SNIA DEVELOPER CONFERENCE
SDC 25
By Developers FOR Developers

Hyatt Regency Santa Clara, CA
September 15-17, 2025

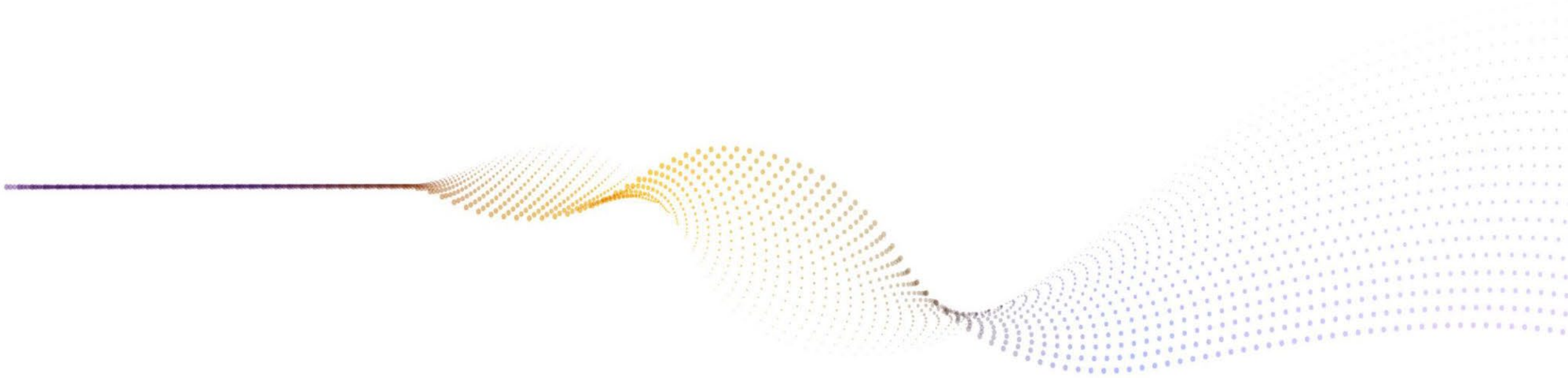
Compute Express Link (CXL) as Scalable and highly cost effective Memory architecture for Modern Workloads

Pramod Peethambaran

Director of Engineering, Data Fabric Solution,
MSL, Samsung Semiconductors Inc.
https://semiconductor.samsung.com/about-us/locations/us-rnd-labs/memory-labs/data-fabric-solutions/

Contents

- Introduction
- Memory Expansion - Current Challenges
- Revisiting Memory Hierarchy with CXL
- Recap CXL
- Samsung Cognos Memory Orchestration
- Performance enhancing value add feature
- Test results
- Conclusion & future work



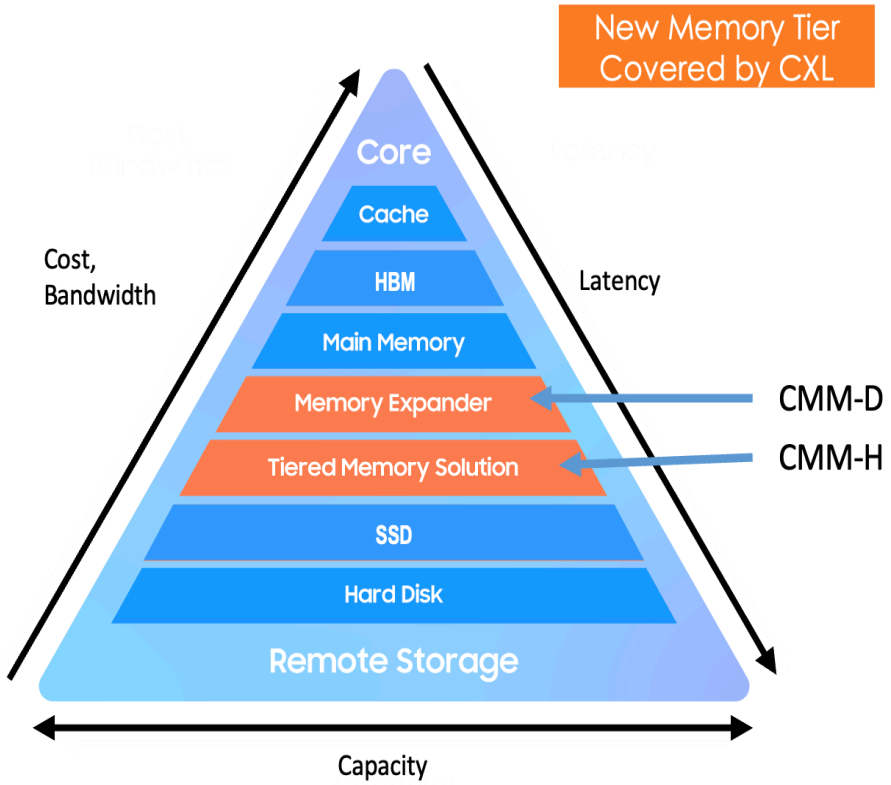
Introduction

Memory Expansion - Current Challenges

- CPU memory channel limitations
  - Physical limitations of CPU memory channels - # of pins, signal integrity and noise, Routing complexity, etc.
  - Higher Manufacturing Cost
- Inefficient scale-out
  - Adding more nodes for memory – higher cost for network, Compute and space
- Memory Stranding
  - Dedicated memory – inefficient resource utilization

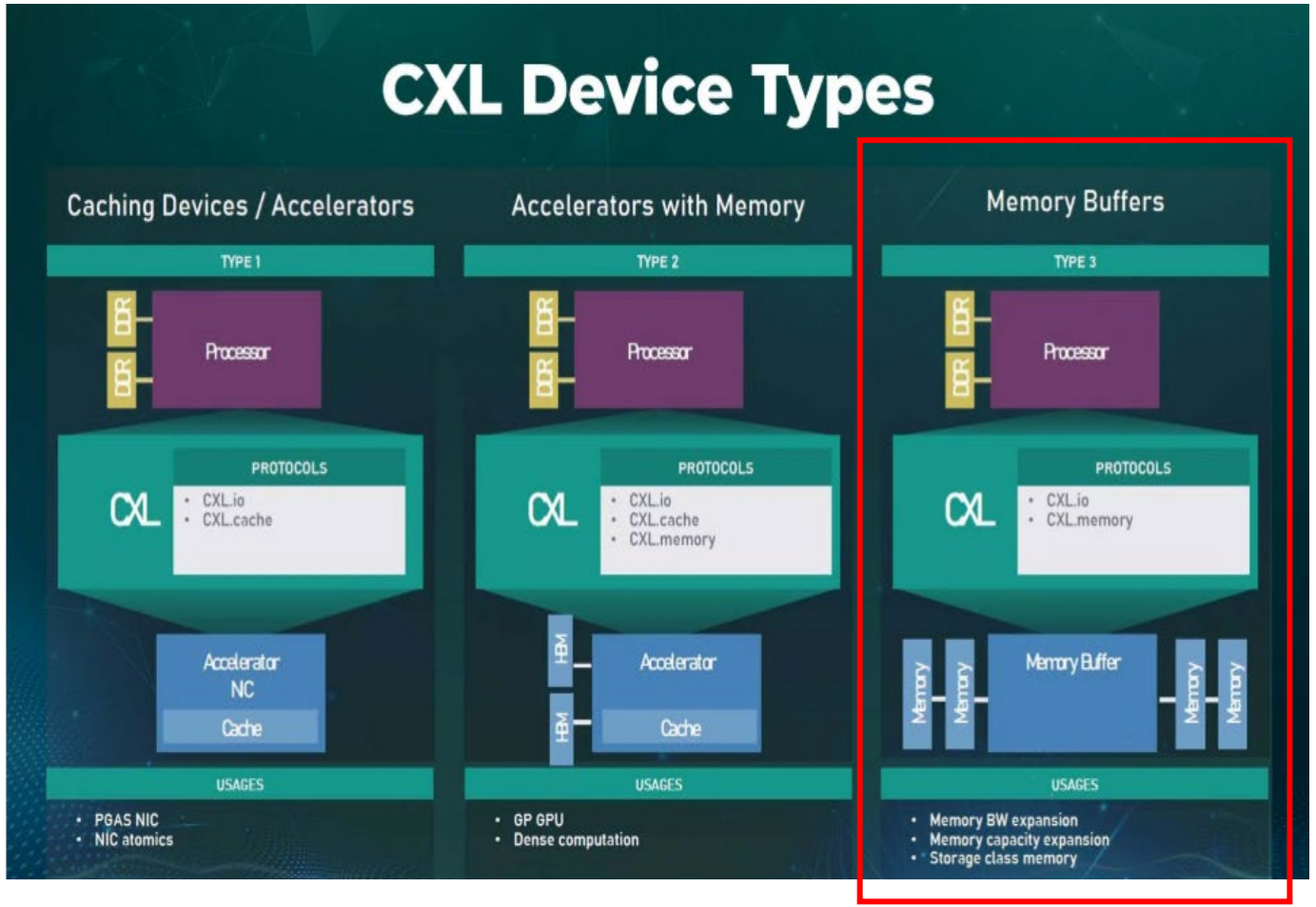
Revisiting Memory Hierarchy with CXL

- Fastest memory on the Top of the pyramid
- Highest capacity on the bottom of the pyramid
- Cost increases as we move towards the top
- CXL comes in the middle balancing cost and performance

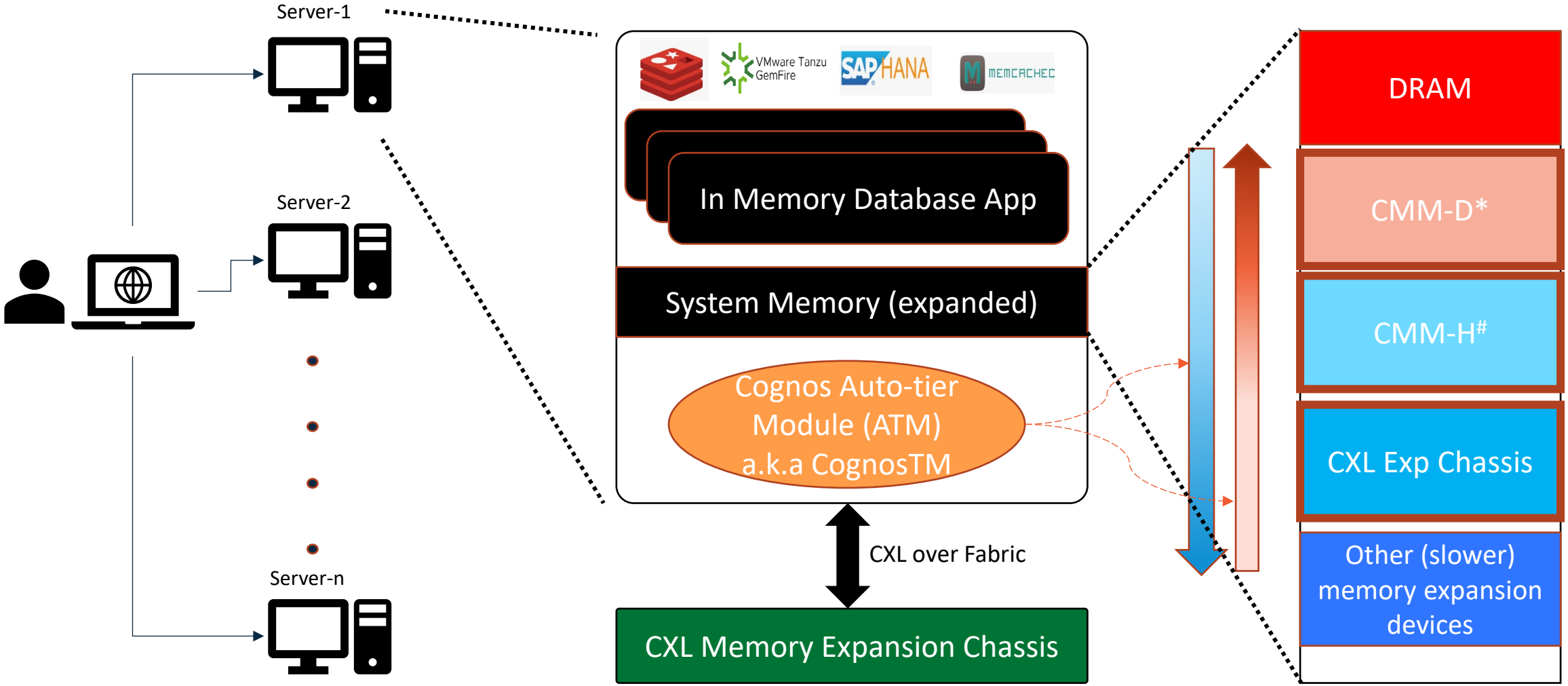


Recap CXL

- Open Standard, built on PCIe as the physical interface
- Protocols
  - CXL.IO
  - CXL.Cache
  - CXL.mem
- Device Types
  - Type-1
  - Type-2
  - Type-3



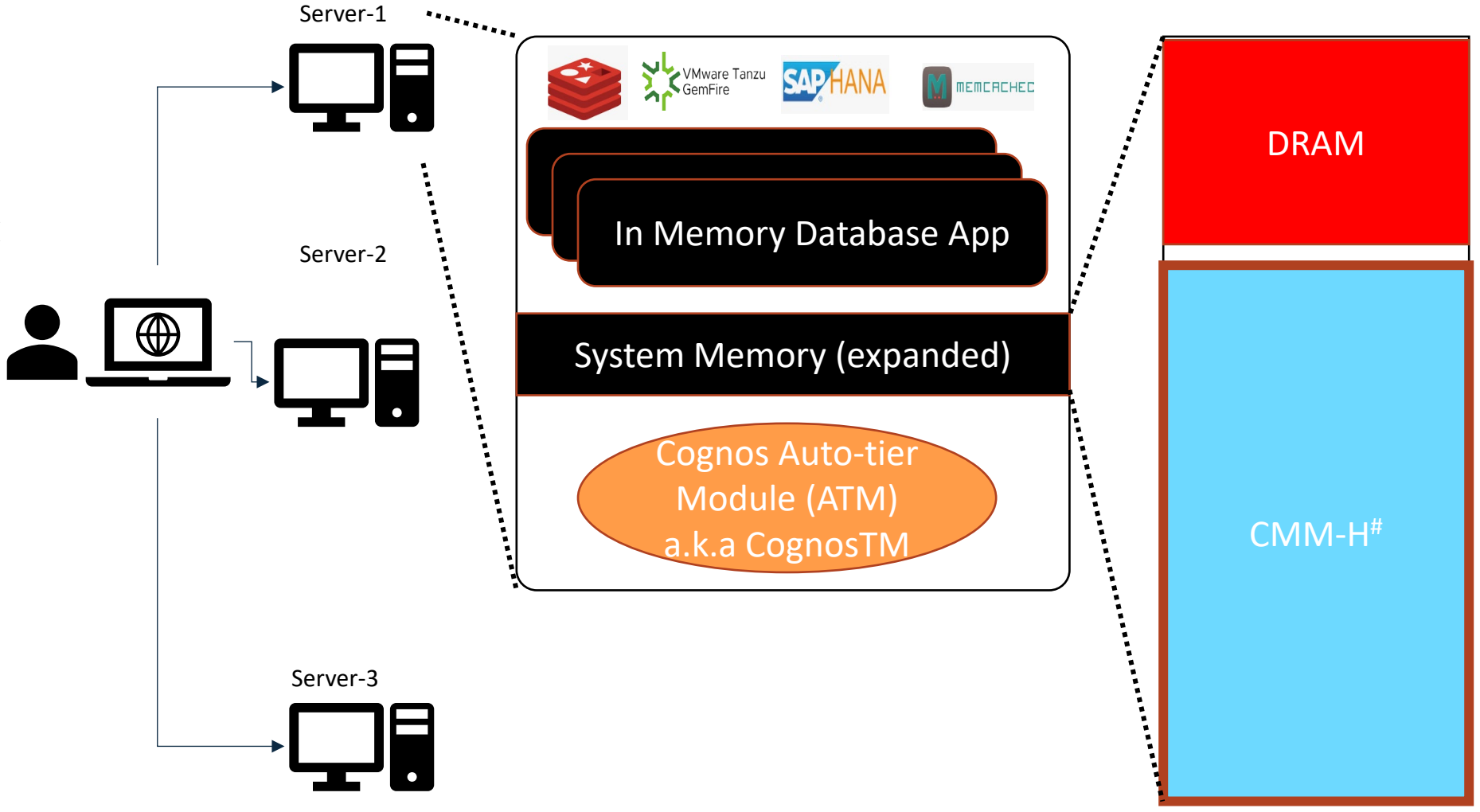
Samsung Cognos Memory Orchestration/Tiering – IMDB as Application (General view)



** Samsung DRAM-Only CXL device, # Samsung Hybrid (DRAM and NAND) CXL Research device*

Overall Goal and Solution

- Goal
  - **30% TCO reduction**
  - **100 K Ops/sec**
  - **<1ms p99 latency**
- Solution
  - **CMM-H (HW)**
  - **Auto-tier Module (SW)**
  - Analytics (SW)
  - RAS (SW/HW)



 # Samsung Hybrid (DRAM and NAND) CXL Research device

Benchmarking set-up and Configuration

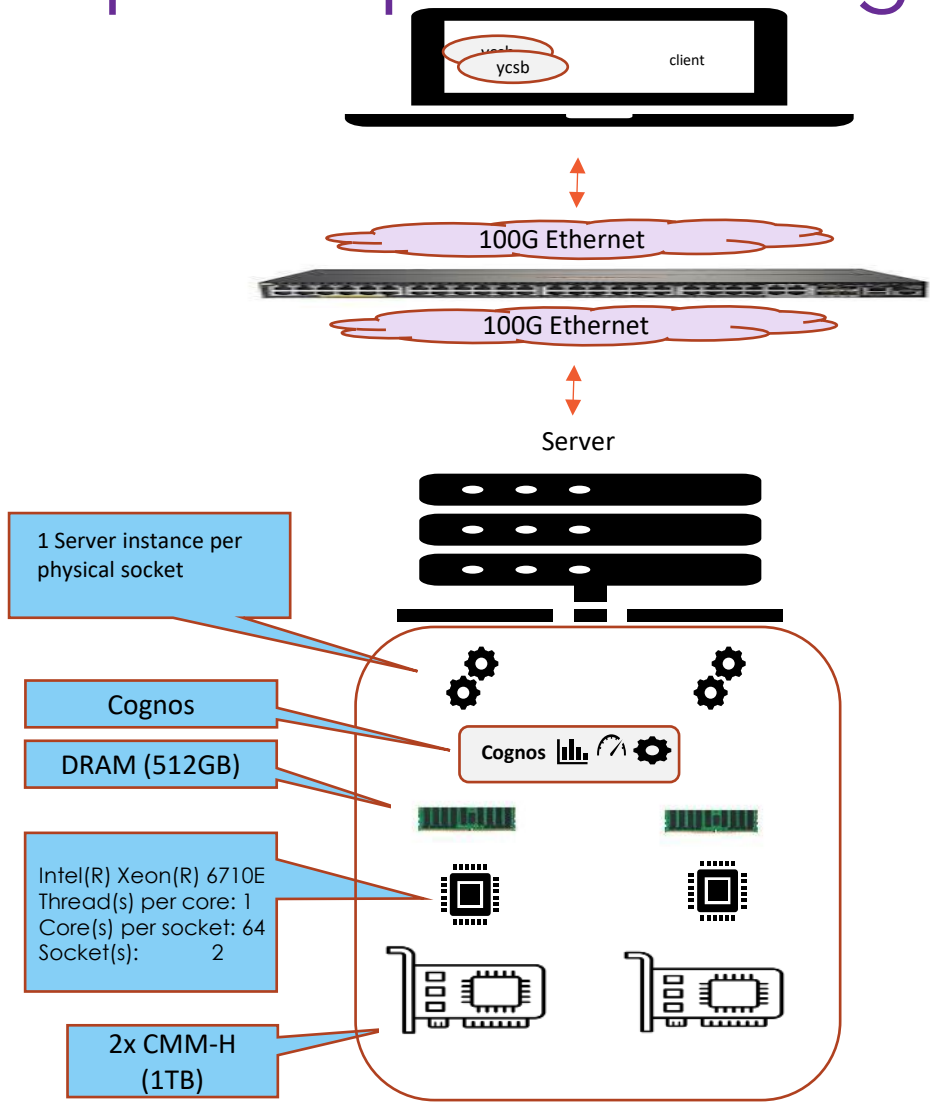
	Single node		3-node	
	Baseline-1	**Cognos-1**	**Baseline-2**	**Cognos-2**
❑ Hardware				
# of nodes	1	1	3	3
CPU	128 Core, Dual socket(64 core each), Intel® Xeon® 6710E	128 Core, Dual socket(64 core each), Intel® Xeon® 6710E	128 Core, Dual socket(64 core each), Intel® Xeon® 6710E	128 Core, Dual socket(64 core each), Intel® Xeon® 6710E
DRAM	512GB	512GB	1.5TB	1.5TB
CXL Memory	NA	2TB, 4TB#	NA	6TB, 12TB#
Network	2x100G	2x100G	2x100G	2x100G
❑ System				
OS	Ubuntu 22.04	Ubuntu 22.04	Ubuntu 22.04	Ubuntu 22.04
❑ Workload (YCSB)*				
Key:Value Size	400:700 B	400:700 B	400:700 B	400:700 B
Read:Write Ratio	2:1	2:1	2:1	2:1
Redundancy	OFF	OFF	1	1
Total workload size	160GB	800GB, 1.5TB#	160GB	800GB, 1.5TB#

** for all cases - Persistence and Overflow is OFF, 1 IMDB server Instance per NUMA node*

 # Config marked GREY to be available later

Single node benchmark setup component diagram

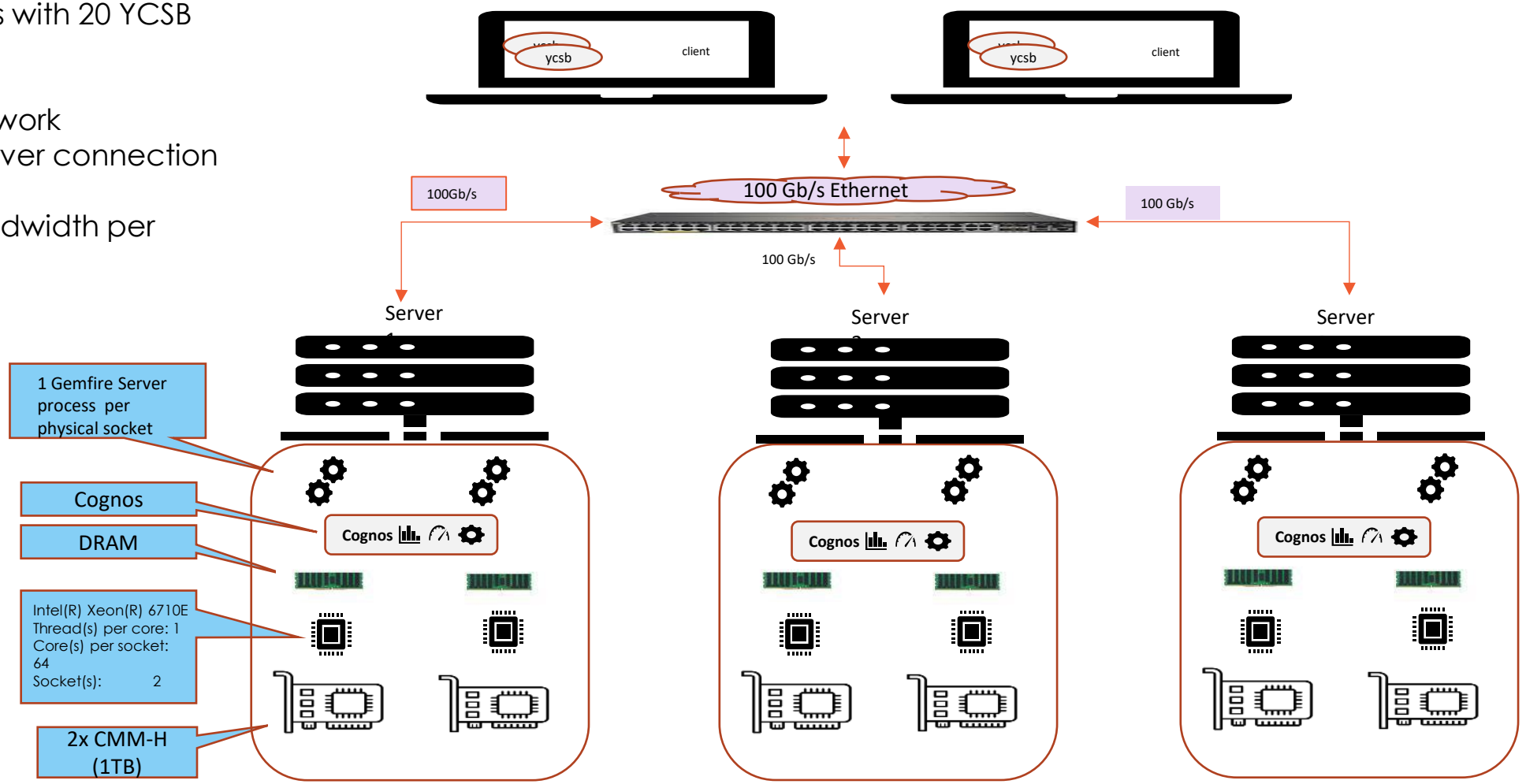
- 1 physical client machines with 10 YCSB instances with 10 threads
- Available **2 * 100Gb/s** network bandwidth per client – server connection
- Used **<1GB/s** network bandwidth per server node.



 # Samsung Hybrid (DRAM and NAND) CXL Research device

3 node Benchmark setup component diagram

- 2 physical client machines with 20 YCSB instances with 10 threads
- Available **2 * 100GB/s** network bandwidth per client – server connection
- Used **<1GB/s** network bandwidth per server node.



 # Samsung Hybrid (DRAM and NAND) CXL Research device

Benchmarking Setup & Workload details

CPU Configuration

Architecture: x86_64
CPU op-mode(s): 32-bit, 64-bit
Address sizes: 52 bits physical, 48 bits virtual
Byte Order: Little Endian
CPU(s): 128
On-line CPU(s) list: 0-127
Vendor ID: GenuineIntel
Model name: Intel(R) Xeon(R) 6710E
CPU family: 6
Model: 175
Thread(s) per core: 1
Core(s) per socket: 64
Socket(s): 2
CPU max MHz: 3200.0000
CPU min MHz: 800.0000
BogoMIPS: 4800.00
Caches (sum of all):
L1d: 4 MiB (128 instances)
L1i: 8 MiB (128 instances)
L2: 128 MiB (32 instances)
L3: 192 MiB (2 instances)
NUMA:
NUMA node(s): 4
DRAM:
NUMA node0 CPU(s): 0-63
NUMA node1 CPU(s): 64-127
CXL:
NUMA node2 CPU(s): N/A
NUMA node3 CPU(s): N/A

Memory Configuration

Type	Size	Numa Node
DRAM	256G	0
DRAM	256G	1
CXL (2.0)	1T	2
CXL (2.0)	1T	3

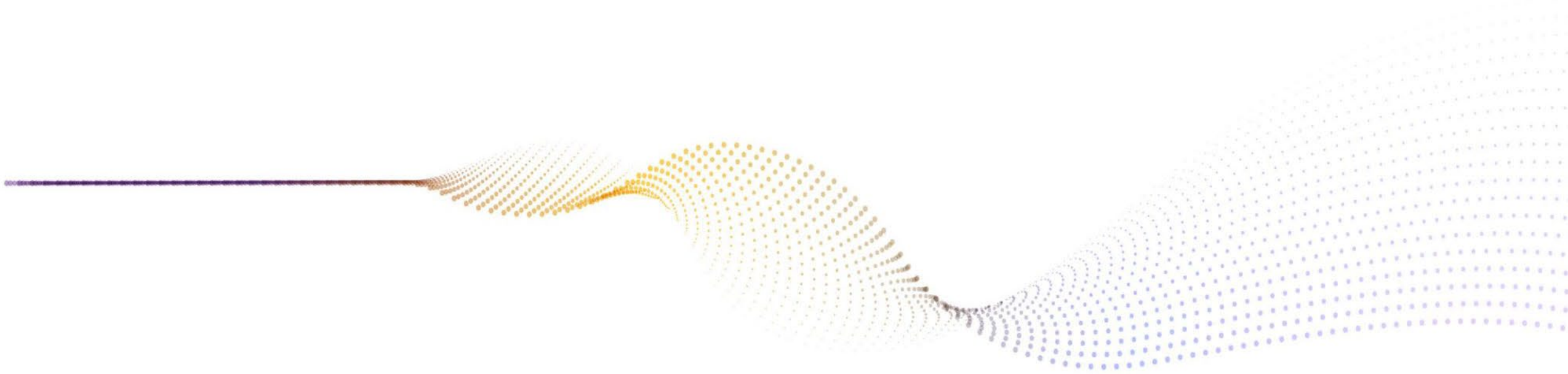
Network

NIC per CPU Socket: 1 * 100 Gb/s, dual ported

Workload Configuration

IMDB Application:
servers per CPU socket: 1
Key:value size: 400:700 (1.1KB)
Total data size: 160G, 800G
Region Persistence: No

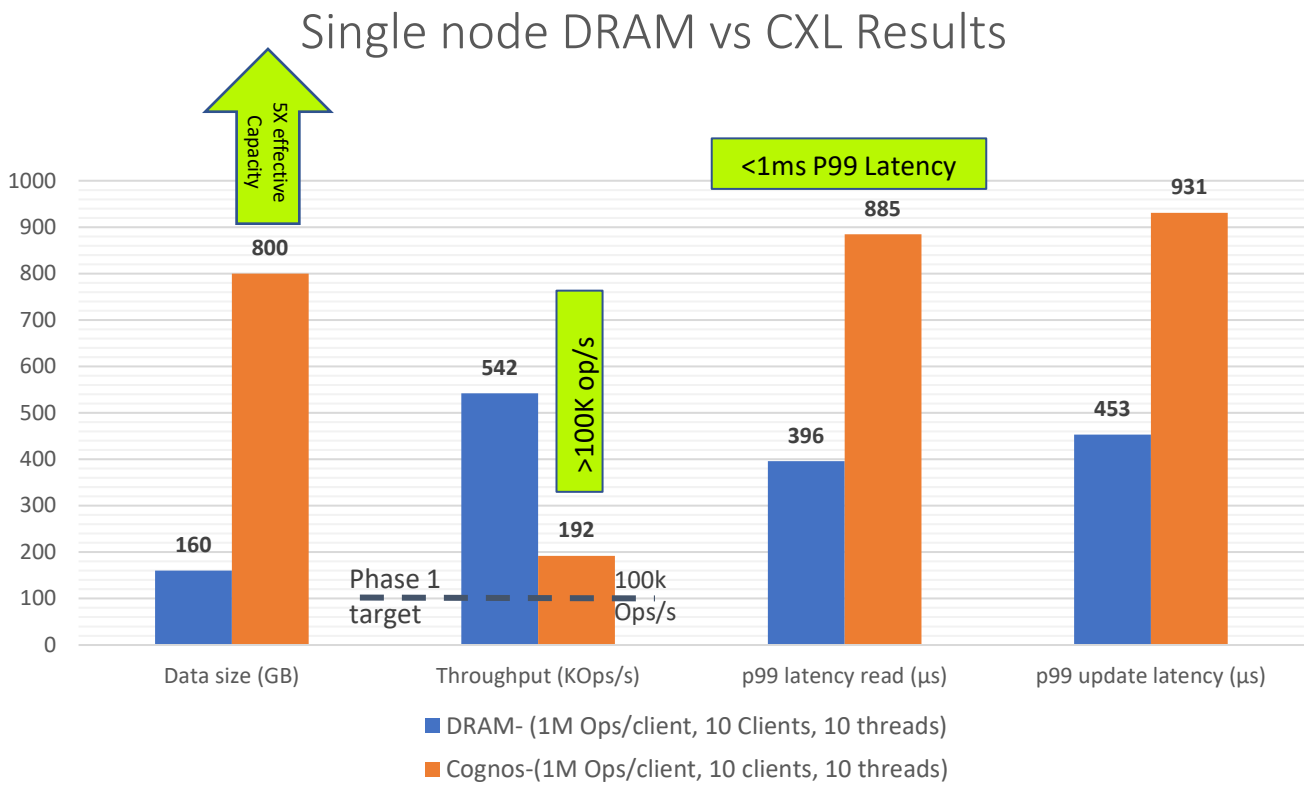
YCSB:
physical nodes: 1
no.of instances: 10
threads per instance: 5, 10
workload ratio (read:update): 66:33
record count (per ycsb instance): 80000000
operation count (per ycsb instance): 200000
no. of iterations per test suite: 32

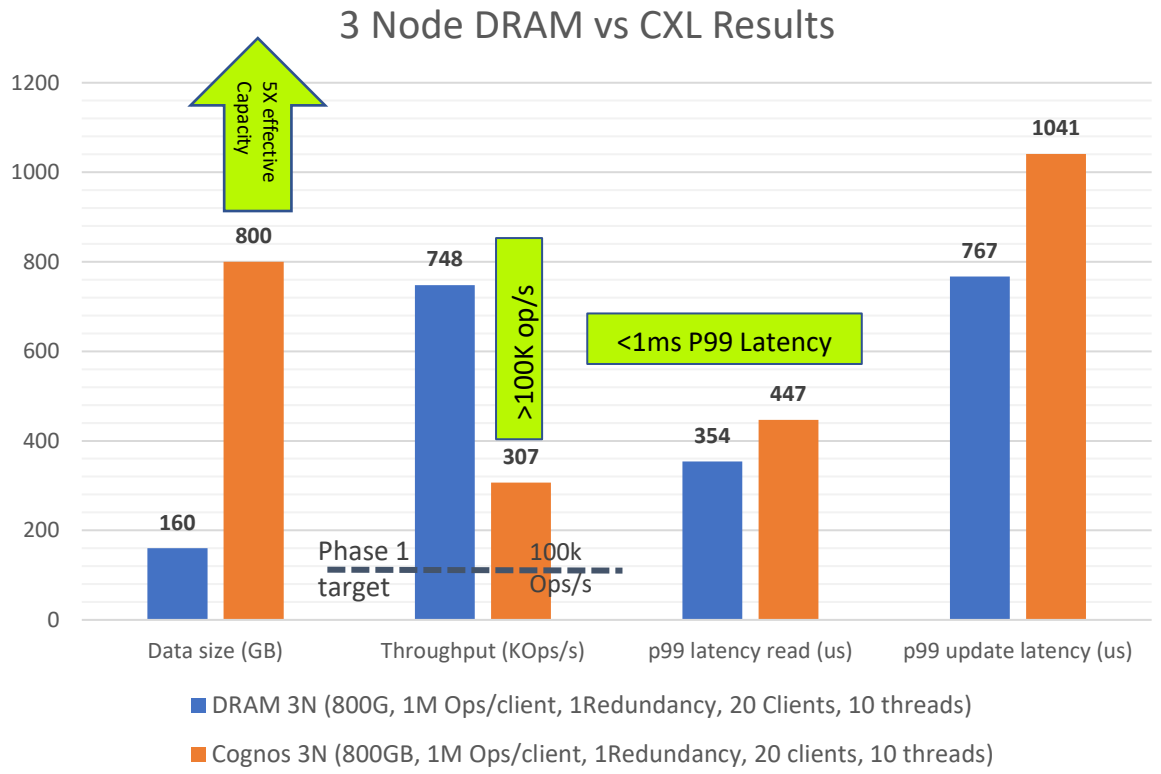


Test Results

Key Takeaways - Highlights

- Usable Memory capacity increased **~5X** (TCO improvement) with CMM-H + CognosTM
- Achieved required SLA of >100K op/s throughput and <= 1ms P99 latency per operation when scaled from single to multinode configuration

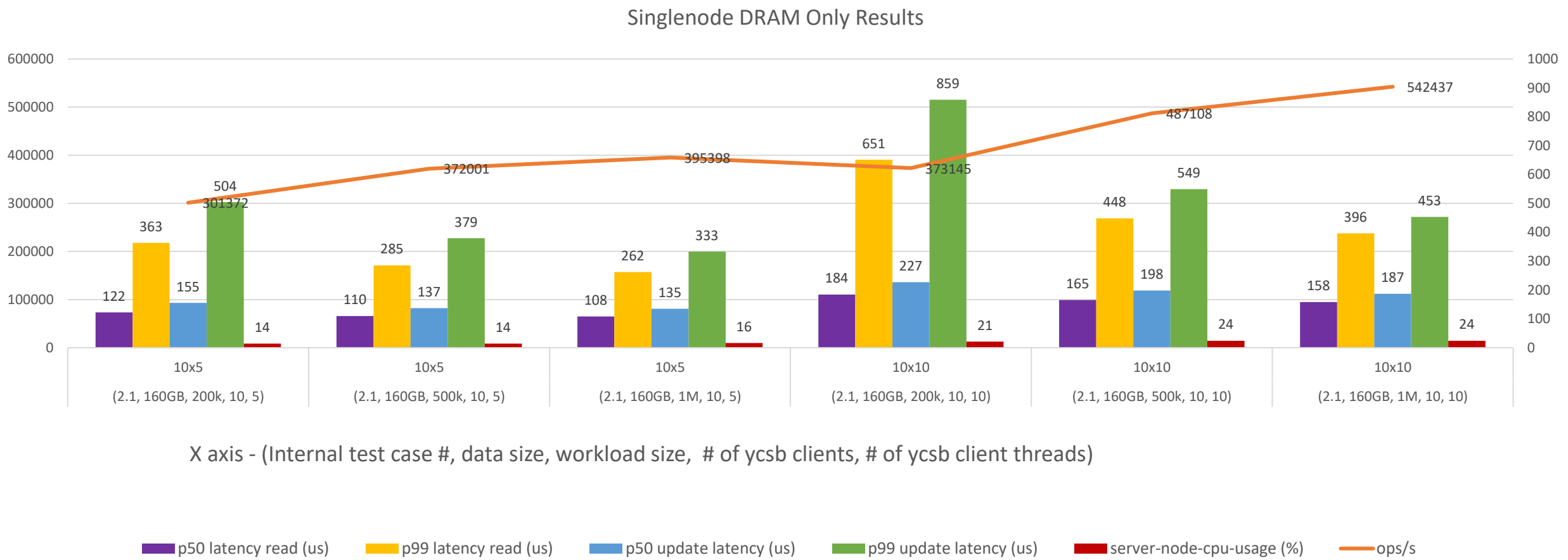




Single node Baseline (DRAM only)

Key Takeaways:

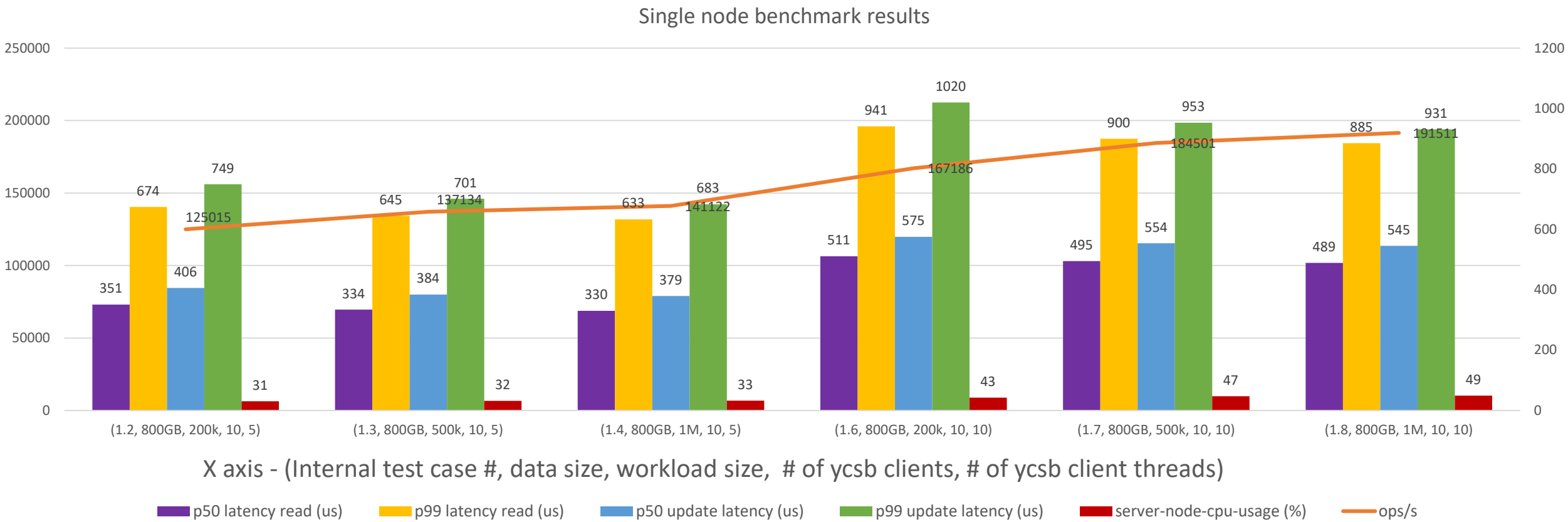
- >500K op/s and < 1ms latency



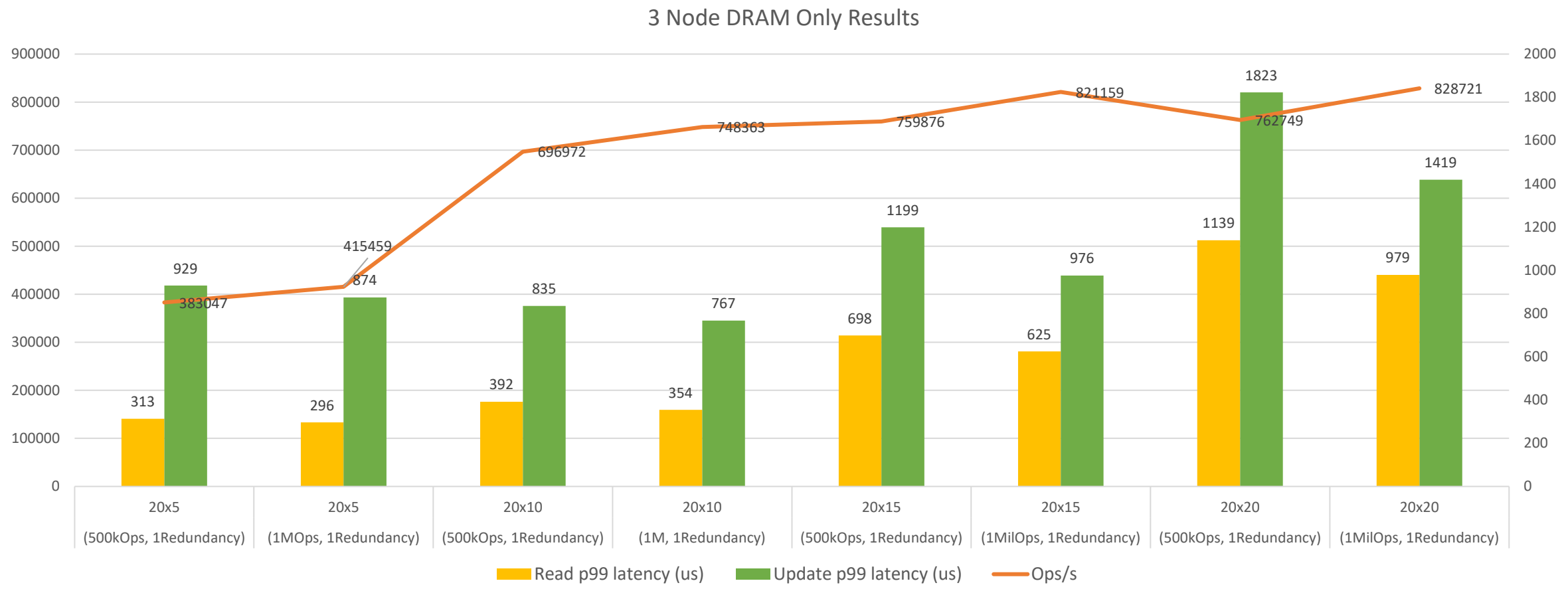
Single Node with Cognos(with CXL)

Key Takeaways:

- Memory capacity increased ~**5X**
- Although ~2.5 times effective degradation against DRAM with Gen 1 CXL device,
  - Achieved >180K op/s throughput and < 1ms P99 latency per operation



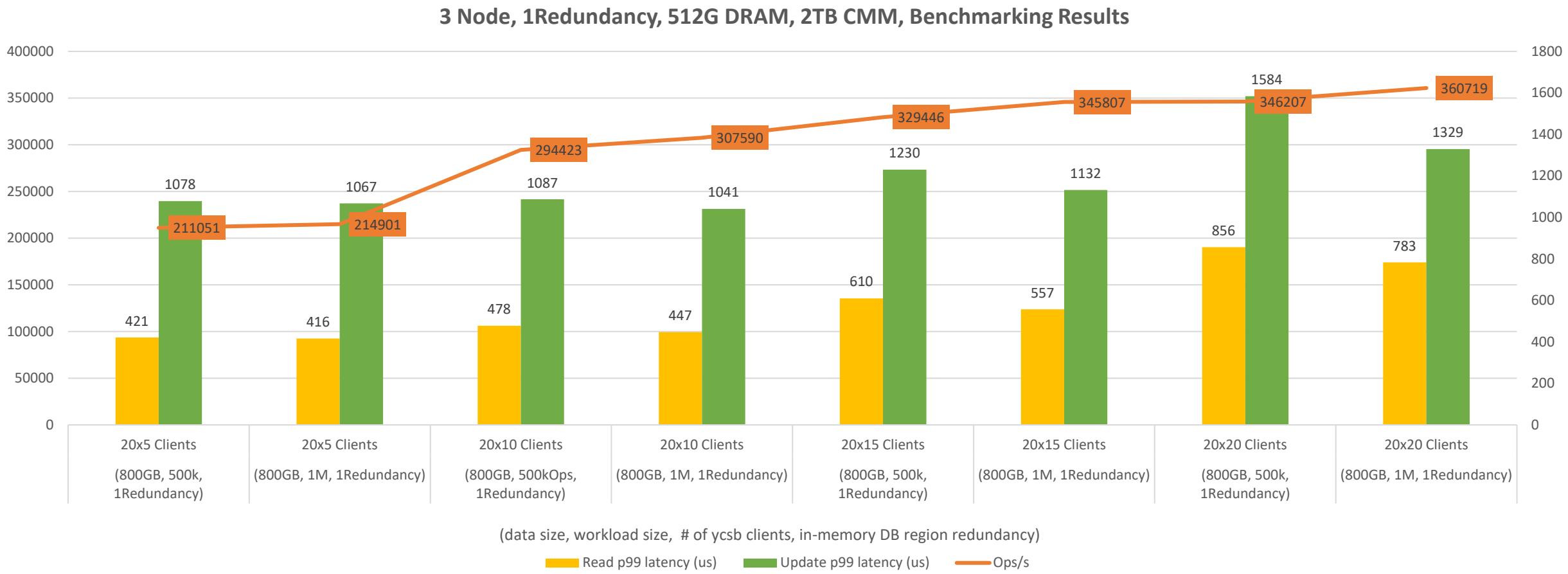
3 Node Baseline (DRAM only)



3 Node with Cognos (With CXL)

Key Takeaways:

- Memory capacity increased ~**5X**
- Achieved >100K op/s throughput and ~1ms P99 latency per operation with Gen 1 CXL device

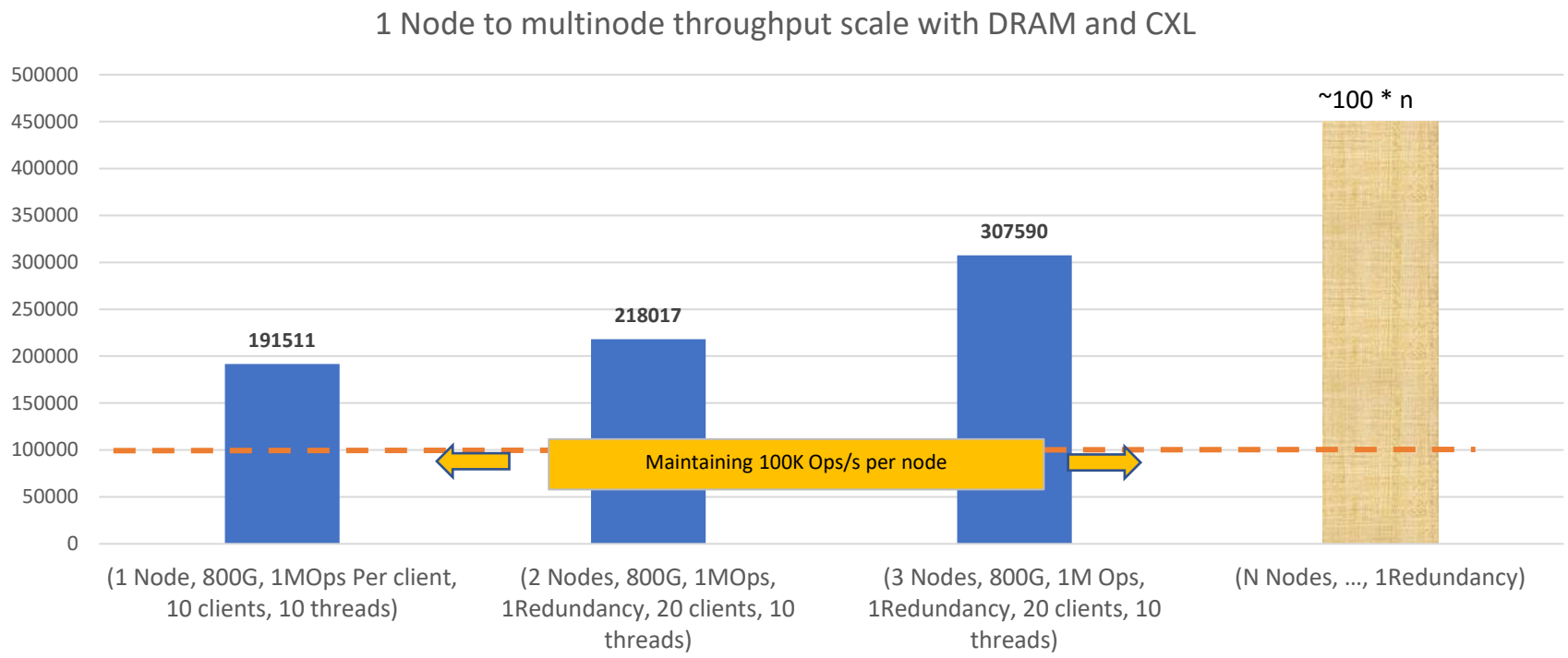


Horizontal scaling extrapolation

Horizontal Scaling (more nodes to cluster)

Based on the benchmarking results for 1, 2 and 3 nodes, following SLA parameters are expected to hold:

- throughput > 100K Ops/s per node, where GemFire redundancy parameter is set to 1
- end to end <=1ms P99 latency per node, where GemFire redundancy parameter is set to 1

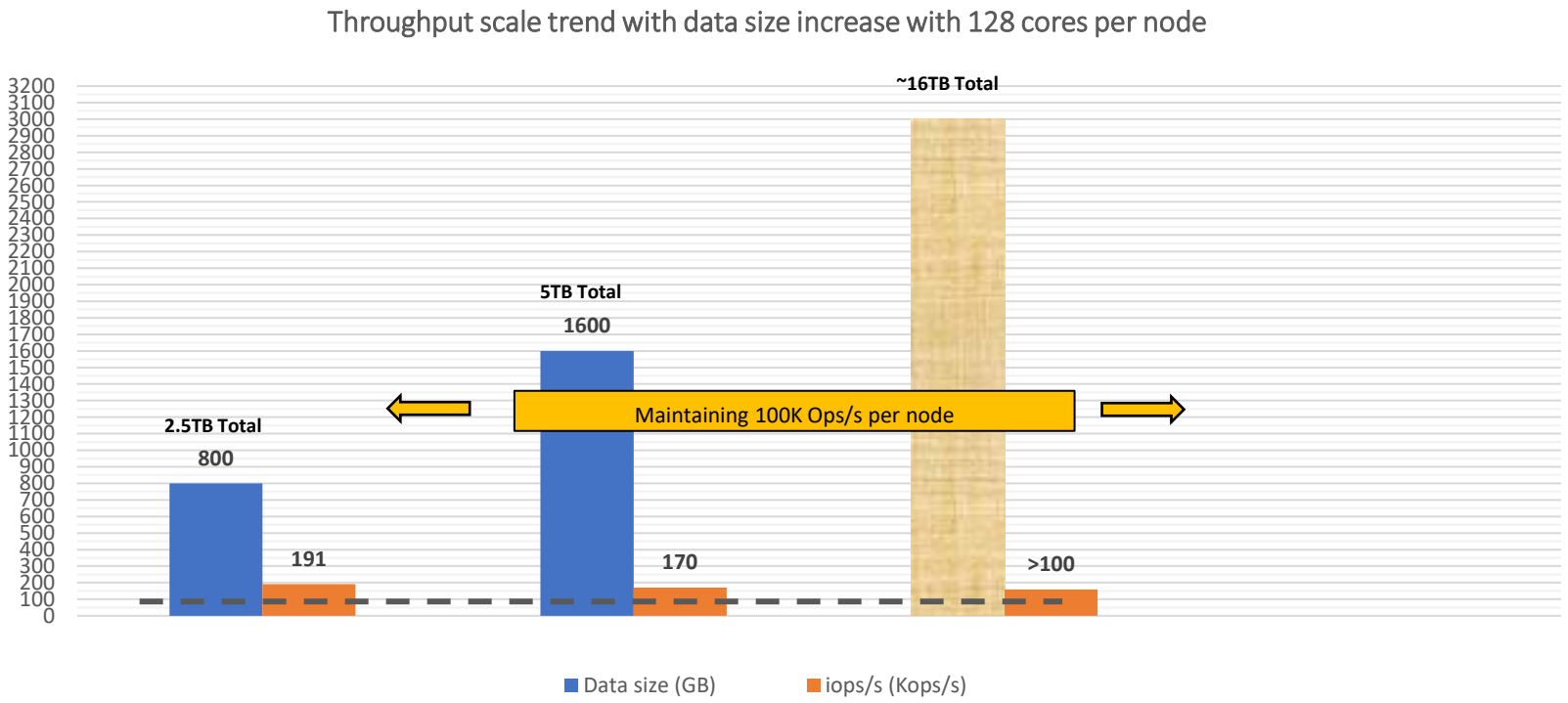


Vertical scaling extrapolation

Vertical Scaling (more memory per node)

Based on the benchmarking results for 160G, 800G & 1600G data size, following SLA parameters are expected to hold:

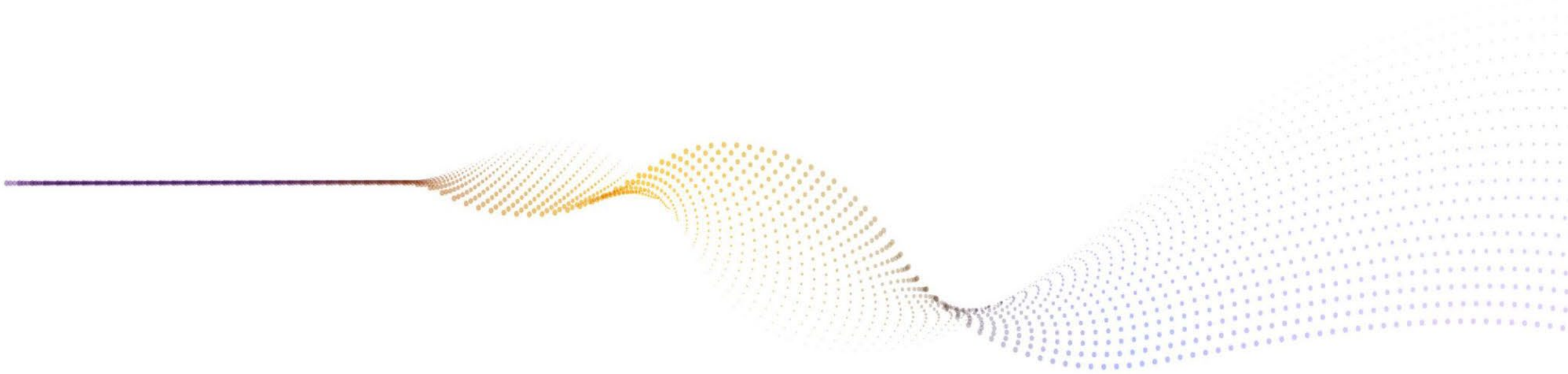
- throughput > 100K Ops/s per node, where GemFire redundancy parameter is set to 1
- end to end <=1ms P99 latency per node, where GemFire redundancy parameter is set to 1



Conclusions and Future work

- Adding CXL memory along with Cognos Auto-tiering is able to increase effective capacity per node by 5x. This reduces # of nodes required for a particular workload, thus **reducing TCO**.
- Memory bound Applications like IMDB, are able to **maintain the expected SLA** of 100K Ops/s and <1ms P99 latency, with higher capacity CXL memory

- Future Work (few, but not limited to)
  - Data Pre-fetch to improve the effectiveness of tiering
  - Device Pooling and Sharing to still reduce the TCO
  - Cognos support for orchestrating memory interleaving based on the Application targeted SLA.
  - Tiering for multiple tiers of heterogenous memory



Thank you for attending!

Please remember to rate this session. You get access the presentations at
http://sniadeveloper.org/conference

https://semiconductor.samsung.com/about-us/locations/us-rnd-labs/memory-labs/data-fabric-solutions/

