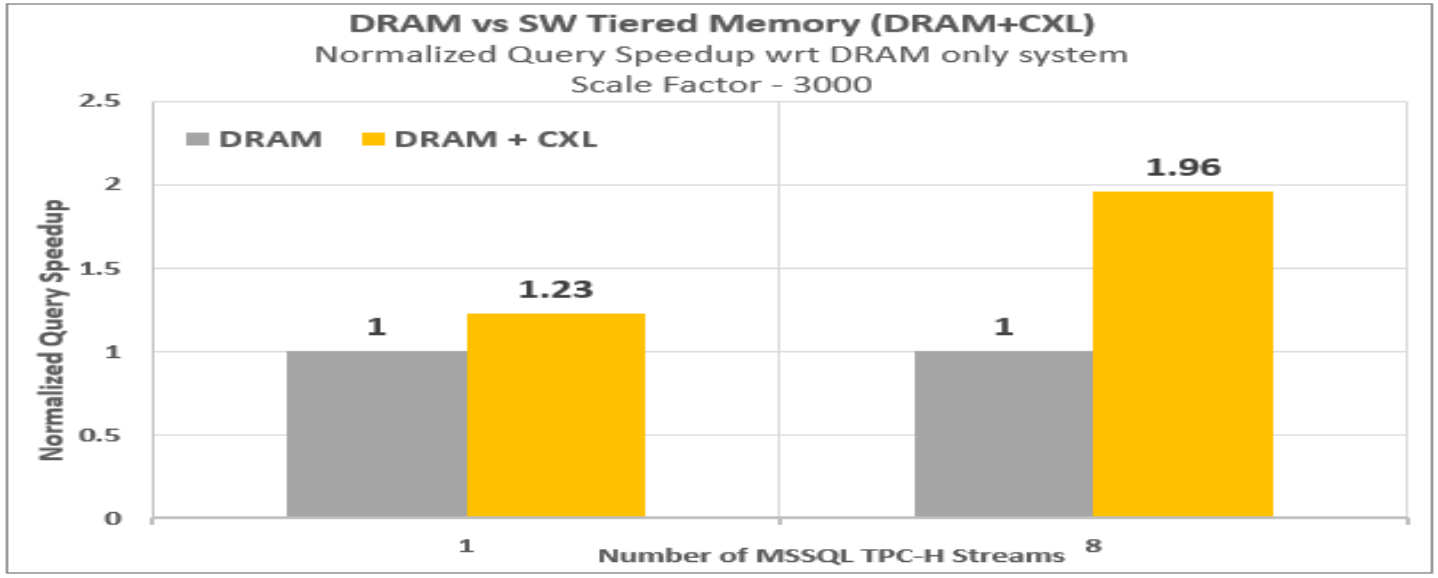
Bandwidth and Capacity solutions

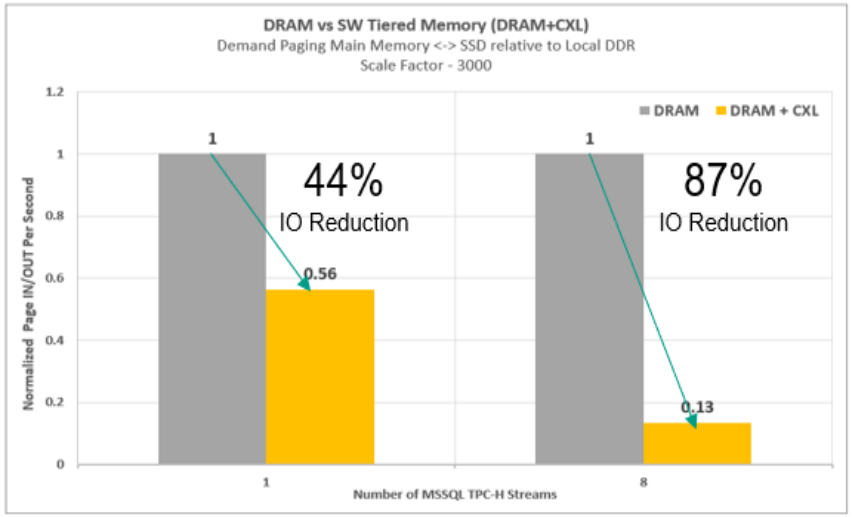
- AMD Bergamo platform equipped with AMD EPYC processors offers wide configurability of NUMA domains. It supports the concept of NPS (NUMA Node Per Socket) for improving the performance of different workloads.
- **CXL memory capacity expansion using memory tiering**
  - NPS1 (12 channels DDR5 - 1 NUMA domain) + (4 CXL CMM - 1 NUMA domain)
  - NPS1 – Each socket is in a single NUMA domain, with all the cores in the socket and its associated memory connected to the socket in one NUMA domain.
  - Hot data near main memory(NPS1), warm data in CXL and cold data in storage media .
  - Kernel can profile page hotness and manage promoting CXL hot pages to Local main memory.
- **CXL – Software + HW Heterogenous interleaving**
  - NPS4 (4 NUMA domains) + CXL NUMA domain (1 NUMA domain) in a 4:1 interleaved fashion with numactl .
  - NPS4 :Each socket is partitioned into 4 NUMA quadrants/domains. Each NUMA domain has 3 memory channels and memory is interleaved across these 3 memory channels in each quadrant.

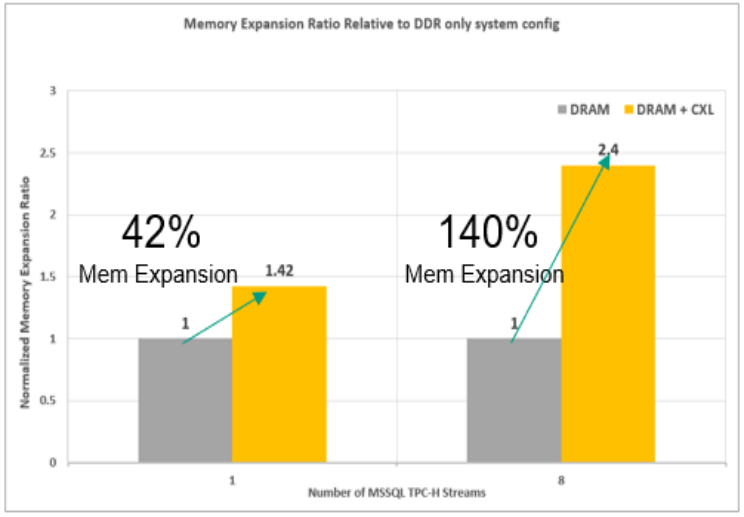


Workload Performance Analysis : MSSQL + TPC-H

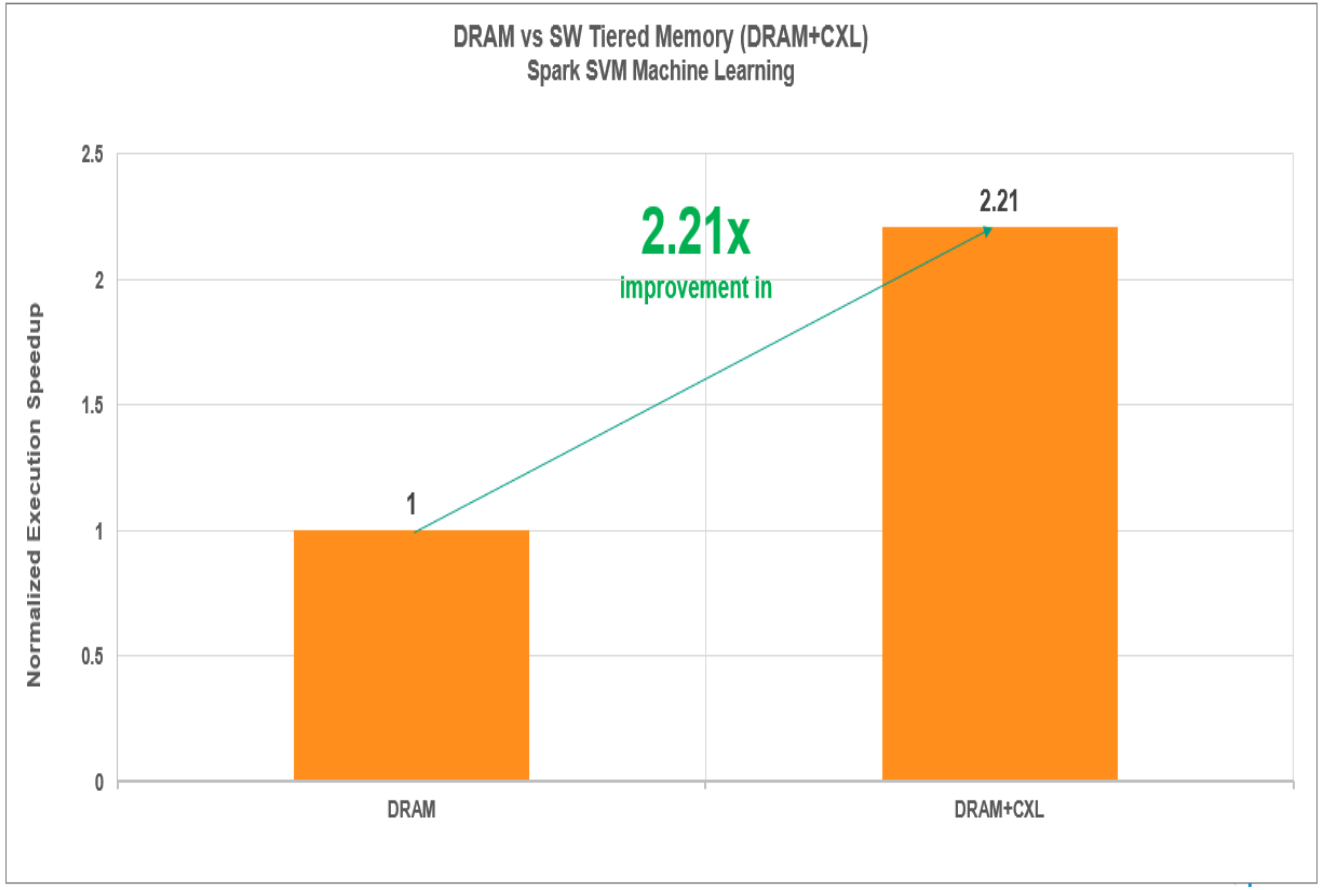
Hardware Setup	
Processor	AMD EPYC 9754 128-core (Bergamo)
Memory	768 GB DRAM – Near (Tier 1) Memory (12 x 64GB Micron DDR5 DIMMs) 1024 GB – Far (Tier 2) Memory (4 x 256GB Micron CZ120 CMMs)
Storage	8x Micron 7450 NVME SSD





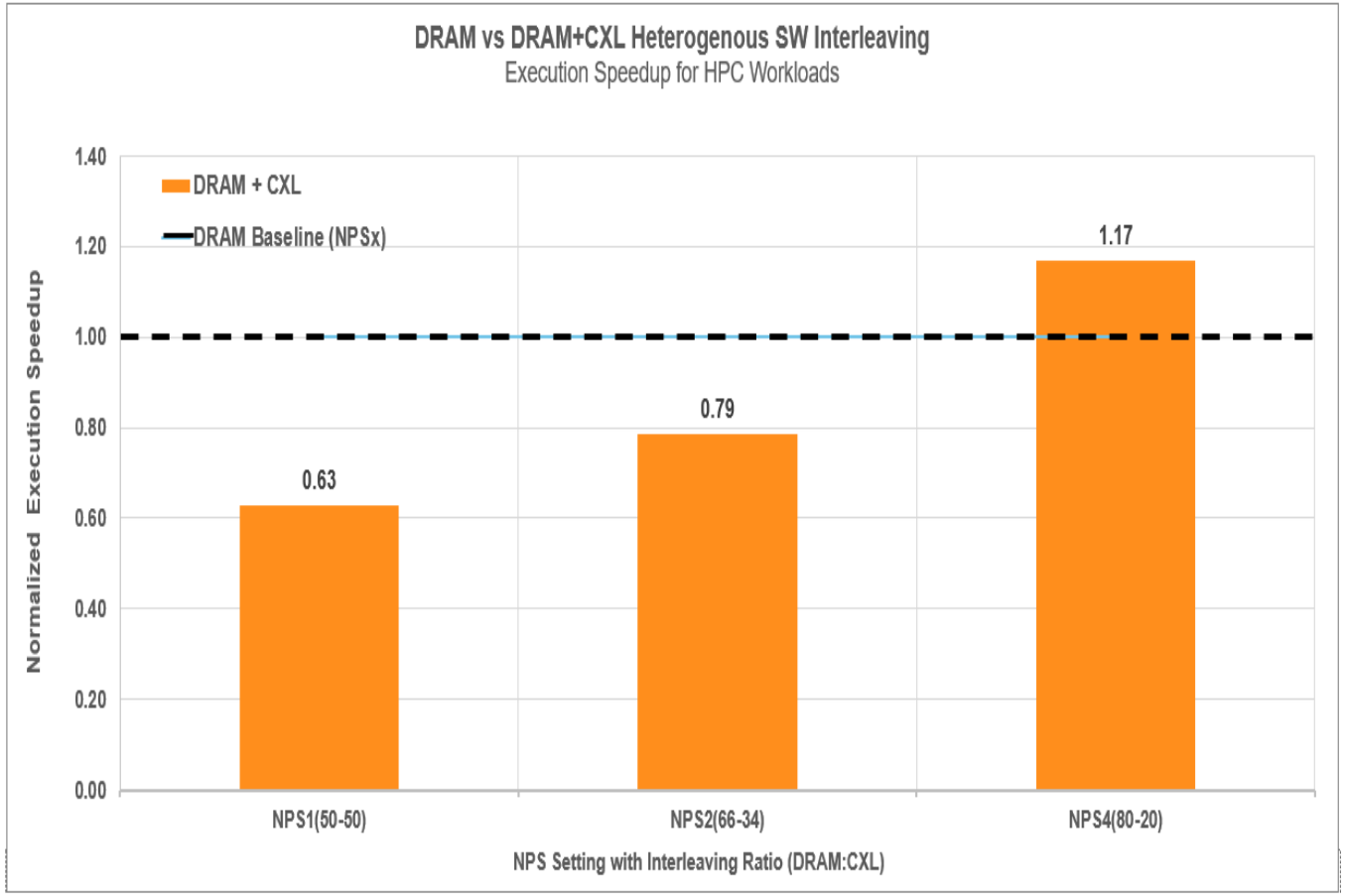


Workload Performance Analysis : Spark ML SVM



Hardware Setup	
Processor	AMD EPYC 9754 128-core (Bergamo)
Memory	768 GB DRAM – Near (Tier 1) Memory (12 x 64GB Micron DDR5 DIMMs) 1024 GB – Far (Tier 2) Memory (4 x 256GB Micron CZ120 CMMs)
Data Set	360GB

Workload Performance Analysis : Clover Leaf



Hardware Setup	
Processor	AMD EPYC 9754 128-core (Bergamo)
Memory	768 GB DRAM – 4 NPS NUMA nodes (each 3 x 64GB Micron DDR5 DIMMs) 1024 GB – 1 NUMA Memory node (4 x 256GB Micron CZ120 CMMs)
Interleaving Ratio	50%-50% (NPS1), 66%-34% (NPS2) and 80%-20% (NPS4)

Conclusion / Next Steps

Conclusions :

- CXL memory can provide a solution to increased Memory Bandwidth and Capacity requirements .
- CXL memory can help in bandwidth expansion using SW + HW based heterogenous interleaving between DDR and CXL memory. Bandwidth sensitive workloads, such as CloverLeaf can benefit this by reducing backend end memory stalls .
- CXL memory when introduced as tiered memory can help in increasing memory capacity and reducing latency impact of Storage media . Capacity sensitive workloads , Such as TPC-H can benefit by significantly reducing the number of I/O transactions due to demand paging.
- Different workloads have different characteristics and sensitivity to metrics such as Latency, Bandwidth and Capacity. To extract the value proposition for CXL memory expansion, the right system configuration must be set to optimize for the workload characteristic.

Next Steps :

- Further improvements in NUMA interleave policies will provide similar result as SW+ HW interleaving with more configurability .
- Application aware and optimized page allocation algorithms can further improve system performance by utilizing various memory tiers and media characteristics .
- CXL memory pooling and Fabric attached memory can help further in defining various memory tiers to reduce system TCO.

Thank You!

Acknowledgement :

Micron Team :- Vinicius Tavares Petrucci , Eishan Mirakhur , Nikesh Agarwal , Su Wei Lim