OCP FUTURE TECHNOLOGIES SYMPOSIUM

OCP Global Summit

November 8, 2021 | San Jose, CA

Big Memory @VMware (Capitola)

Renu Raman

OCTO

vmware®

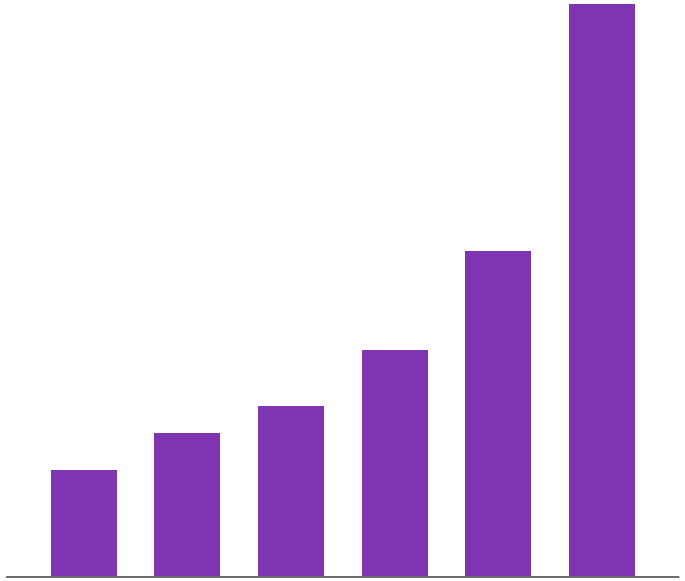
©2021 VMware, Inc.

Disclaimer

- This presentation may contain product features or functionality that are currently under development.
- This overview of new technology represents no commitment from VMware to deliver these features in any generally available product.
- Features are subject to change, and must not be included in contracts, purchase orders, or sales agreements of any kind.
- Technical feasibility and market demand will affect final delivery.
- Pricing and packaging for any new features/functionality/technology discussed or presented, have not been determined.

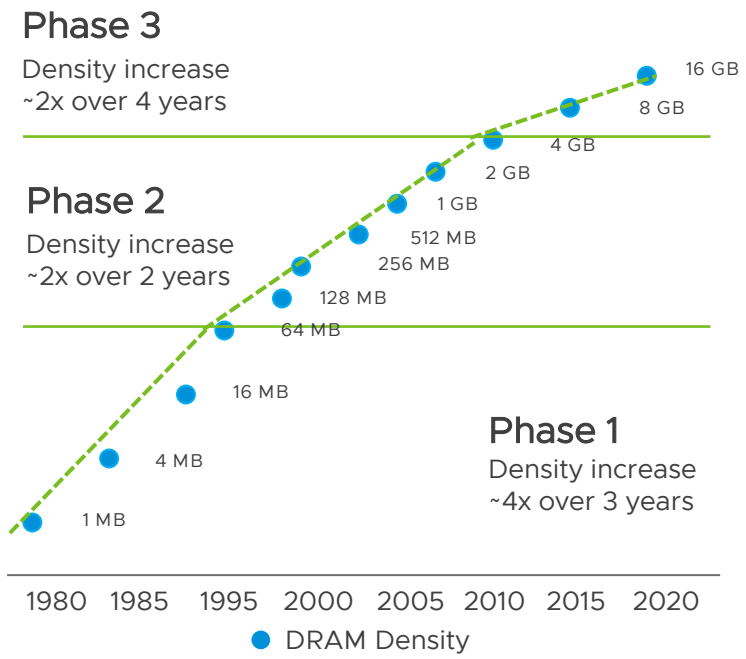
Compute, Memory & Data

Compute demand accelerating

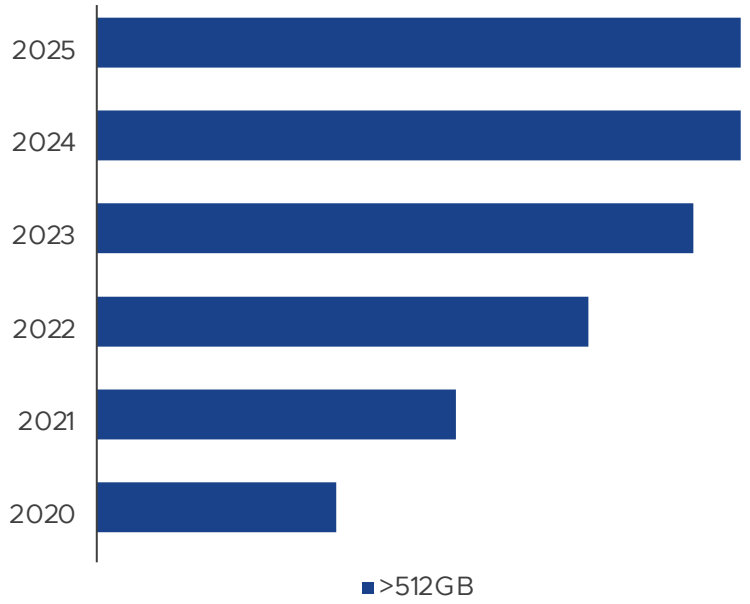


CPU core growth projection over time

DRAM not scaling
Compound annual growth rate in data[2]



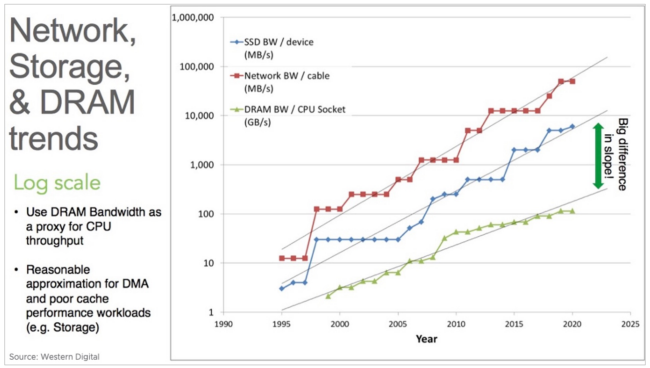
Large-memory systems growing in prevalence

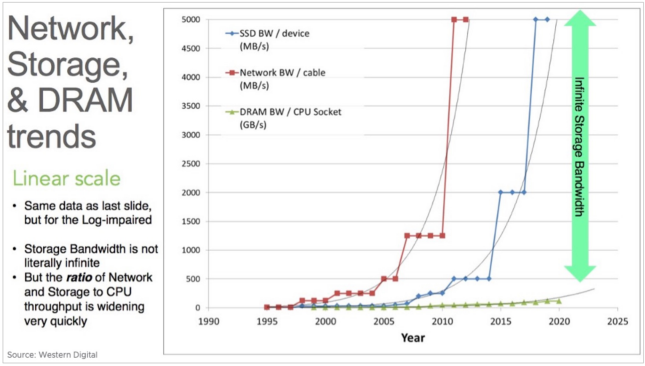


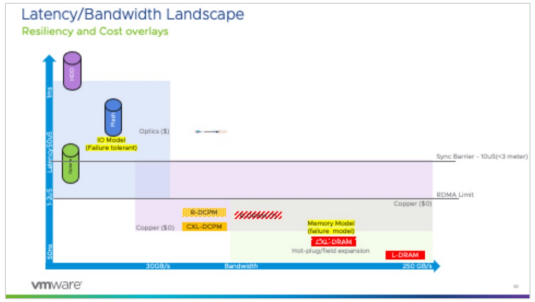
Data-intensive workloads need hot data in memory, but DRAM is expensive and has limited capacity
Enterprise data is vulnerable, but security and encryption can compromise performance

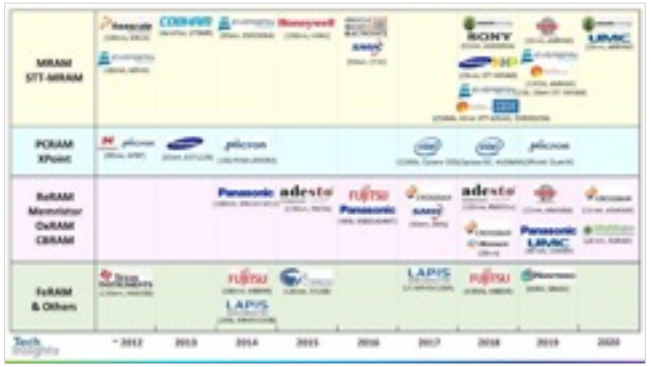
[2] IBM. "3D NAND Technology – Implications to Enterprise Storage Applications." 2015 Flash Memory Summit. 2014. flashmemorysummit.com/English/Collaterals/Proceedings/2015/20150811_FM12_Yoon.pdf.

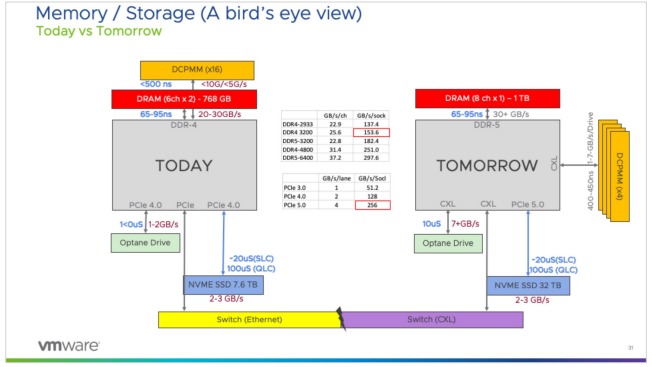
System Trends (Memory vs Network+Storage)











WHY: Addressing key Issues

Simplify, Scale and Resilient

TCO
&
Power

Complex
Memory
Hierarchy

Storage
Vs
Memory

Resiliency

The Solution: VMware Value Proposition

Addressing the growing need of in-memory computing!

WHAT CUSTOMERS HAVE

Inflexible Hardware

Over-Provisioned Memory

Memory is the highest cost factor (50-80%)

Limited Inflexible Capacity

Failure prone

Project Capitola

Software Defined Memory Tiering with uniform consumption model

Cluster level memory pooling, visibility and availability

Fully Integrated with VMware Stack

WHAT CUSTOMERS WANT

Flexible Software

Defined Memory

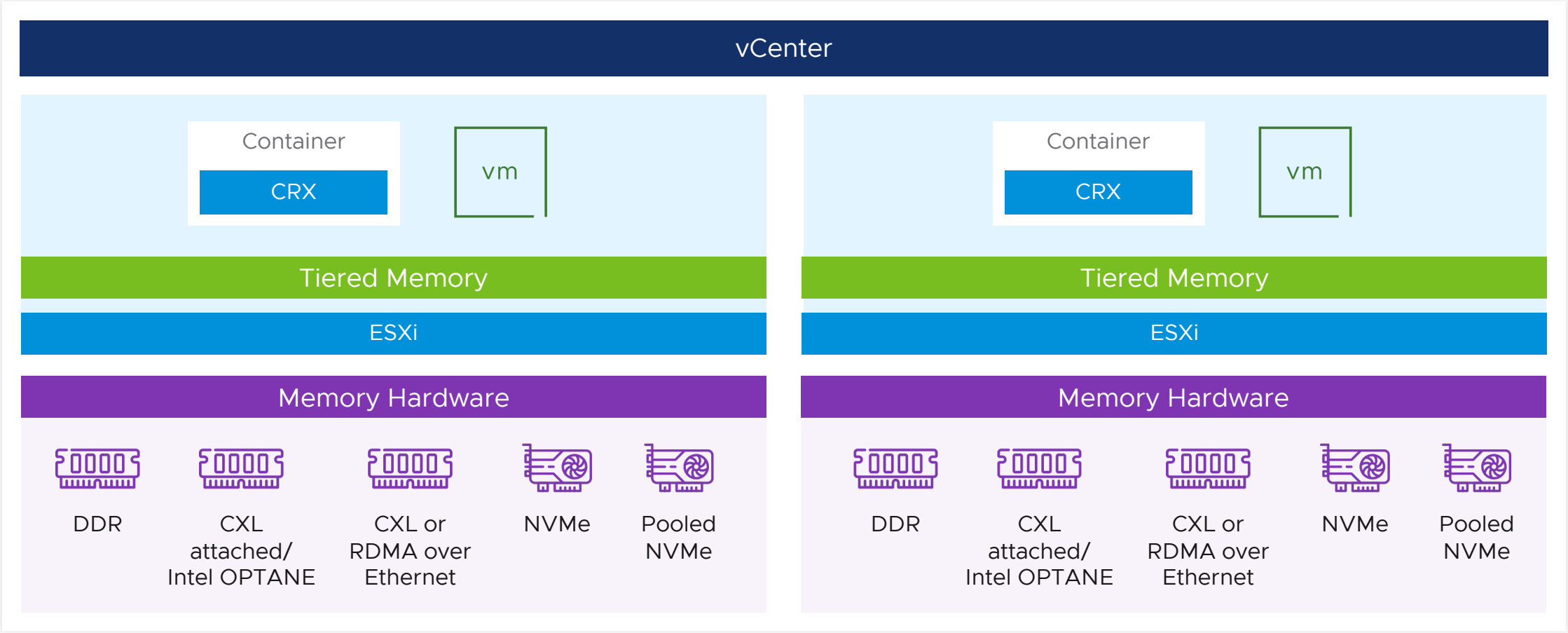
Reduce TCO of memory

Expand Capacity on demand

Faster failure recovery

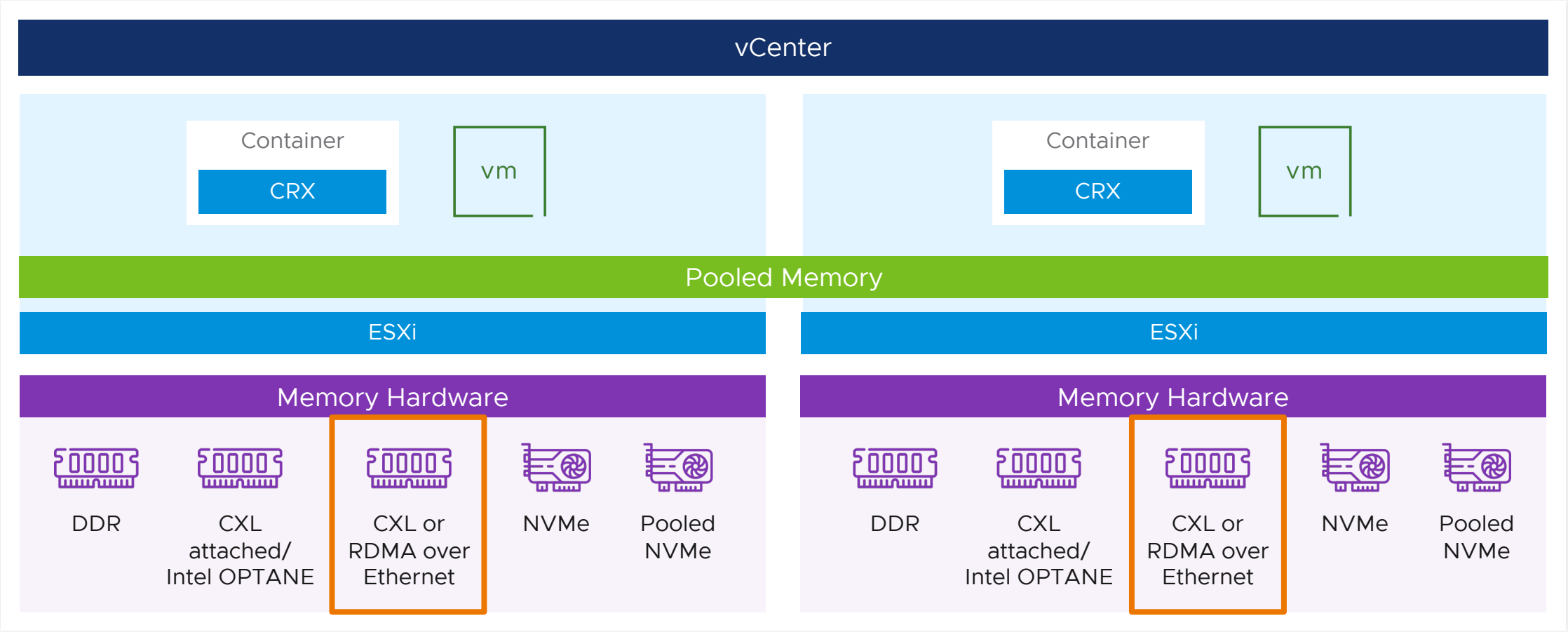
The Solution: How Does it Work?

Phase-1: Local tiering with cluster support



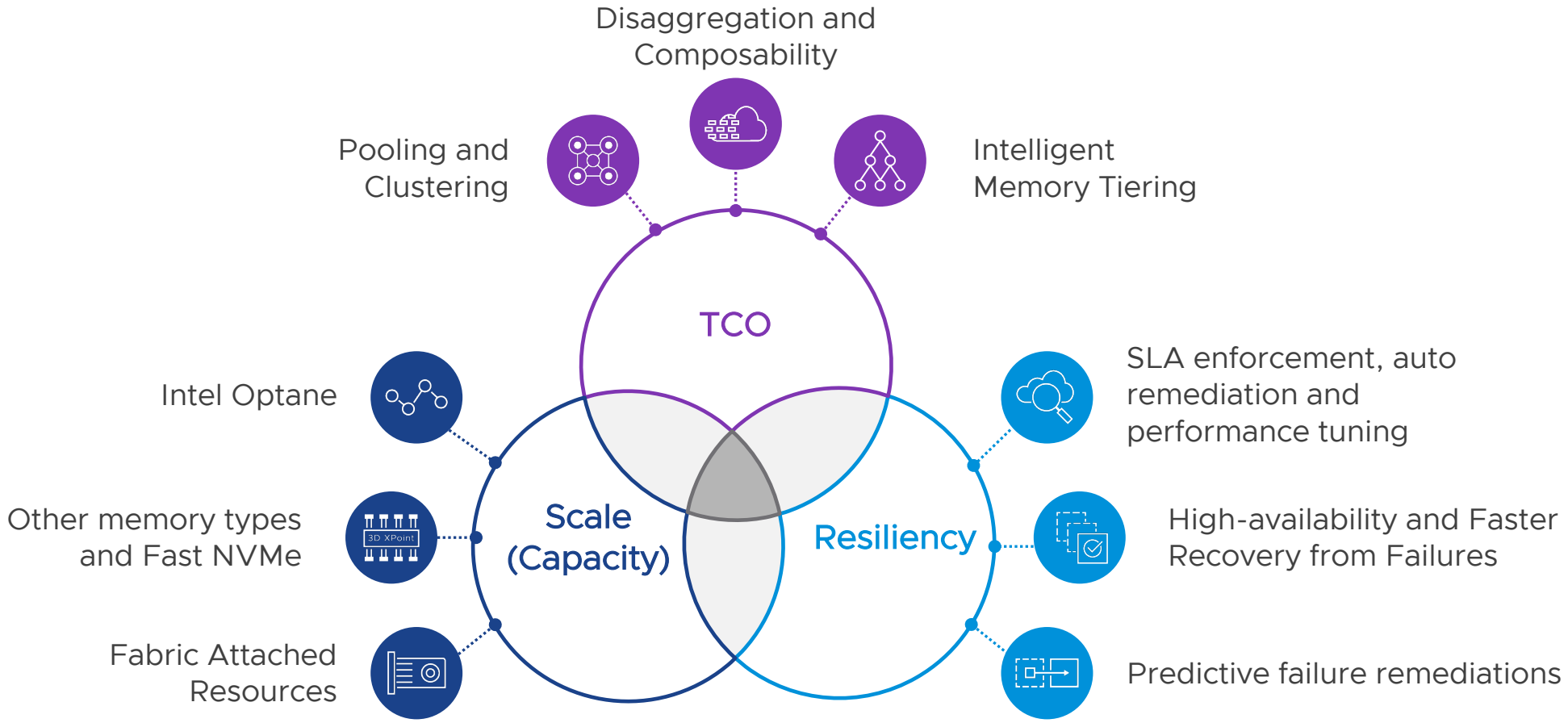
The Solution: How Does it Work?

Phase-2: Cluster wide pooling



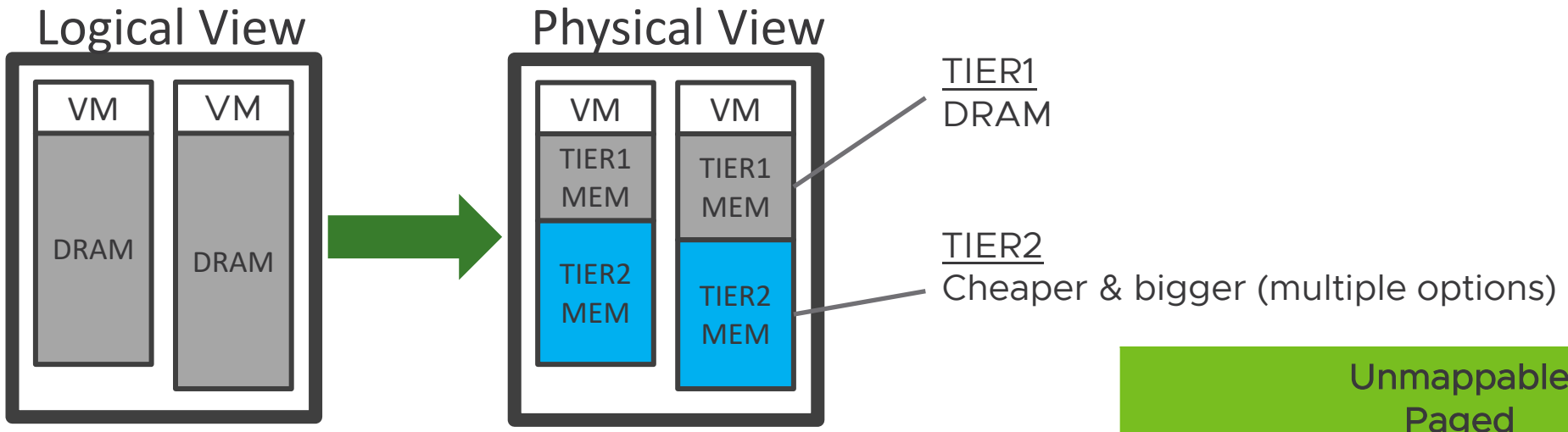
Project Capitola: Various Foundational Technologies In-play

Uniquely positioned to address data-growth and memory-demand with integrated offering



Project Capitola Enables Transparent Tiering

Built in vSphere, requires no modifications to applications or Operating Systems

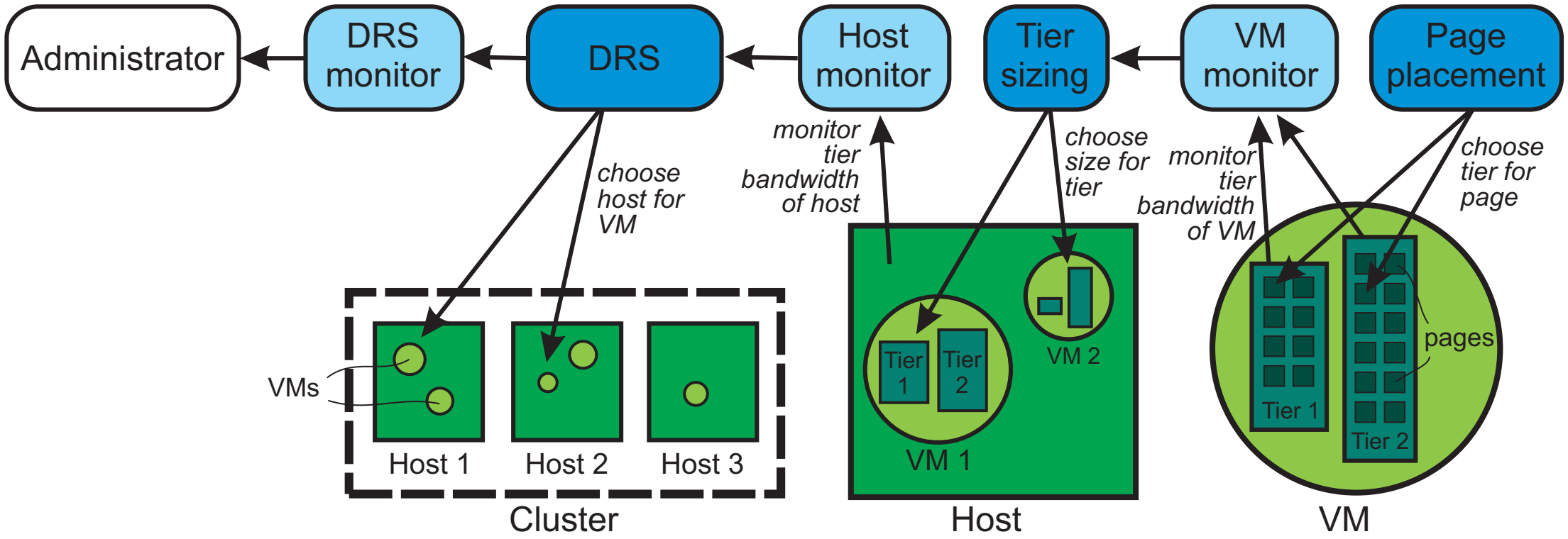


- ✓ vSphere decides what tiers to use and when
- ✓ Memory available to the host is sum of all memory tiers

	Unmappable Paged	Mappable Byte-Addressable
Locally attached	Fast NVMe SSD	High-density mem
Pooled	DRAM High-density mem Fast NVMe SSD	DRAM High-density mem

How it all fits together?

Managed part of end-to-end vSphere workflow!



Project Capitola

Preliminary Results

HammerDB using SQL Server

Configuration Details

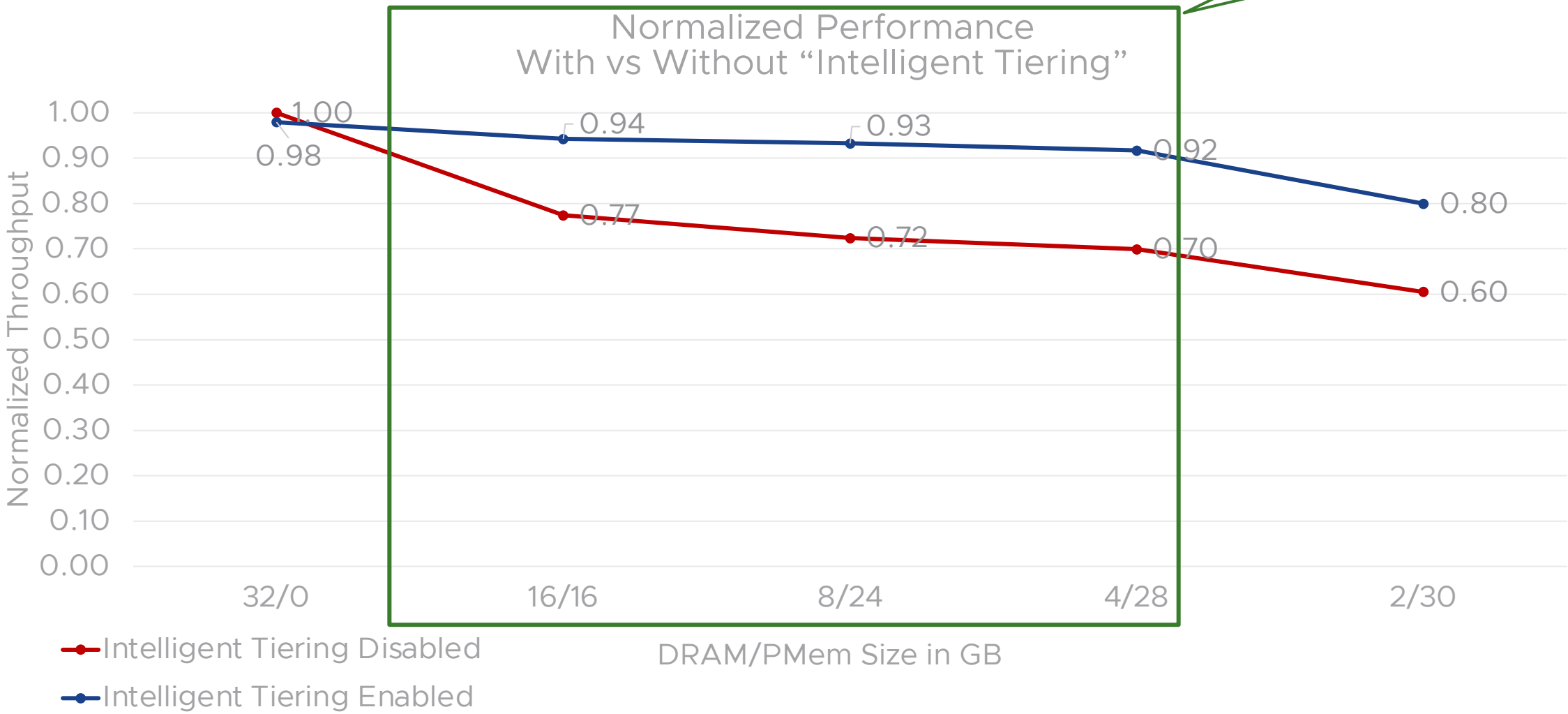
	Hardware Set up
System	DELL PowerEdge R740xd with 2x Intel Xeon Platinum 8280L (28 cores per socket)
Memory	768 GB DRAM (12 x 64 GB DIMMs) 3TB PMEM (12x256GB DIMMs)
Storage	Samsung PM1725b 1.6TB NVMe SSDs
Hypervisor	VMware ESXi™ Future

SQL Server VM	
OS	Windows Server 2016
CPU	8 vCPU
vRAM	32 GB (Max.Mem = 28GB)
NVMe Intel P4600 SSD	100 GB DB, 50 GB Logs

HammerDB	
Virtual Users	500
Warehouses	1000
Warm Up Time	2 mins
Run Time	20 mins
Tests-Profile	TPC_C

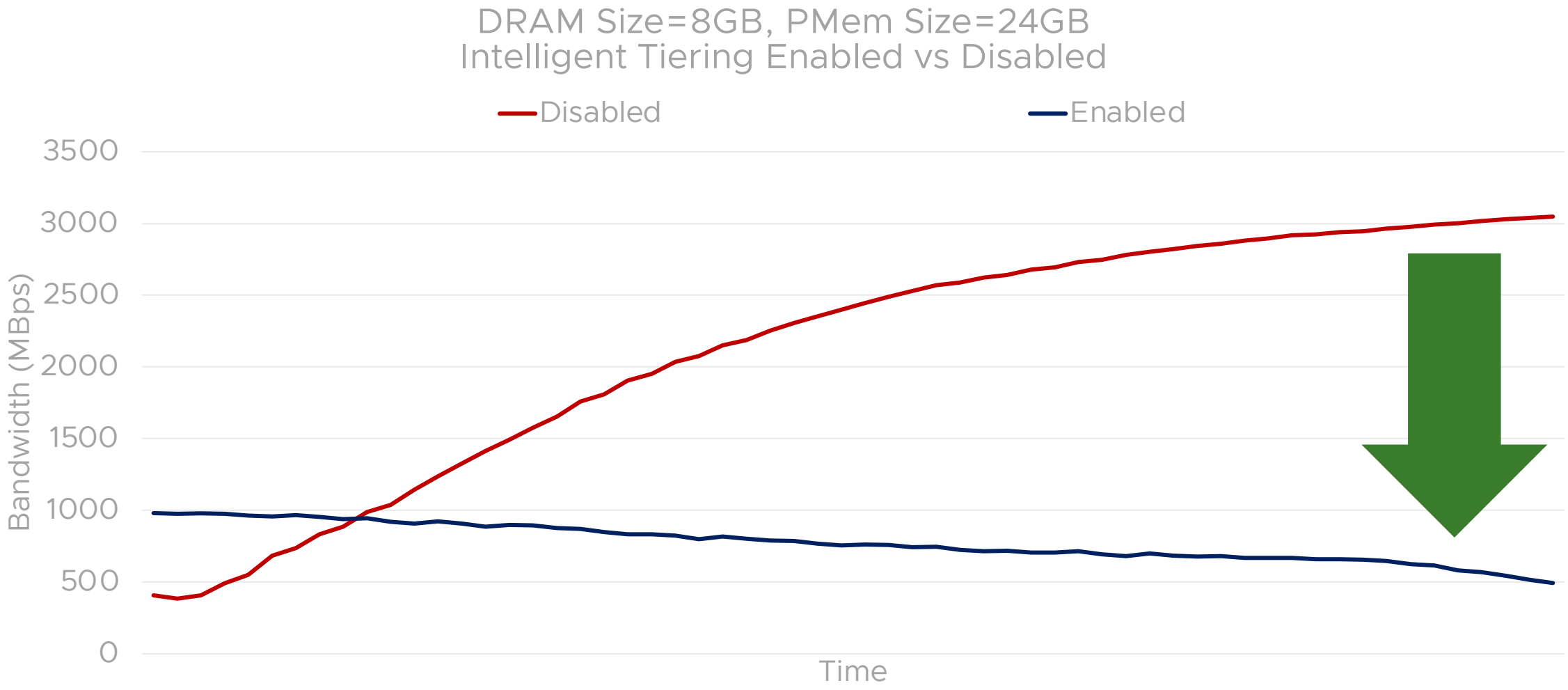
HammerDB using SQL Server

Performance Results (Higher is better)



HammerDB using SQL Server

Bandwidth Comparison (Lower is better)



Java Server Enterprise Workload

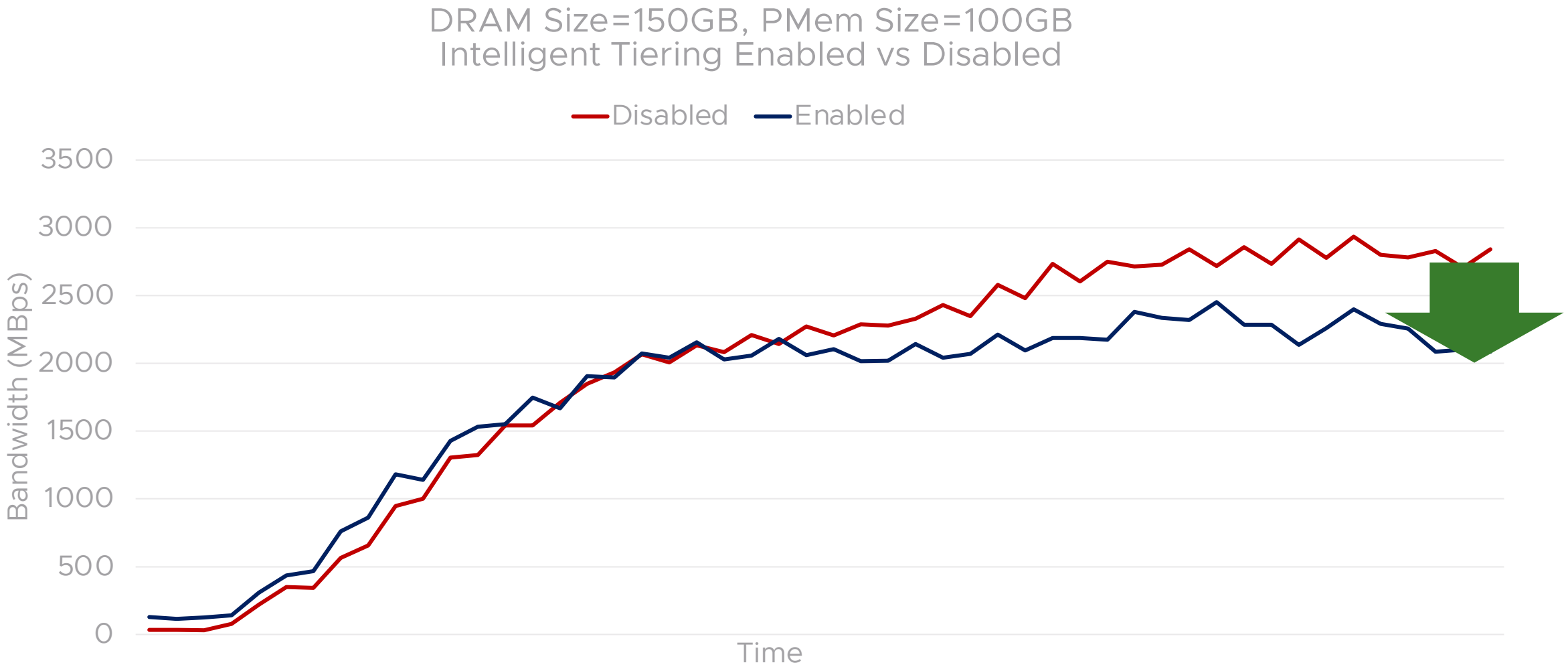
Configuration Details

	Hardware Set up
System	DELL PowerEdge R740xd with 2x Intel Xeon Platinum 8280L (28 cores per socket)
Memory	768 GB DRAM (12 x 64 GB DIMMs) 3TB PMEM (12x256GB DIMMs)
Storage	Samsung PM1725b 1.6TB NVMe SSDs
Hypervisor	VMware ESXi™ Future

Number of VMs	1
vCPUs per VM	8
Memory Size per VM	250 GB
Java Heap Size	200 GB

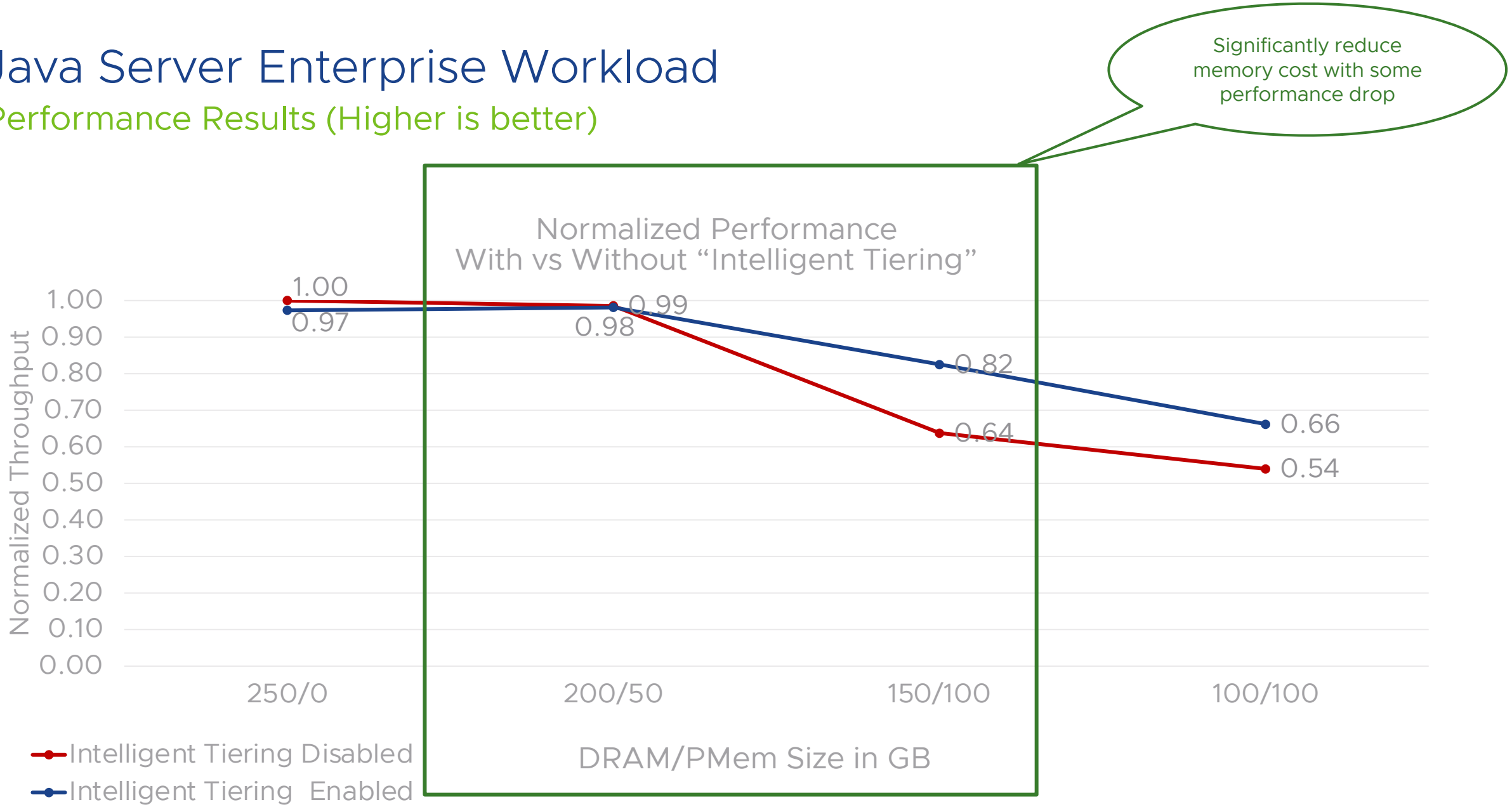
Java Server Enterprise Workload

Bandwidth Comparison (Lower is better)

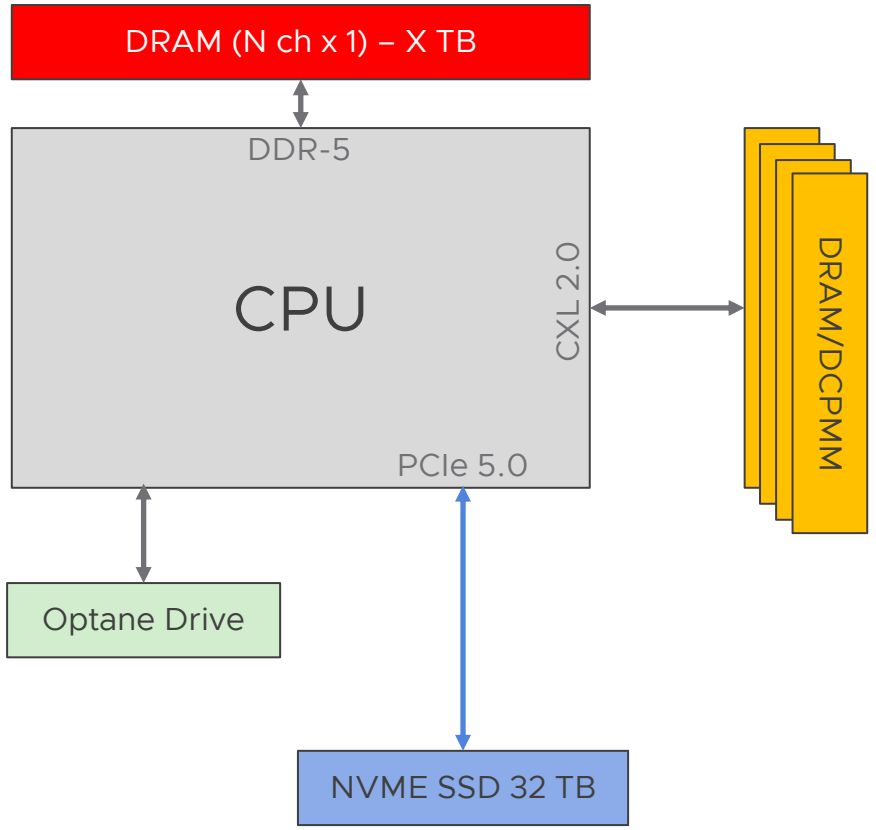


Java Server Enterprise Workload

Performance Results (Higher is better)



CXL: The beginning of the Memory Centric Journey



	DRAM	DRAM + CXL DRAM
Capacity Expansion	Limited	Expanded
BER		Expected to be same
BW Expansion	300 GB/s	+200 GB/s
Latency	65-180ns	100-250 (1-2 NUMA hops)
Hot Add/Remove	NA	Add is easier than remove
Asset Preservation	NA	Potential
Cost	1X	<=1x (Expected)
Computational memory	NA	Enabler
Resiliency / Avail.		Needs Attention

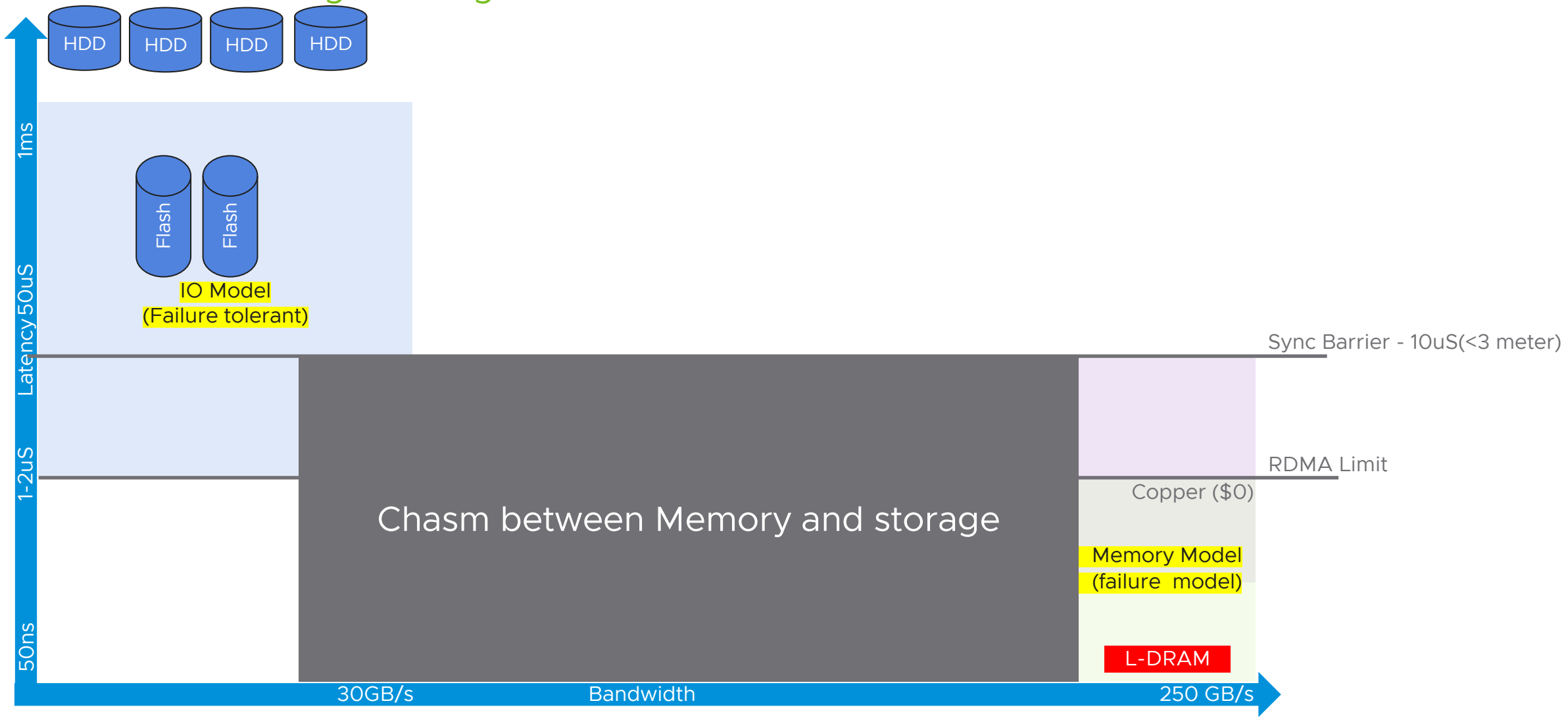
OCP FUTURE TECHNOLOGIES SYMPOSIUM

2021 OCP Global Summit | November 8, 2021, San Jose, CA

EMBEDDED SLIDES

Memory vs Storage Chasm

Performance and Programming model



Compute – Memory Density Roadmap

4KW Chassis	2020	2022	2025	power
Cores	448	672	1024	1680
DRAM (TB)	16	8	16	1920
PMEM (TB)	0	32	128	0
Flash (TB)	128	256	512	320

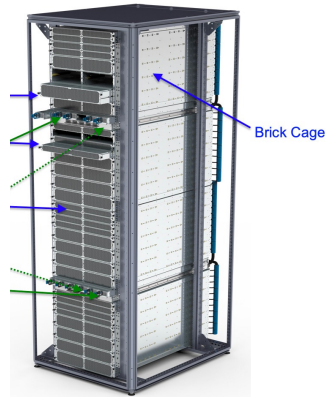
4K Watt
12U
10 (2S) nodes

Accelerator (GPU/FPGA not included yet)

15 KWatt
42 U
40 (2S) nodes

15KW Rack	2020	2022	2025	Power
Cores	1792	2688	4096	6720
DRAM (TB)	64	32	64	7680
PMEM (TB)	0	128	512	0
Flash (PB)	512	1024	2048	1280

Total Mem/Core	37	61	144

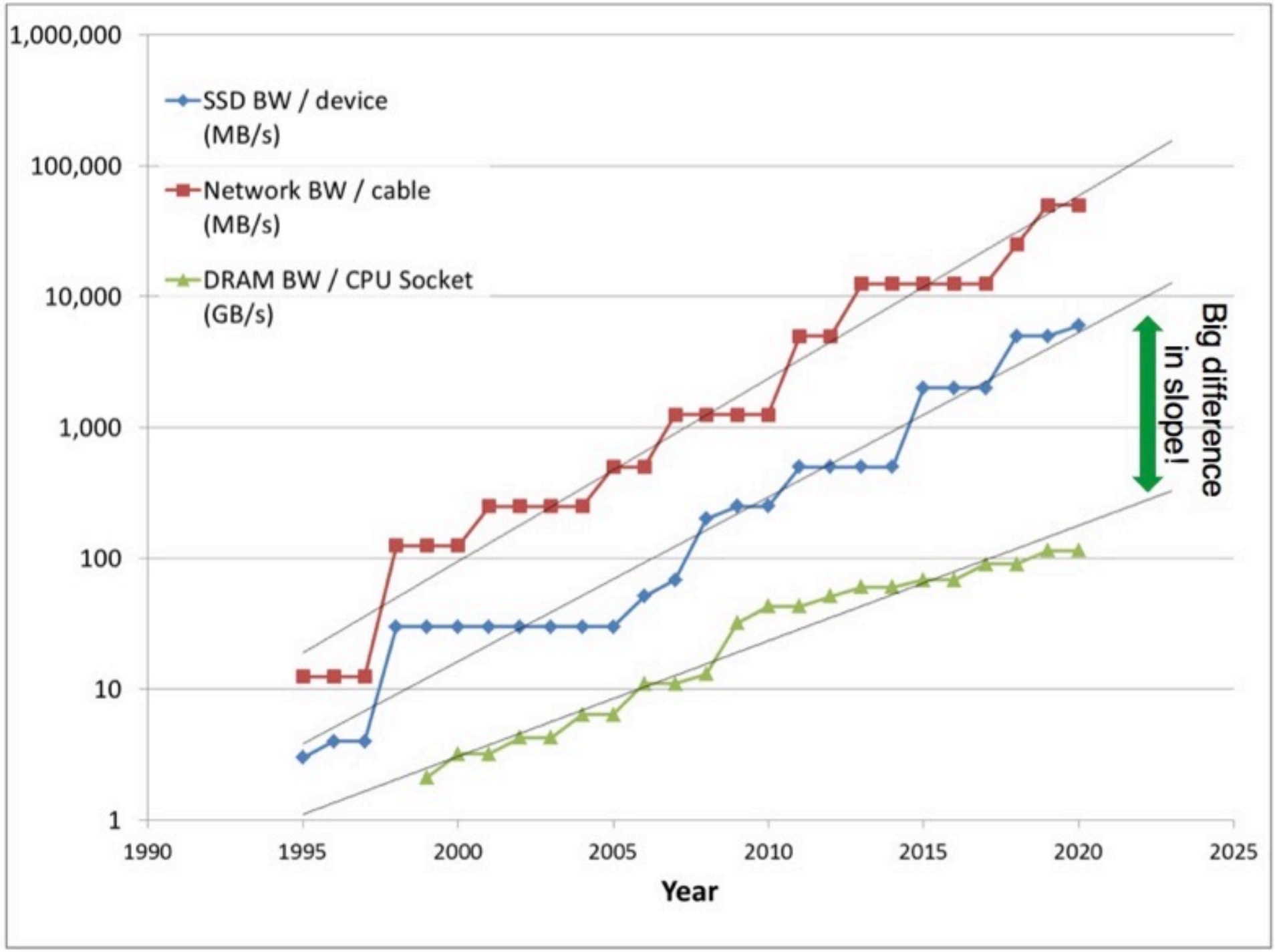


A single rack has enough scale. Latency and programmer simplicity are problems to solve

Network, Storage, & DRAM trends

Log scale

- Use DRAM Bandwidth as a proxy for CPU throughput
- Reasonable approximation for DMA and poor cache performance workloads (e.g. Storage)



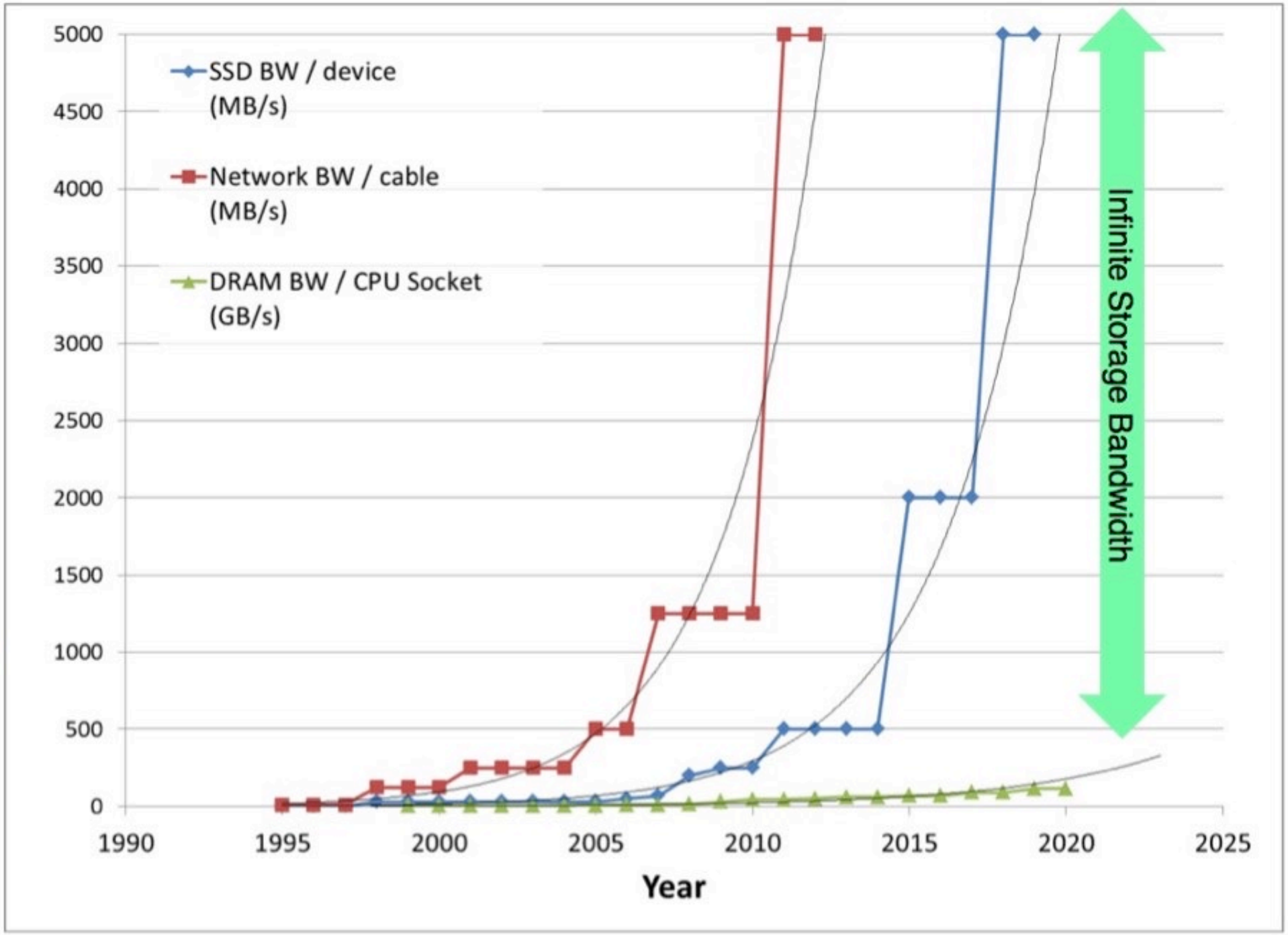
Source: Western Digital

Network, Storage, & DRAM trends

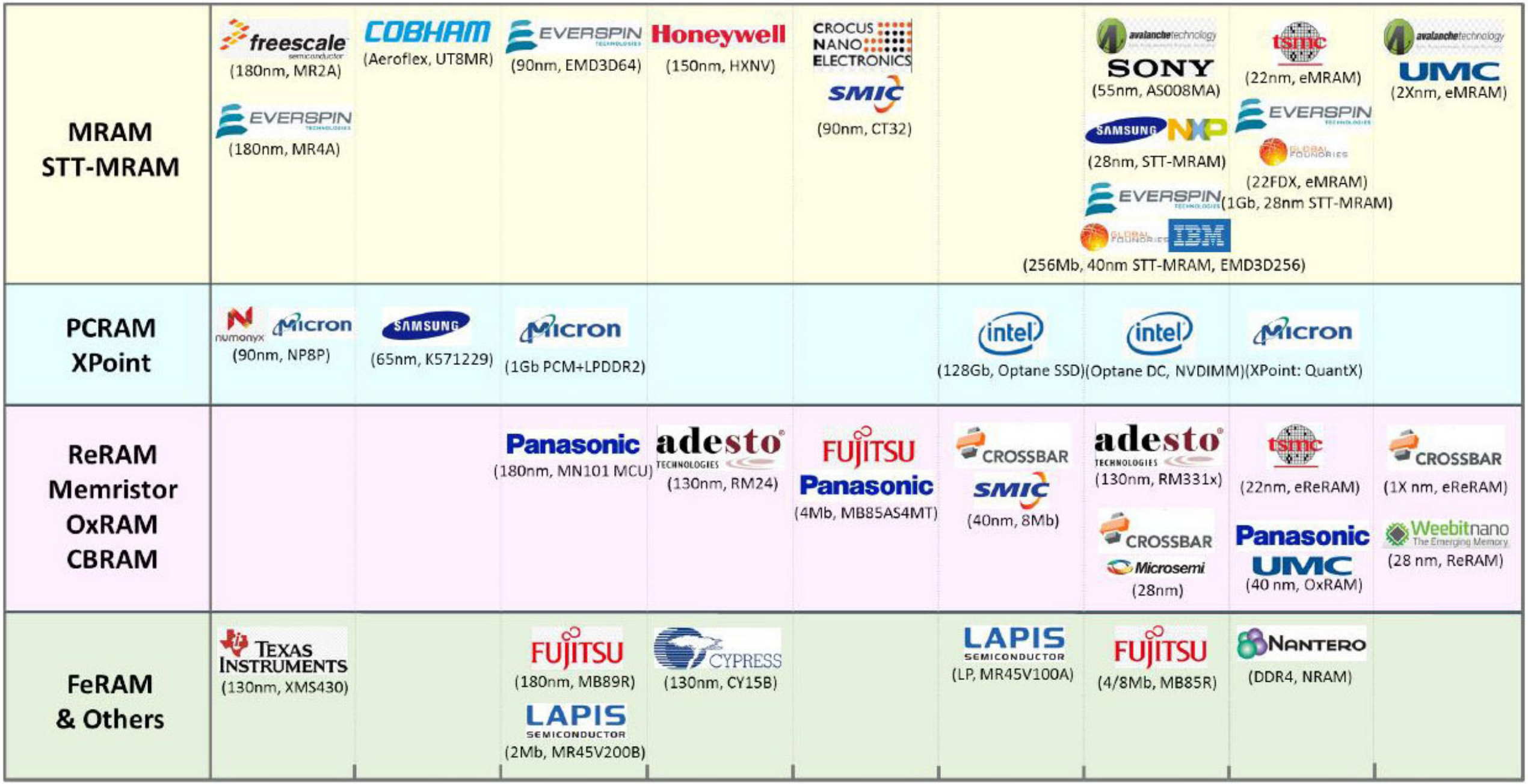
Linear scale

- Same data as last slide, but for the Log-impaired

- Storage Bandwidth is not literally infinite
- But the ***ratio*** of Network and Storage to CPU throughput is widening very quickly

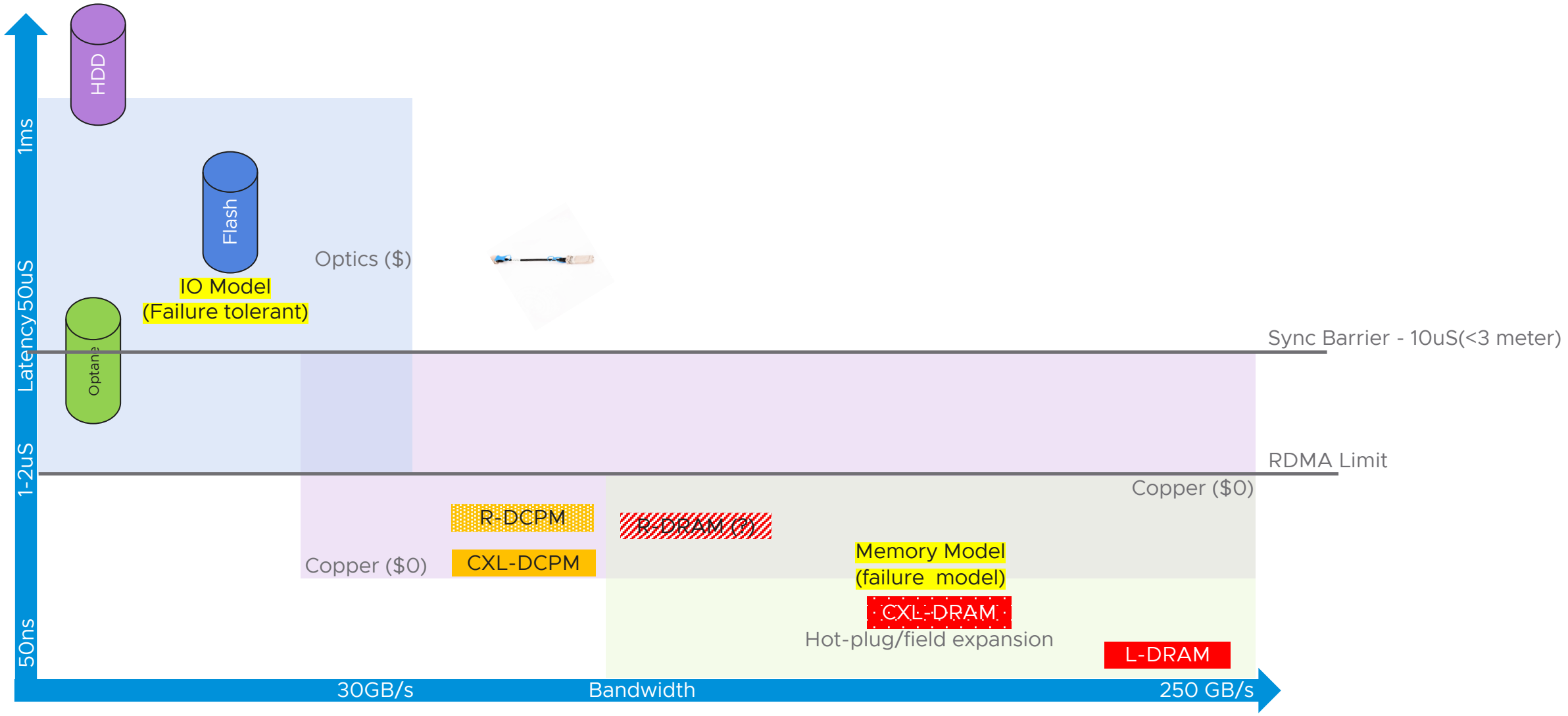


Source: Western Digital


MRAM
STT-MRAM
freescale semiconductor
(180nm, MR2A)
EVERSPIN
(180nm, MR4A)
COBHAM
(Aeroflex, UT8MR)
EVERSPIN
(90nm, EMD3D64)
Honeywell
(150nm, HXNV)
CROCUS NANO ELECTRONICS
SMIC
(90nm, CT32)
avalanchetechnology
SONY
(55nm, AS008MA)
SAMSUNG NXP
(28nm, STT-MRAM)
EVERSPIN
(1Gb, 28nm STT-MRAM)
GLOBAL FOUNDRIES IBM
(256Mb, 40nm STT-MRAM, EMD3D256)
tsmc
(22nm, eMRAM)
EVERSPIN
GLOBAL FOUNDRIES
(22FDX, eMRAM)
avalanchetechnology
UMC
(2Xnm, eMRAM)
PCRAM
XPoint
numonyx Micron
(90nm, NP8P)
SAMSUNG
(65nm, K571229)
Micron
(1Gb PCM+LPDDR2)
intel
(128Gb, Optane SSD)
intel
(Optane DC, NVDIMM)
Micron
(XPoint: QuantX)
ReRAM
Memristor
OxRAM
CBRAM
Panasonic
(180nm, MN101 MCU)
adesto TECHNOLOGIES
(130nm, RM24)
FUJITSU
Panasonic
(4Mb, MB85AS4MT)
CROSSBAR
SMIC
(40nm, 8Mb)
adesto TECHNOLOGIES
(130nm, RM331x)
CROSSBAR
Microsemi
(28nm)
tsmc
(22nm, eReRAM)
Panasonic
UMC
(40 nm, OxRAM)
CROSSBAR
(1X nm, eReRAM)
Weebitnano The Emerging Memory
(28 nm, ReRAM)
FeRAM
& Others
TEXAS INSTRUMENTS
(130nm, XMS430)
FUJITSU
(180nm, MB89R)
LAPIS SEMICONDUCTOR
(2Mb, MR45V200B)
CYPRESS
(130nm, CY15B)
LAPIS SEMICONDUCTOR
(LP, MR45V100A)
FUJITSU
(4/8Mb, MB85R)
NANTERO
(DDR4, NRAM)
~ 2012
2013
2014
2015
2016
2017
2018
2019
2020
Tech Insights

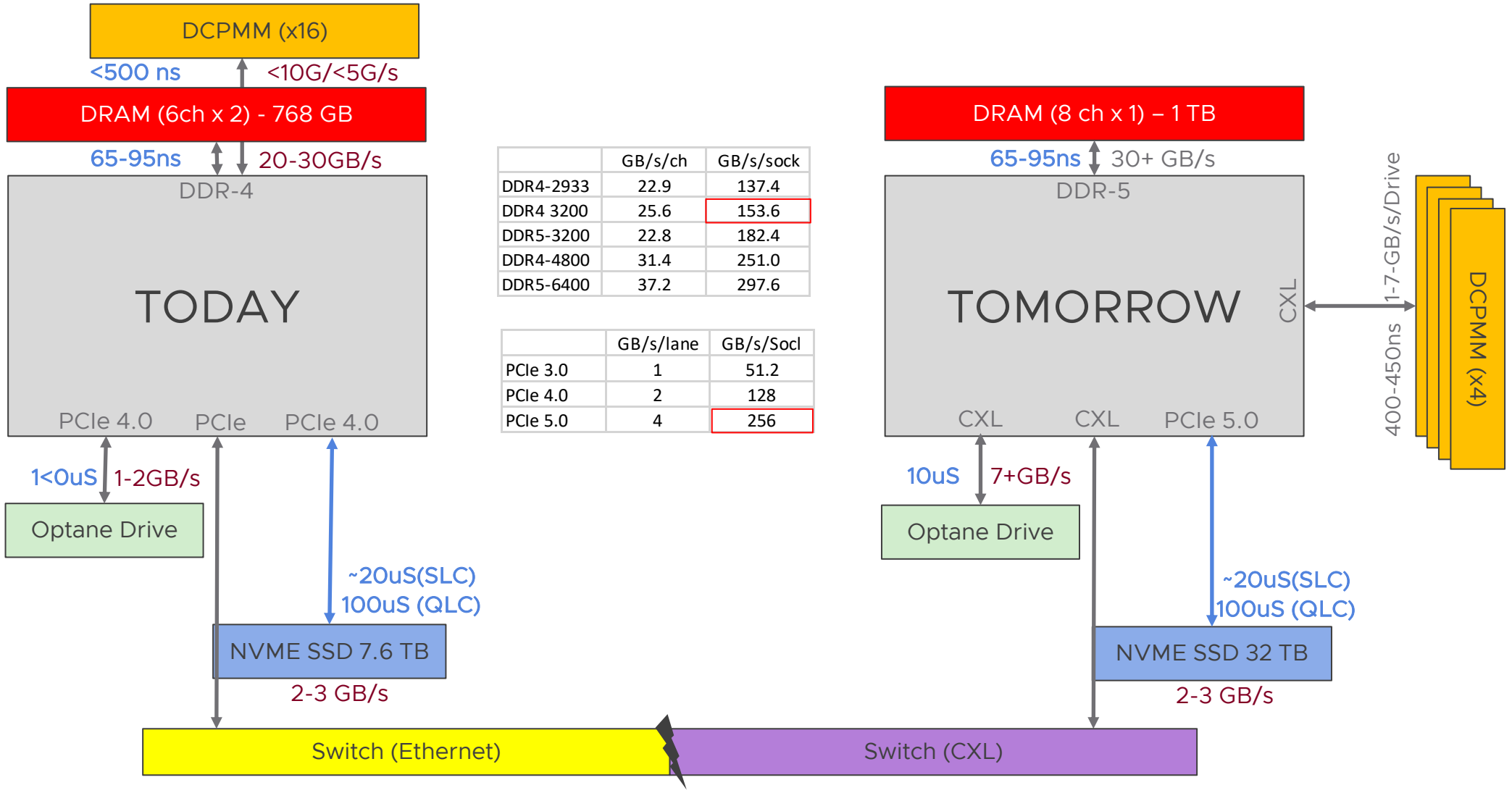
Latency/Bandwidth Landscape

Resiliency and Cost overlays



Memory / Storage (A bird's eye view)

Today vs Tomorrow

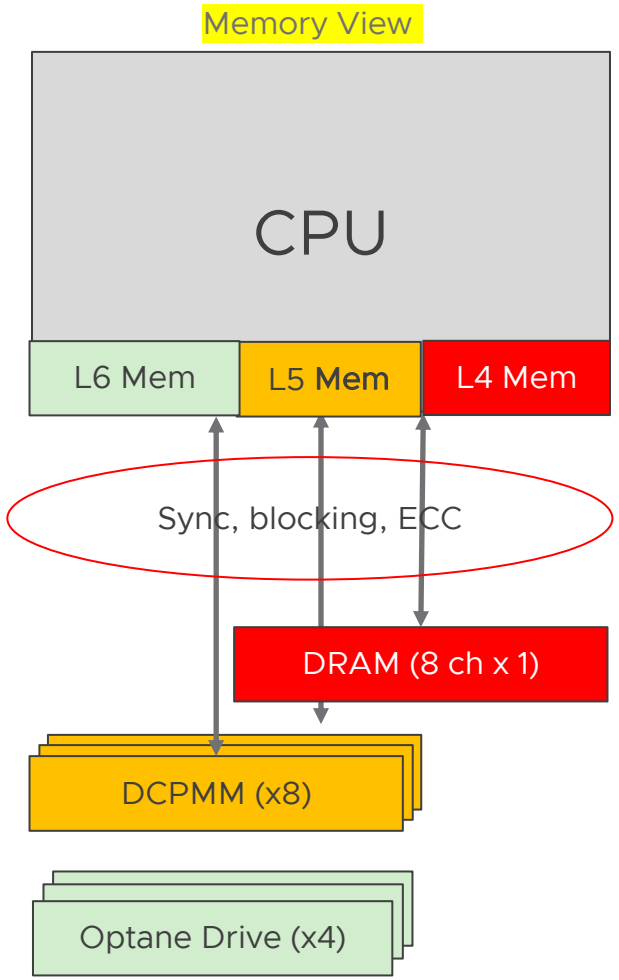


	GB/s/ch	GB/s/sock
DDR4-2933	22.9	137.4
DDR4 3200	25.6	153.6
DDR5-3200	22.8	182.4
DDR4-4800	31.4	251.0
DDR5-6400	37.2	297.6

	GB/s/lane	GB/s/Socl
PCIe 3.0	1	51.2
PCIe 4.0	2	128
PCIe 5.0	4	256

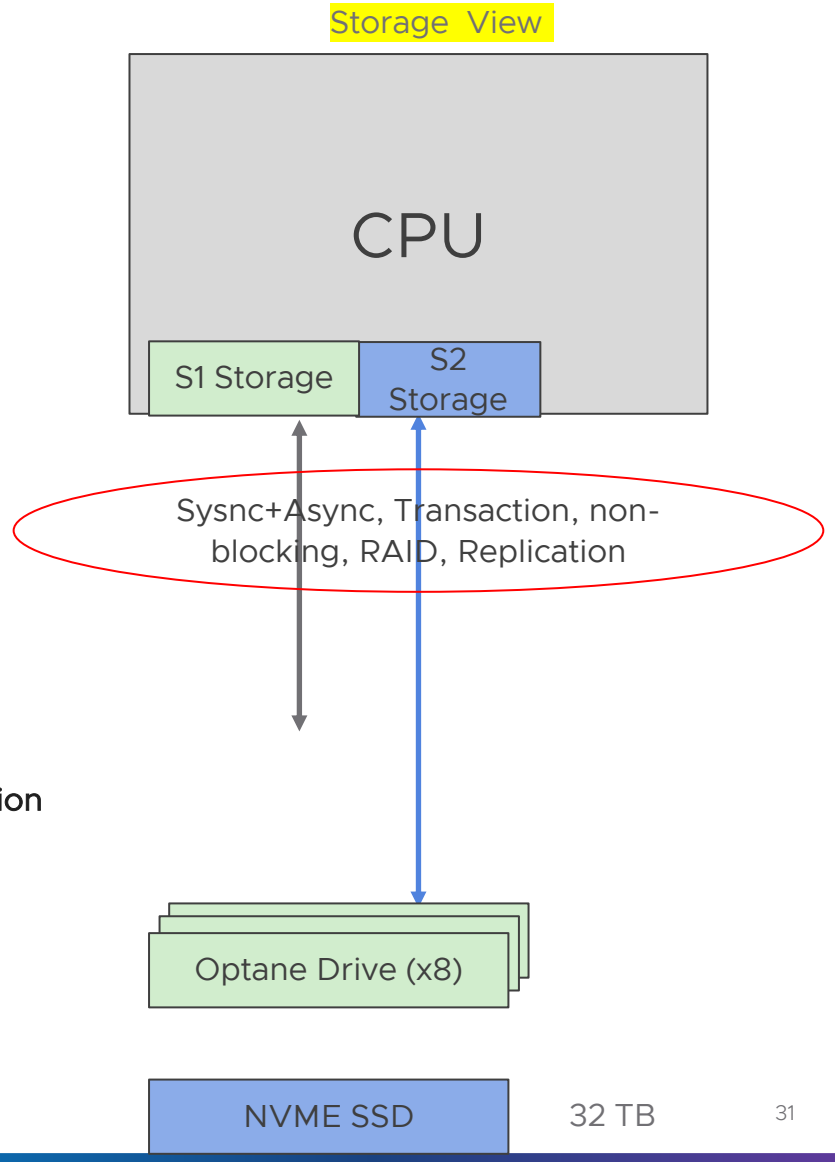
Memory Centric vs Storage Centric View

Logical View



Failure Modes/issues

1. Media related failure rates
2. Transaction model
3. Resilience Model (ECC Codes vs RAID/Replication



OCP FUTURE TECHNOLOGIES SYMPOSIUM

2021 OCP Global Summit | November 8, 2021, San Jose, CA