SNIA CMSI | COMPUTE, MEMORY, AND STORAGE

Persistent Memory, CXL, and Memory Tiering - Past , Present & Future

Live Webcast

June 27, 2023 10:00 am PT

SNIA Legal Notice

- The material contained in this presentation is copyrighted by the SNIA unless otherwise noted.
- Member companies and individual members may use this material in presentations and literature under the following conditions:
  - Any slide or slides used must be reproduced in their entirety without modification
  - The SNIA must be acknowledged as the source of any material used in the body of any document containing material from these presentations.
- This presentation is a project of the SNIA.
- Neither the author nor the presenter is an attorney and nothing in this presentation is intended to be, or should be, construed as legal advice or an opinion of counsel. If you need legal advice or a legal opinion please contact your attorney.
- The information presented herein represents the author's personal opinion and current understanding of the relevant issues involved. The author, the presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.

NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.

What Does SNIA Do?

- SNIA is a non-profit global organization dedicated to developing standards and education programs to advance storage and information technology.

Industry Leading Organizations

180

Active Contributing Members

2,500

IT End Users & Storage Pros Worldwide

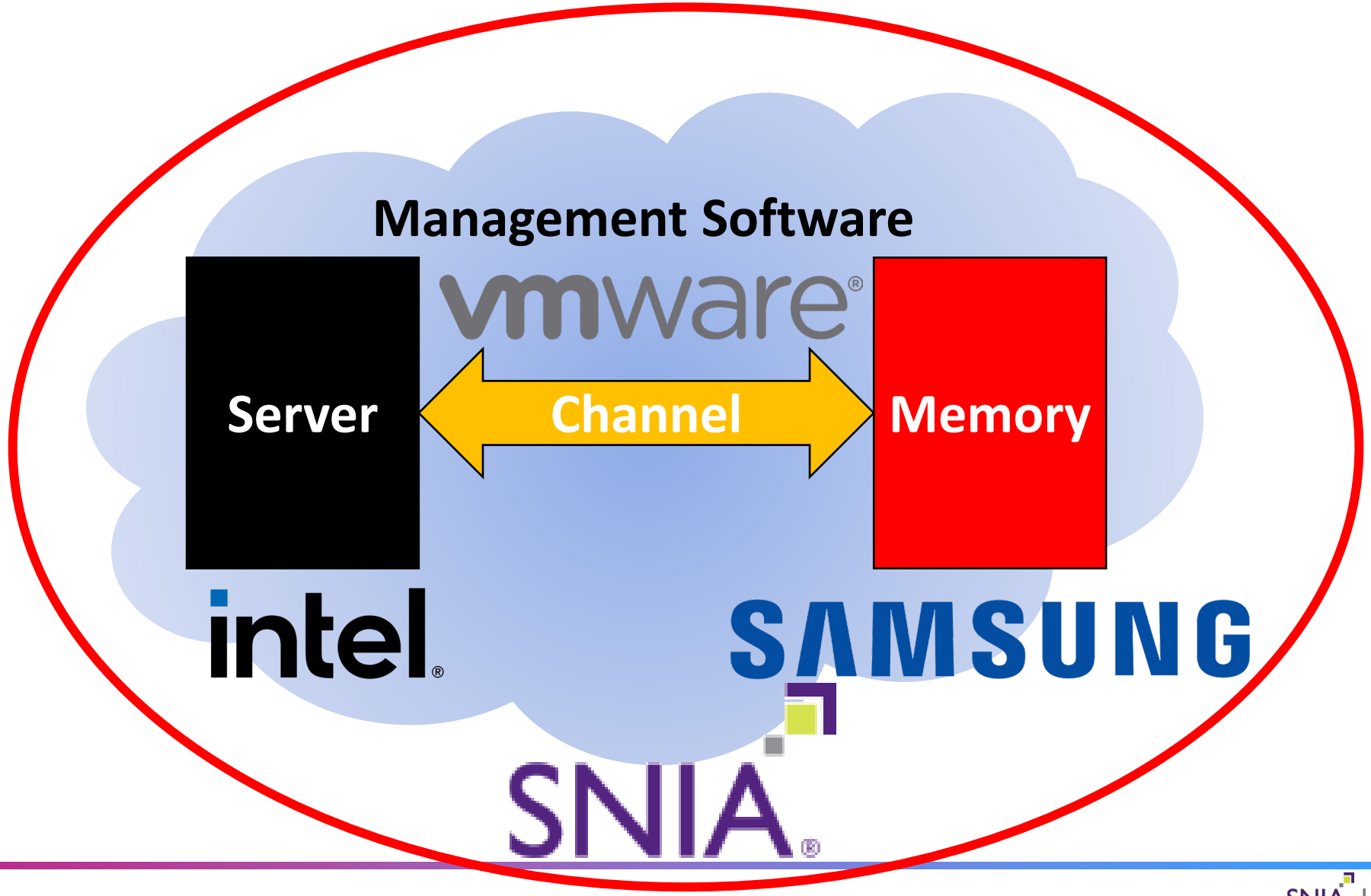
50,000

Who is CMSI?

- Part of SNIA, the SNIA Compute, Memory, and Storage Initiative is a community of storage professionals and technical experts who support:
  - The industry drive to combine processing with memory and storage
  - The creation of new compute architectures and software to analyze and exploit the explosion of data creation over the next decade
- CMSI's four Special Interest Groups – Computational Storage, DPU, Persistent Memory, and Solid State Drives – evangelize and educate on these technologies to the industry.

www.snia.org/cmsi

Our Panel



Our Presenters

Andy Rudoff
Panelist
Sr. Principal Engineer
Intel Labs

Bhushan Chitlur
Panelist
Sr. Principal Engineer
Intel Datacenter and AI

David McIntyre
Panelist
Director, Product Planning and Business Enablement
Samsung

Sudhir Balasubramanian
Panelist
Sr. Staff Architect & Global Oracle Practice Lead
VMware

Arvind Jagannath
Panelist
Product Line Manager for vSphere Platform,
VMware

What CXL Brings to the System

- Larger memory
  - Memory size is no longer limited by capacitance & power issues
    - CXL-attached memory will have longer latency
- Shared Memory
  - Processors can easily hand messages or even whole data sets back & forth
- Disaggregated Memory
  - Much more efficient use of available resources. No “Stranded” memory
- NUMA Support (Nonuniform Memory Architecture)
  - Volatile, Persistent, Slow, Fast – We take them all!

Andy Rudoff

Past & Present

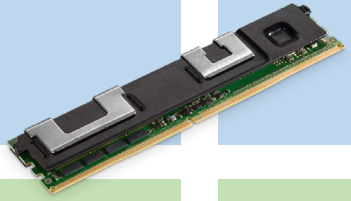
	Persistent	Volatile
Non-Transparent (Apps aware)	I PMem Programming Model	II NUMA interfaces Memory map file or device
Transparent (No app awareness)	III Storage Interfaces	IV HW Memory Tiering OS Memory Tiering Other variants on tiering

Past & Present

	Persistent	Volatile
Non-Transparent (Apps aware)	I PMem Programming Model 	II NUMA interfaces Memory map file or device
Transparent (No app awareness)	III Storage Interfaces	IV HW Memory Tiering OS Memory Tiering Other variants on tiering

Past & Present

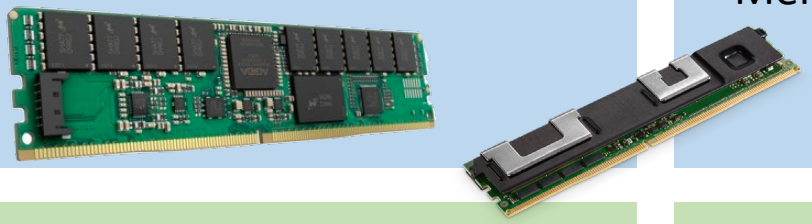
	Persistent	Volatile
Non-Transparent (Apps aware)	I PMem Programming Model	II NUMA interfaces Memory map file or device
Transparent (No app awareness)	III Storage Interfaces	IV HW Memory Tiering OS Memory Tiering Other variants on tiering



Past & Present

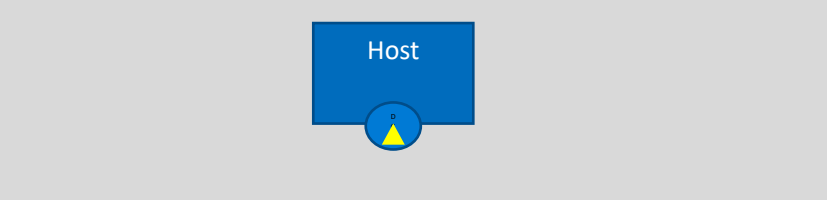
- Carry Forward to CXL-attached memory:
  - PMem Programming model
  - Memory Tiering
  - Tier Detection
    - HMAT
    - CDAT
  - Helper libraries
    - memkind
    - memkind2

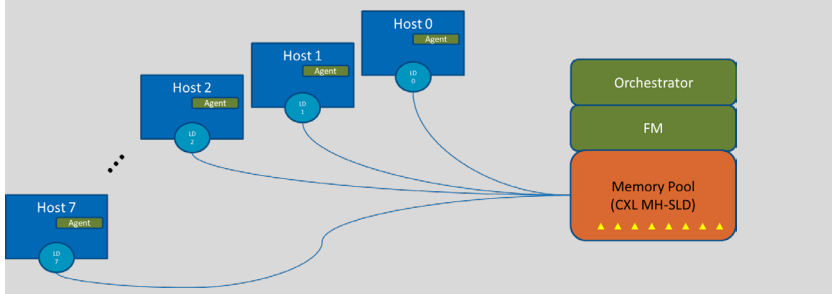
	Persistent	Volatile
Non-Transparent (Apps aware)	I PMem Programming Model	II NUMA interfaces Memory map file or device
Transparent (No app awareness)	III Storage Interfaces	IV HW Memory Tiering OS Memory Tiering Other variants on tiering

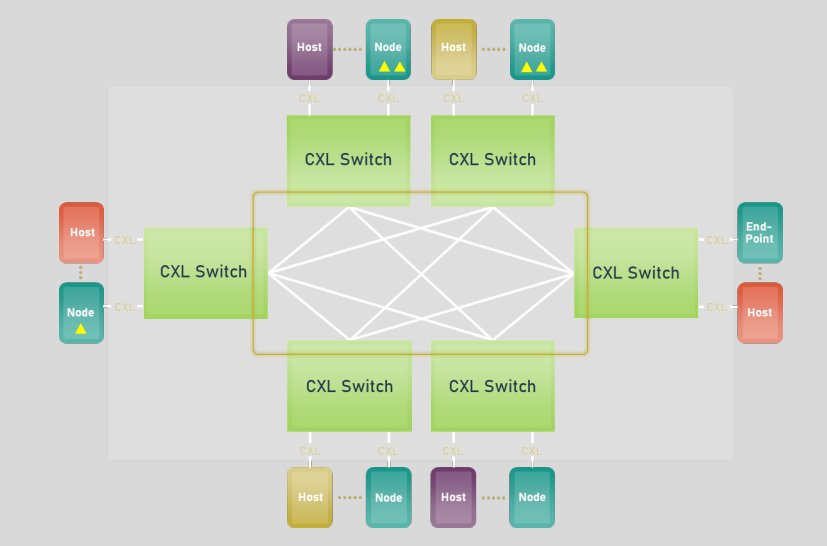


Future

- Memory Pooling
  - Host sees Dynamic Capacity Device (DCD)
  - Scale from 1 host to rack to data center
- Memory Sharing
  - Leveraging CXL 3.0 HW Coherency
- More interesting hybrid devices
  - Enabled by CXL flexibility
  - Near Memory Compute (NMC)



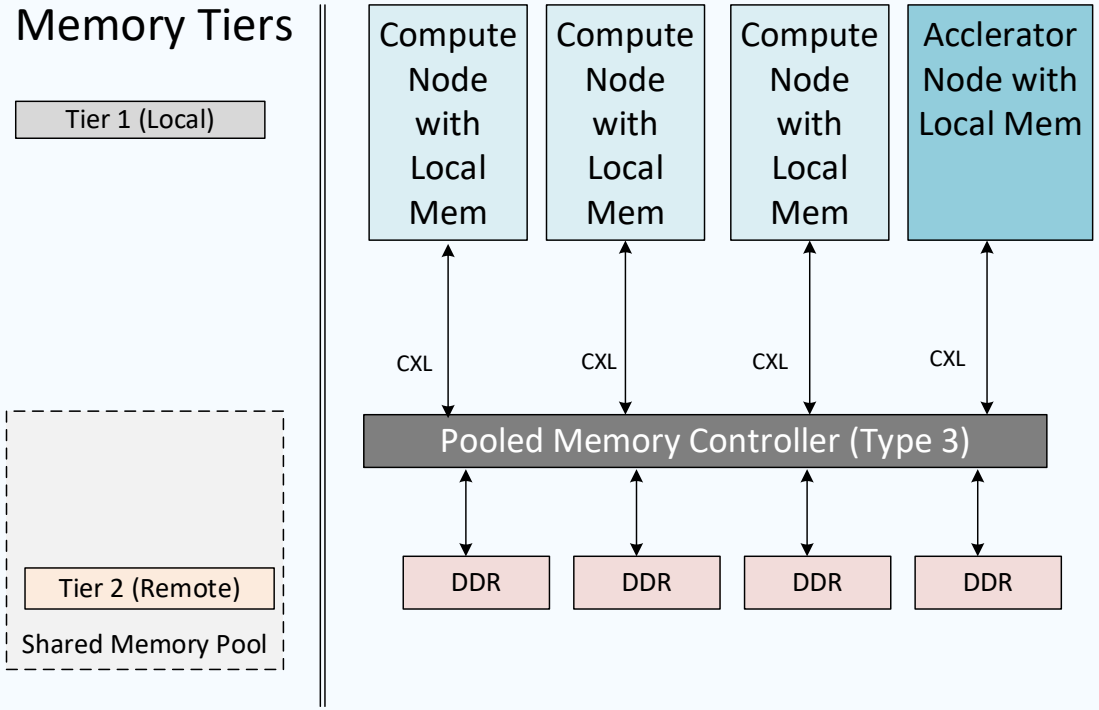




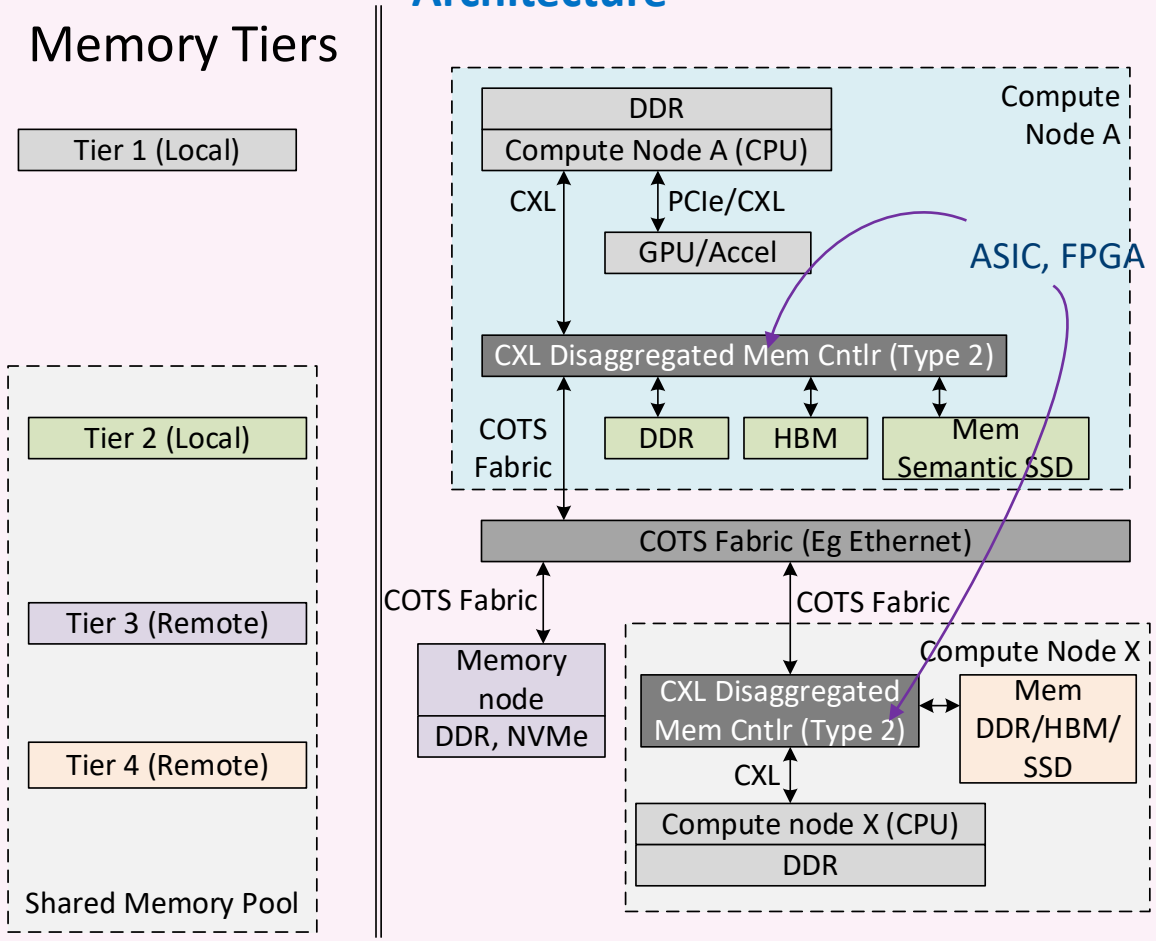
Bhushan Chithur

Embracing Heterogeneity and Disaggregated Memory Topologies

Centralized Disaggregated Memory Architecture



Heterogenous Disaggregated Memory Architecture



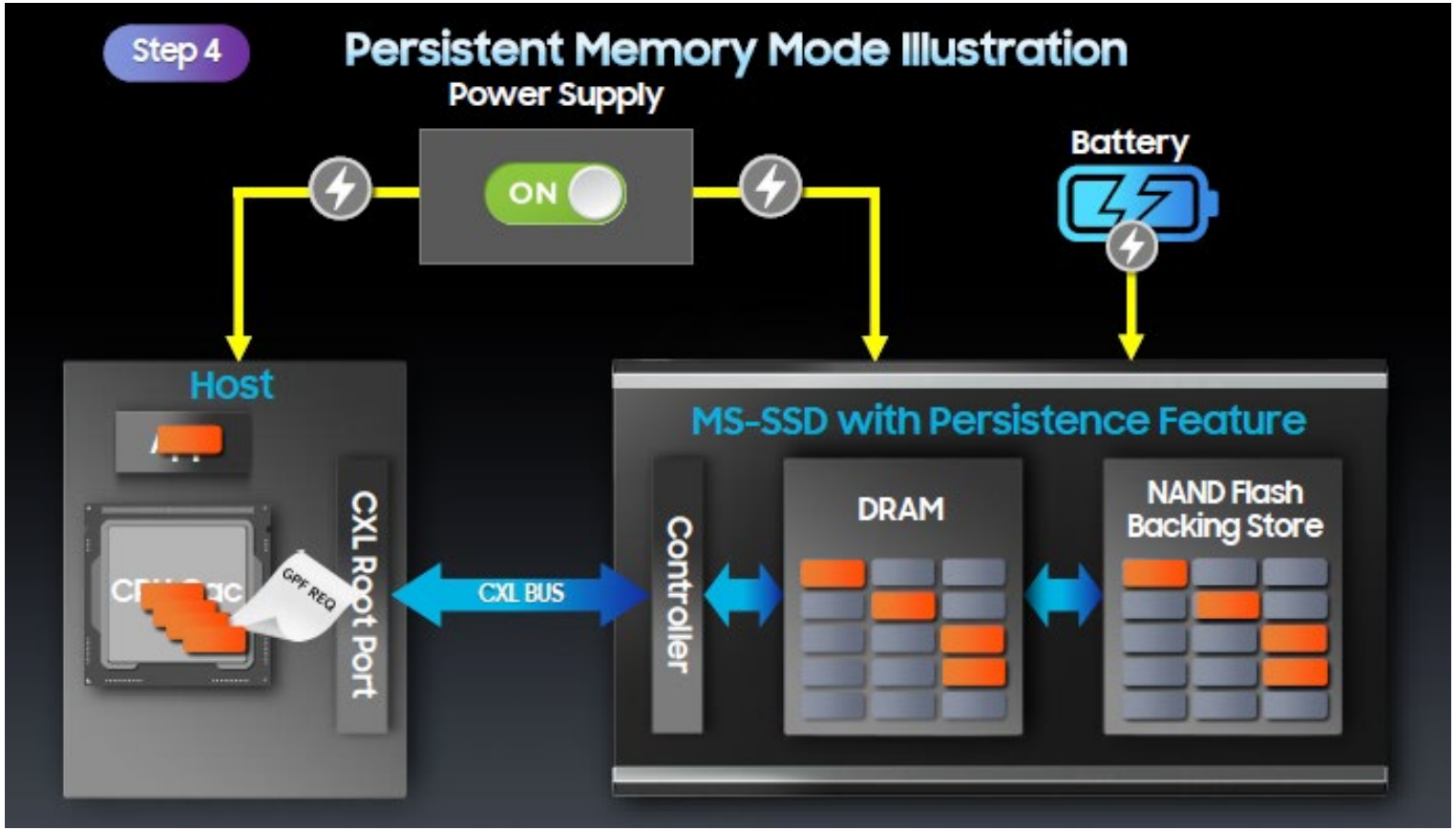
David McIntyre

Memory Class Solution Optimized for AI/ML

- Dual Mode Support
  - NVMe IO mode and CXL memory mode 20x greater 128-byte read performance
    - * Compared to PCIe Gen4 NVMe SSD
- Small Granularity Access
  - Min. 64-byte data transfers (fine/coarse grained access)
- Better System TCO
  - Larger capacity with NAND Flash
  - Lower latency with Internal DRAM cache

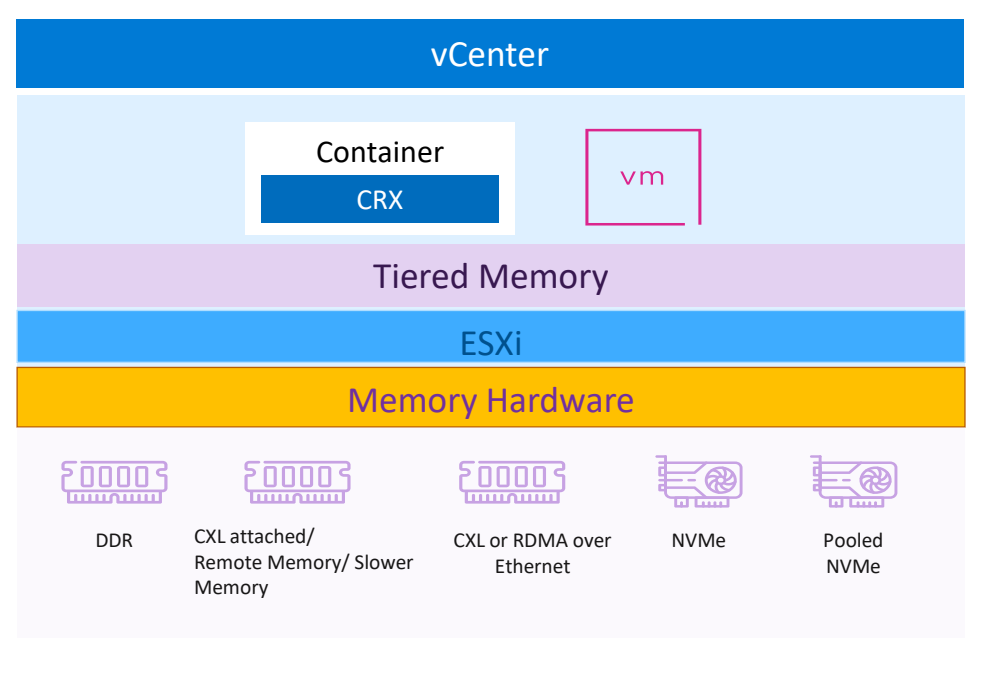
MS SSD: Persistence Mode

- New Persistence Applications
  - In-Memory Databases
  - Metadata
  - Transactional Logs
  - Lookup tables
- Capacity increasing
- MS SSD: 128GB+
- Linking MS SSDs together



Arvind Jagannath

VMware Memory Tiering



Benefits

- Higher Density - core utilization
- Lower TCO
- Larger bandwidth

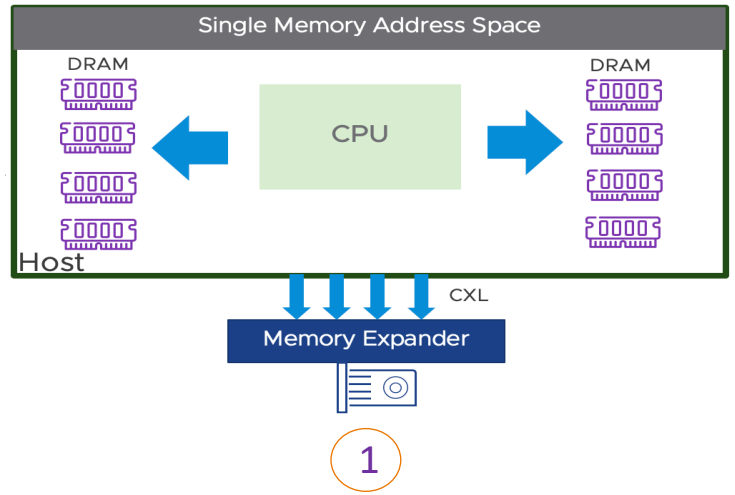
Value over traditional approaches

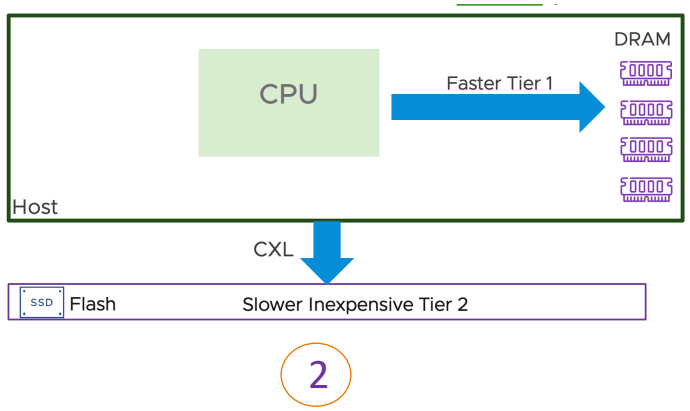
- Virtualization
  - Independent underlying hardware changes
- Transparent – Single volatile memory address
  - No Guest or Application changes
    - Run any Operating System
  - ESX internally handles page placement
- DRS and vMotion to mitigate risks
  - Tiering/device heuristics fed to DRS
- Ensure Fairness across workloads
  - Consistent performance
- Min Configuration changes
  - No special tiering settings
- Minimum Performance Degradation
- Processor specific monitoring
  - vMMR monitors at both VM- and Host-levels

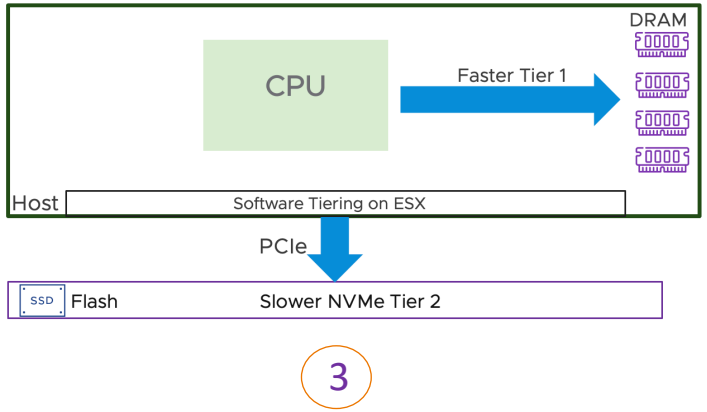
Key Use Cases emerging with CXL

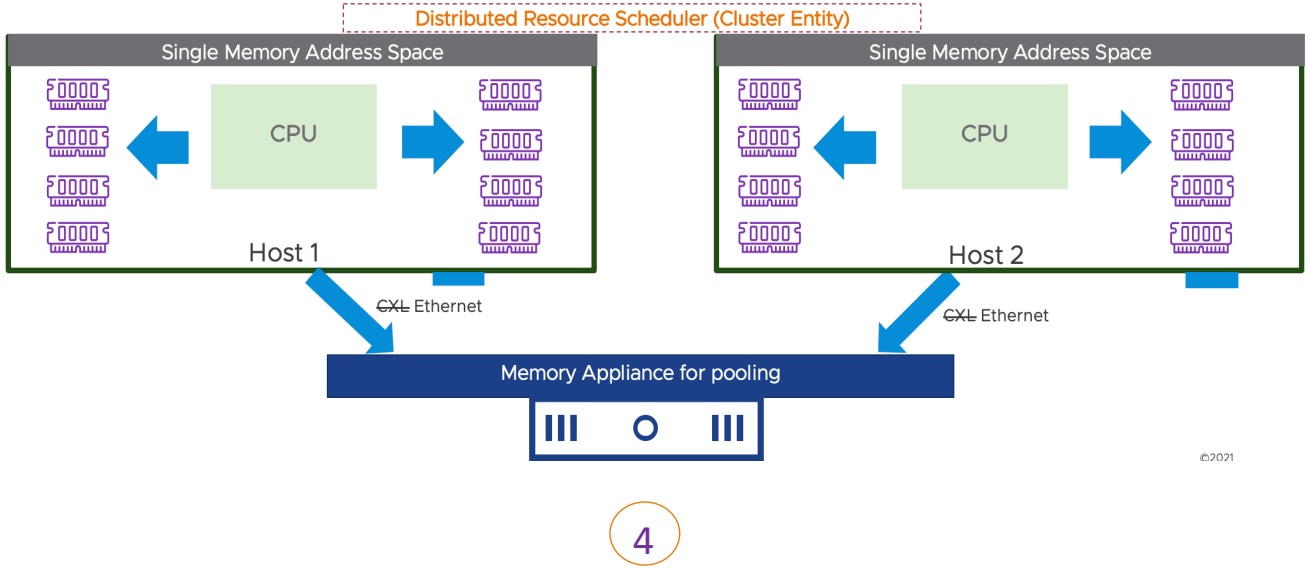
Memory Expansion with NUMA-like latencies	Memory Tiering	Memory pooling across hosts on a cluster using memory appliances	Memory sharing	CXL switching and shared access (future)
-Increase capacity/scale -Flat (non-tiered) expansion -Consolidate server memory -Improve bandwidth -Improve core utilization	Lower TCO – combinations of lower cost memory with DRAM	Consolidate memory usage on a cluster	Utilize stranded memory on hosts	Disaggregation and Composability

Deployment Options









Sudhir Balasubramanian

VMware Software Memory Tiering - NUMA and Memory Details

sc2esx64.vslab.local ACTIONS

Summary Monitor Configure Permissions VMs Resource Pools

Host Details

Hypervisor:	VMware ESXi
Model:	SYS-2049U-TR4
Processor Type:	Intel(R) Xeon(R) Platinum 8 260L CPU @ 2.40GHz
Logical Processors:	192
NICs:	10
Virtual Machines:	2
Memory Tiering:	Software
State:	Connected
Uptime:	10 days

Hardware

CPU	192 CPU(s) x Intel(R) Xeon(R) Platinum 8260L CPU @ 2.40GHz
Memory	4557.15 GB
Virtual Flash Resource	43.2 GB / 119.75 GB
Networking	10 Network(s)
Storage	5 Datastore(s)

Software Memory Tiering Server

- **Total Memory 4.5 TB**
  - 1.5 TB DRAM
  - 3 TB Tier 2 Memory

DRAM Mode Server

- **Total Memory 1.5 TB**
  - 1.5 TB DRAM

sc2esx65.vslab.local ACTIONS

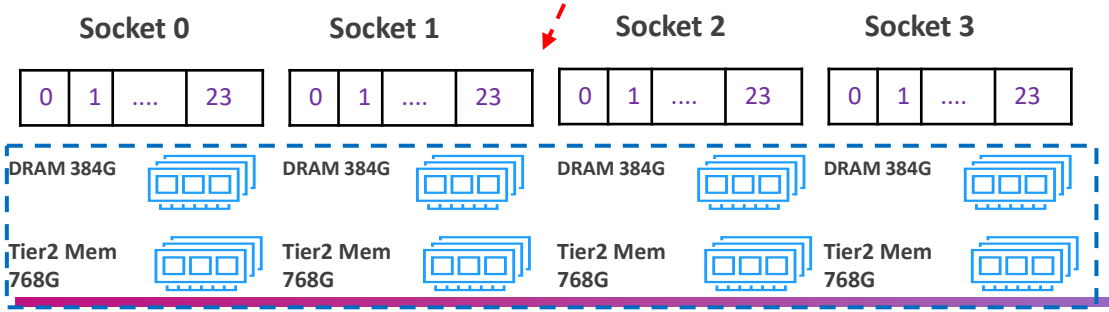
Summary Monitor Configure Permissions VMs Resource Pools

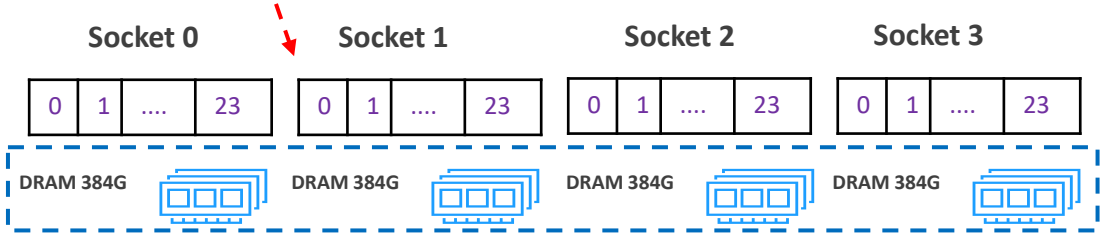
Host Details

Hypervisor:	VMware ESXi,
Model:	SYS-2049U-TR4
Processor Type:	Intel(R) Xeon(R) Platinum 826 0L CPU @ 2.40GHz
Logical Processors:	192
NICs:	10
Virtual Machines:	1
State:	Connected
Uptime:	18 days

Hardware

CPU	192 CPU(s) x Intel(R) Xeon(R) Platinum 8260L CPU @ 2.40GHz
Memory	1534.66 GB
Virtual Flash Resource	43.19 GB / 119.75 GB
Networking	10 Network(s)
Storage	5 Datastore(s)





Oracle Database – Details

- OEL 8.5 UEK ,Oracle 21.5
- Oracle Standalone DB , ASM & ASMLIB
- Oracle on VMware Best Practices Followed

Oracle Memory

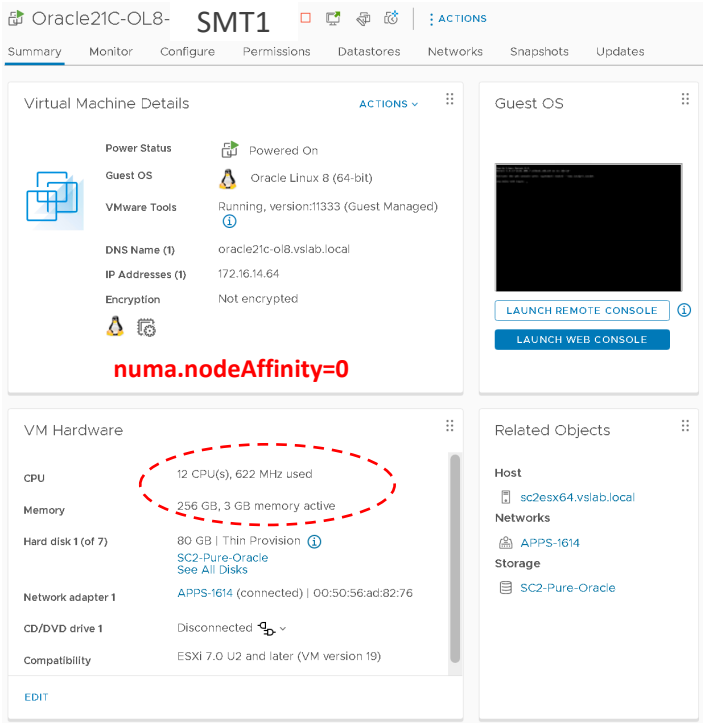
- VM
  - vCPUS=12, RAM = 256G
- Oracle
  - SGA =192G, PGA=30G

Oracle Memory

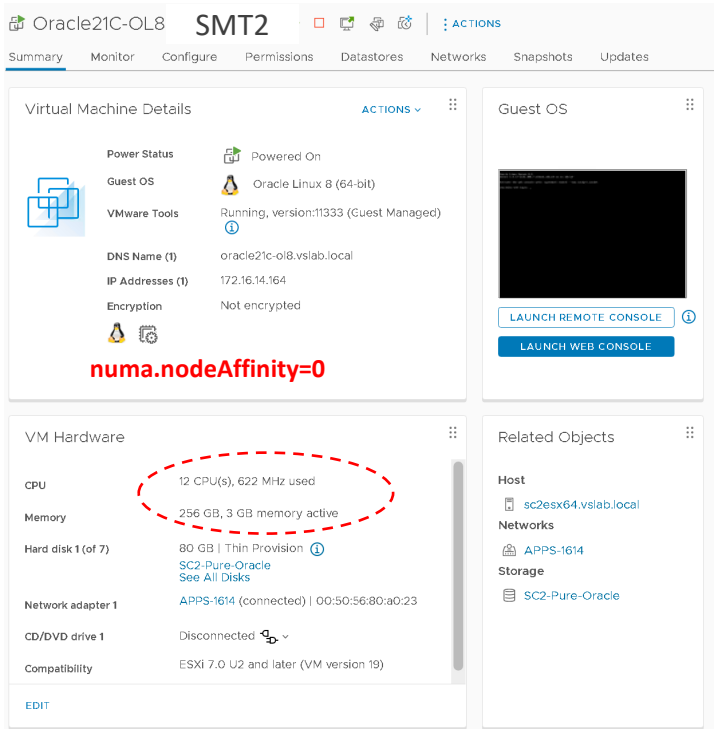
- VM
  - vCPUS=12, RAM = 256G
- Oracle
  - SGA =192G, PGA=30G

Oracle Memory

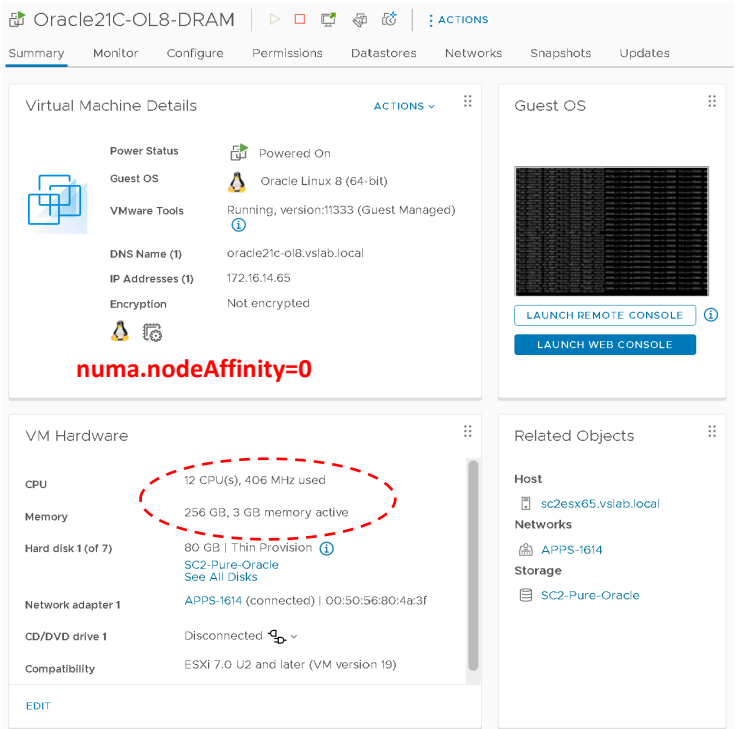
- VM
  - vCPUS=12, RAM = 256G
- Oracle
  - SGA =192G, PGA=30G



SW Memory Tiering VM – SMT1



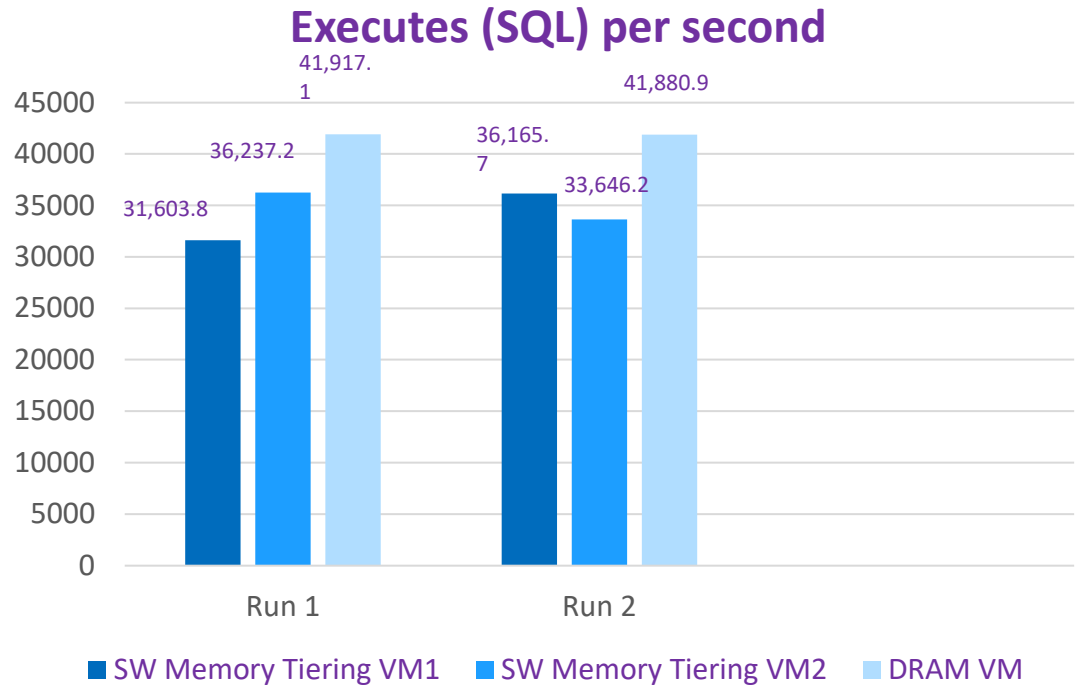
SW Memory Tiering VM – SMT2

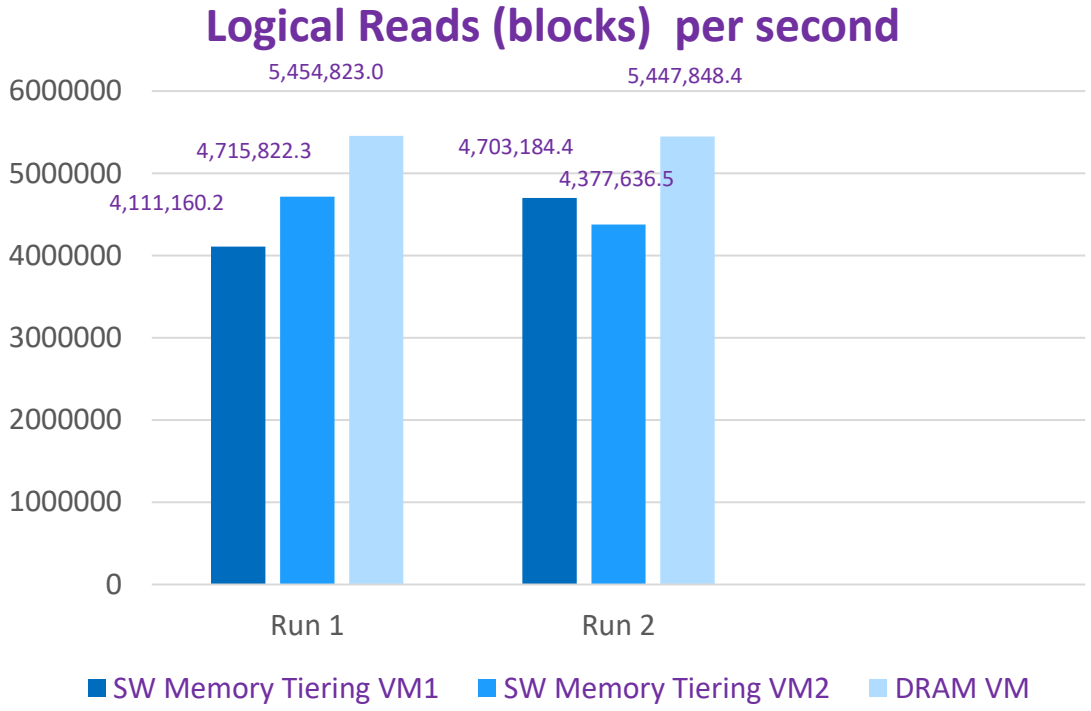


DRAM Mode VM1

Goal – Run '2 SW Memory Tiering' VM's on SMT Server on 1 NUMA node v/s '1 DRAM VM' on DRAM only Server on 1 NUMA node – Can we double our workload performance with lower TCO ?

Oracle Workload on SW Memory Tiering & DRAM Mode - Metrics





- Load Generator chosen as SLOB 2.5.4.0
  - UPDATE_PCT=0 (READ only test - performance comparison between SW Memory Tiering v/s DRAM Mode)
  - RUN_TIME=1200 secs(20mins)

- Test Results
  - **Executes(SQL) / second**
  - Run 1
    - **Aggregate SW Mem Tier VM1 + VM2 = 69,841/sec**
    - **DRAM Mode VM - 41,917.1/sec**
  - Run 2
    - **Aggregate SW Mem Tier VM1 + VM2 = 69,811.9/sec**
    - **DRAM Mode VM - 41,880.9/sec**

- Test Results
  - **Logical Reads (blocks) per second**
  - Run 1
    - **Aggregate SW Mem Tier VM1 + VM2 = 8,826,982.5/sec**
    - **DRAM Mode VM - 5,454,823.0/sec**
  - Run 2
    - **Aggregate SW Mem Tier VM1 + VM2 = 9,080,820.9/sec**
    - **DRAM Mode VM - 5,447,848.4/sec**

Attendee Actions

- Ask your questions via the Question Box!
- Please rate this webcast and provide us with feedback
- A Q&A from this webcast will be posted to the SNIA Compute, Memory, and Storage Blog
- Learn more:
  - **Visit us Live!**
    - **Flash Memory Summit, August 8-10, 2023, Santa Clara CA**
    - **SNIA Storage Developer Conference, September 18-21, 2023, Fremont CA**
  - **Online**
    - This webcast and many other videos and presentations on today's topics are in the SNIA Educational Library
    - View our SNIA YouTube playlists on **CXL** and **Memory**
- Join SNIA and the Persistent Memory Special Interest Group
  - www.snia.org/join
  - https://www.snia.org/technology-focus/persistent-memory

Questions?

Thank you for attending!

Follow us on Twitter @sniacmsi

Learn more at www.snia.org/educational-library

