



Flash Memory Summit

Enterprise Flash Storage Annual Update

Flash, It's not just for tier 0 anymore
Or
Flash is the new black

Santa Clara, CA
August 2019

Howard Marks
Technologist Extraordinary
and Plenipotentiary





Flash Memory Summit

Your not so Humble Speaker

- 30+ years of consulting & writing for trade press
- Occasional blogger at TechTarget
- Recently Chief Scientist DeepStorage, LLC.
 - Independent test lab and analyst firm
- Technologist Extraordinary and Plenipotentiary
 - VAST Data
 - I promise to keep the sales down, I'm new to it anyway





Agenda

- Review 2017-2018 events, predictions
 - Flash is just normal
 - The shift from SSD to NVMe
 - NVMe over fabrics the new lingua franca
 - Is Tier 0 sustainable
 - 3D Xpoint and Storage Class Memory
- A look at a few illustrative examples



Flash Memory Summit

A Decade+ of Enterprise Flash



2007

- Rackmount SSDs
- Texas Memory
- Violin Memory
- Fast but niche



2010

- SSDs in DISK arrays
- High cost
- Endurance fears
- Hybrids emerge



2014

- Flash goes commercial
- All Flash Arrays
- Costs = high performance HDD



2017

- Flash is mainstream
- Full data services & data reduction
- Cost effective for primary storage



2020

- Democratizing flash
- Data intensive applications
- 3D Xpoint starts small/fast cycle again



The Tipping Point Tipped

- 2017
 - Enterprise SSD 25X capacity HDD \$/GB
- 2019
 - 1 TB SSD < \$100
 - Enterprise SSD 10-12¢/GB (3.5X)
 - WD exits 10 & 15K RPM HDDs
 - SK Hynix announces 128 layer 128 Tb chip



Intel 660p Series M.2 2280 1TB PCIe NVMe 3.0 x4 3D2, QLC Internal Solid State Drive (SSD) SSDPEKNW010T8X1

\$94.99 (25 Offers)

Save: 51%

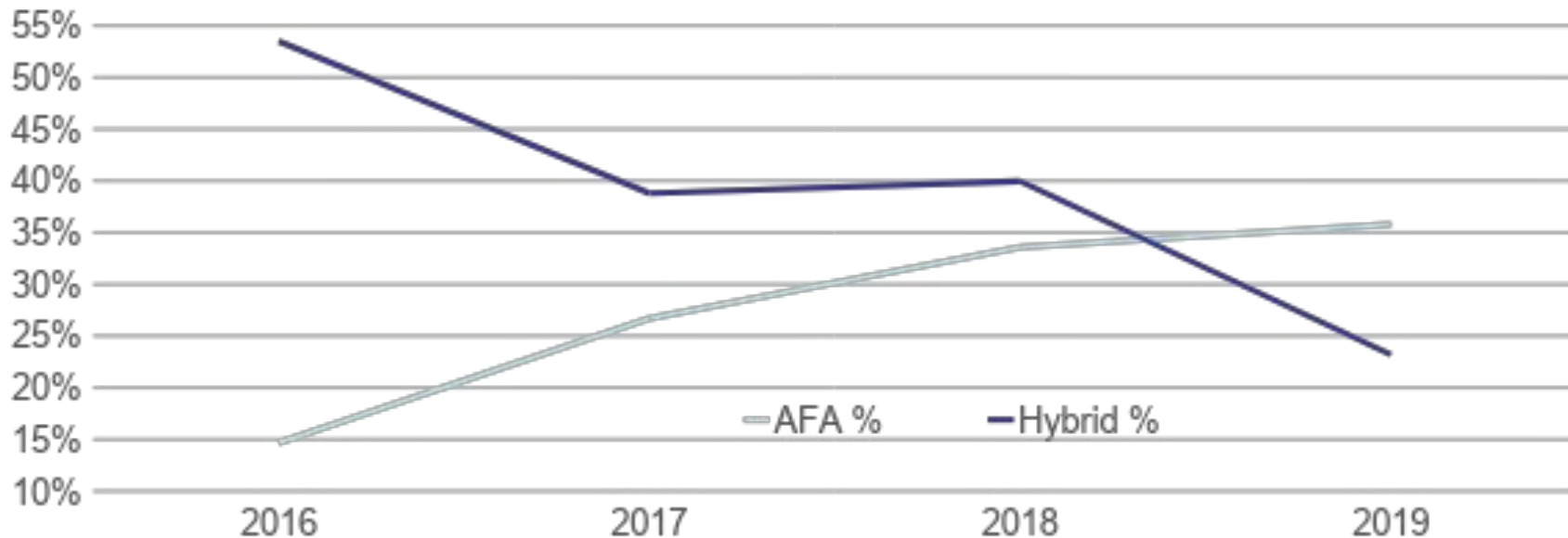
Free Shipping

	GB	Cost	\$/GB	HHD Multiple
Intel P3700	1600	\$405	0.25	8.44
Intel P3520	2000	\$535	0.27	8.92
16TB HDD	16000	\$480	0.03	
Micron 5120 ION	7680	\$800	0.10	3.47



The Tipping Point Tips, Part Deux

- AFA market share passes hybrids





All Flash Player Joins the Big Boys

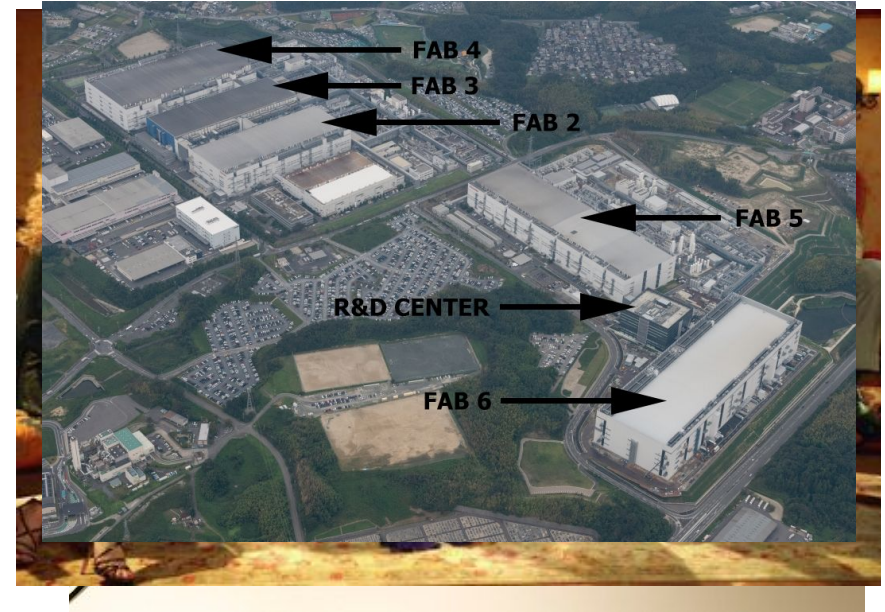
Company	1Q19 Revenue	1Q19 Market Share	1Q18 Revenue	1Q18 Market Share	1Q19/1Q18 Revenue Growth
1. Dell Technologies ^a	\$2,355.9	34.4%	\$2,219.6	34.0%	6.1%
2. NetApp	\$894.9	13.0%	\$890.1	13.6%	0.5%
3. HPE/New H3C Group ^b	\$745.4	10.9%	\$652.2	10.0%	14.3%
4. Hitachi	\$452.7	6.6%	\$457.9	7.0%	-1.1%
T5. IBM*	\$320.0	4.7%	\$364.1	5.6%	-12.1%
T5. Pure Storage*	\$289.5	4.2%	\$236.4	3.6%	22.4%
Rest of Market	\$1,800.3	26.2%	\$1,709.0	26.2%	5.3%
Total	\$6,858.6	100.0%	\$6,529.3	100.0%	5.0%

Source: IDC Worldwide Quarterly Enterprise Storage Systems Tracker, June 6, 2019.



The Toshiba Memory Soap Opera

- 2006 Toshiba buys Westinghouse
- 3/2017 Westinghouse chapter 11 (API000 Reactors \$9B loss)
- 9/2109 Toshiba sells memory unit \$18B
 - WAIT – Western Digital/SanDisk sue
 - Lawyers make money, waste time
- Sale Closes 6/2018
 - 9/2018 Fab 6 opens at Yokkaichi
- June 16 Power failure at Yokkaichi
 - 6+ EB NAND production lost
- Also June CEO Yasuo Naruke, goes on sick leave





Flash Memory Summit

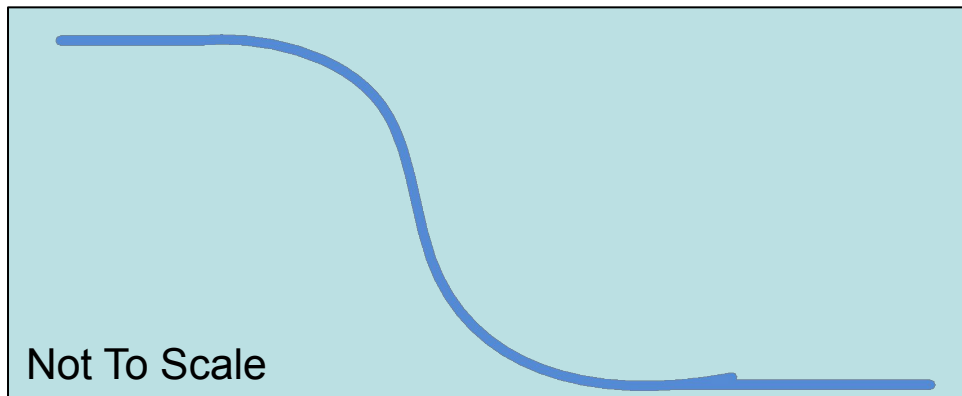
Toshiba becomes Kioxia

- IPO planed for 9/2019
 - Before both power failure and CEO illness
- Rebrand effective October 1, 2019.
- Kioxia:
 - Kioku meaning “memory”
 - Japanese
 - Axia meaning “value”
 - Greek
- pronunciation :
kee-ox-ee-uh



The Party's Over, again

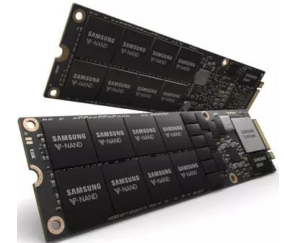
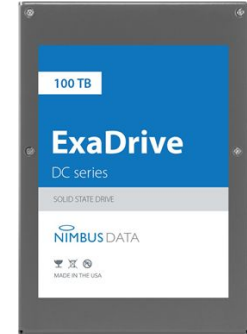
- 2008-2015 SSD \$/GB $-30\%/yr$
- 2016-2018 maybe 30% total
- Last year I said “Expect 30+% CAGR”
- I thought:
 - Supply is easing
 - 96 layer+ QLC
 - Process improvements
 - New fabs
- Fabs cut back starts
- Next 2 quarters flat, back to 25-30% CAGR





Enterprise SSD Evolution

- Data center NVMe \approx SAS/SATA volume
- SSDs and HDDs now both \approx 16 TB
- Greater Differentiation
 - Performance and cost vary 5X or more
 - SLC returns as SCM
- New form factors remain proprietary
 - M.2 didn't work in data centers
 - Samsung NGSFF
 - Intel Ruler



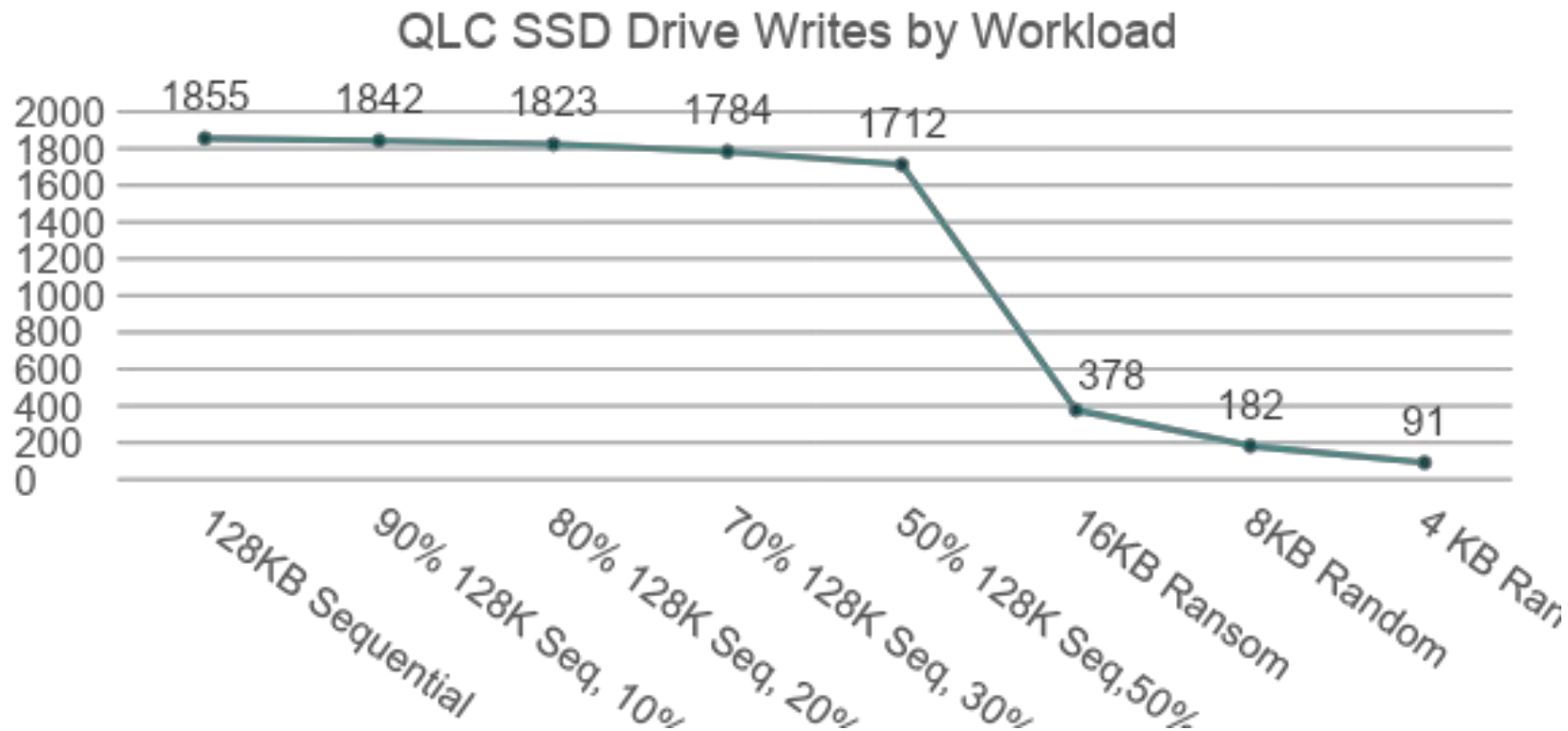


SSD Differentiation

- Storage Class Memory SSDs
 - More on this later
- Dual-port enterprise
 - DRAM/Supercap
- Single port enterprise
 - NVMe and SATA for HCI, HPC, Etc.
- Low cost single port
 - Hyperscaler's tail



QLC SSD Endurance by Workload





Open Channel SSDs

The disaggregation of flash storage

Today - Monolithic model

Hardware managed (SSD)



Denali model

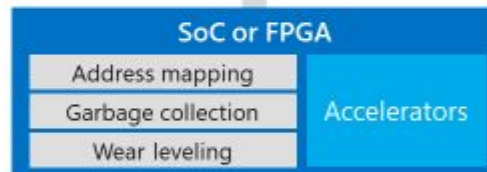
Software defined (Direct)



pBLK interface



Software defined (Offloaded)





PCIe Advances



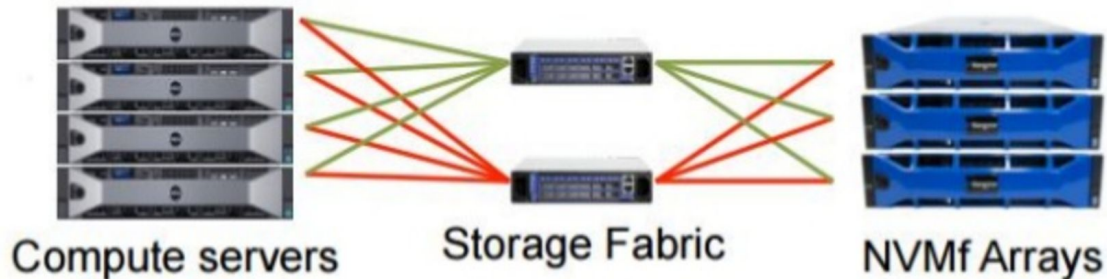
- PCIe 4.0
 - Doubles bandwidth/lane to 2GBps
 - Driven by 100Gbps Ethernet & NVMe
 - Power systems shipping now
 - x86 Next server chipset release
- PCIe 5.0 close on its heels
 - .7 version issued May 2018
 - Adoption planned Q1 2019
 - 400Gbps Ethernet \approx x16 slot
 - Servers and such 2020?

	Spec Date	Raw	Bandwidth per lane	x8 Gbps
PCIe 1	2003	2.5GT/s	250MB/s	16
PCIe 2	2007	5.0GT/s	500MB/s	32
PCIe 3	2010	8.0GT/s	984MB/s	64 (63.04)
PCIe	201	16GT	1969MB/	126



NVMe Over Fabrics (NVMe-oF)

- Extends/encapsulates NVMe semantics over
 - Ethernet with RMDA
 - Fibre Channel
 - Infiniband (no products yet announced)
 - TCP
- Adds name spaces and discovery
- 10-50 μ sec protocol and network overhead





NVMe-oF Models

- JBOF
 - Just Fabric-SSD bridges
 - HA optional
- JBOF+
 - Adds slice/dice and RAID
 - Also manage in client models
- NVMe-oF Array
 - All the abstractions and services of SCSI over Fibre Channel
 - Lower latency of NVMe-oF



NVMeOF Pioneers Shakeout

- Mangstor
 - Reborn as EXTEN
 - Software NVMe-oF JBOF+
- Apeiron – 40Gbps Ethernet switch in JBOF
- E8 – Dual controller array – basic services
 - Acquired by AWS
- Excellero – Low CPU SDS, RDMA



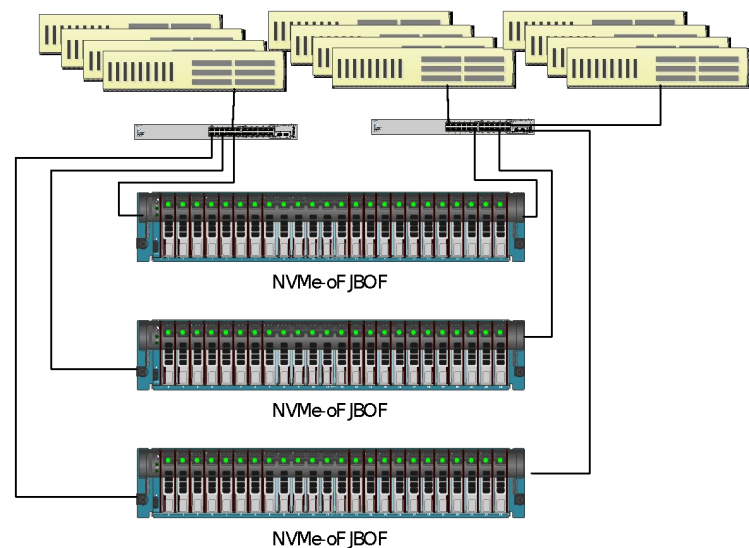
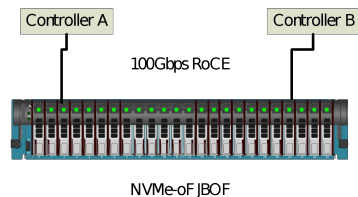
I feel much better





NVMe-oF Use Cases

- Intra-storage system SAS replacement
- HPC/skunkworks/Rackscale
 - RDMA to JBOFs
- Hyperscale
 - TCP to expand to data center scale
- Enterprise
 - Primarily arrays
 - NVMe runs over Fibre Channel for these customers





NVMe Over Fibre Channel

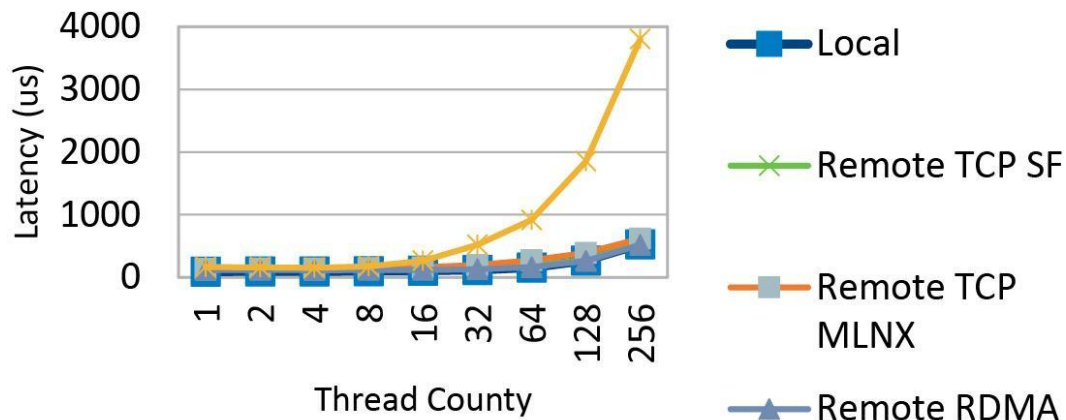
- Fibre Channel
 - Zero copy vs RDMA
 - Flow and congestion control
- Gen5 (16) and Gen6 (32Gbps) Fibre Channel
- One fabric for SCSI and NVMe
- Keeps storage network in storage domain
- The safe move in enterprise



NVMe over TCP

- Encapsulates NVMe verbs in TCP
- Relies on TCP low control
- NIC offload optional
- No switch config requirements
- Nominal latency addition
- Supporters:
 - SolarFlare
 - Cavium
 - Toshiba

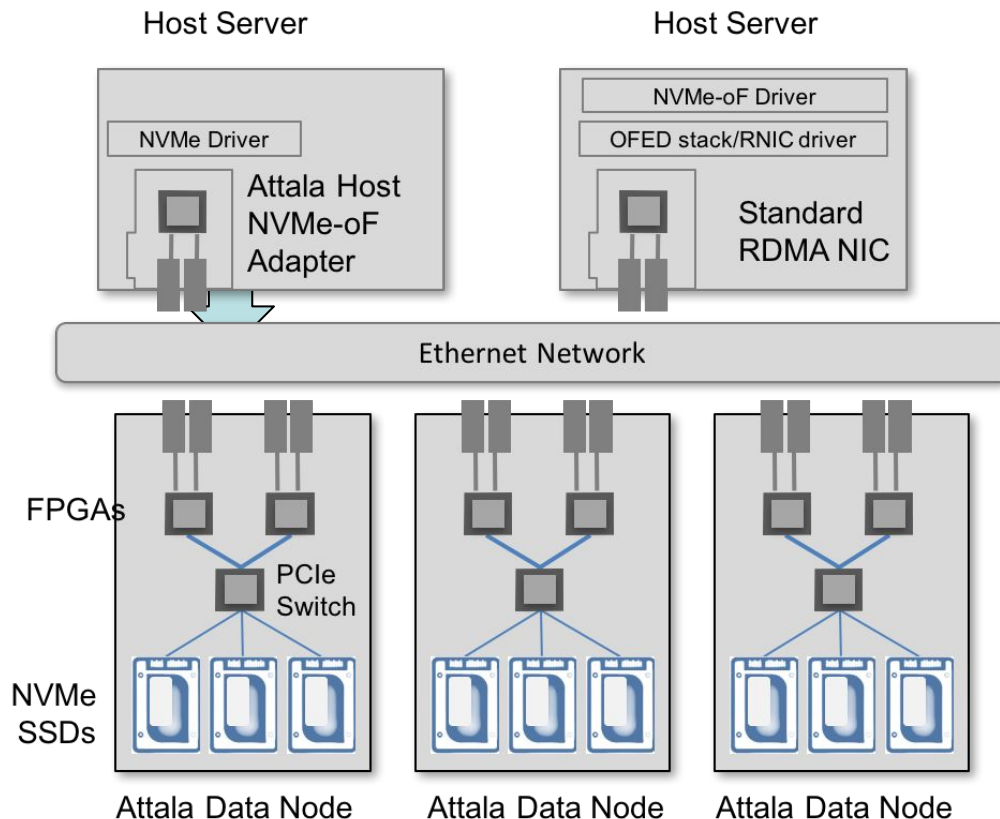
LATENCY - Sustained 4K Random Read



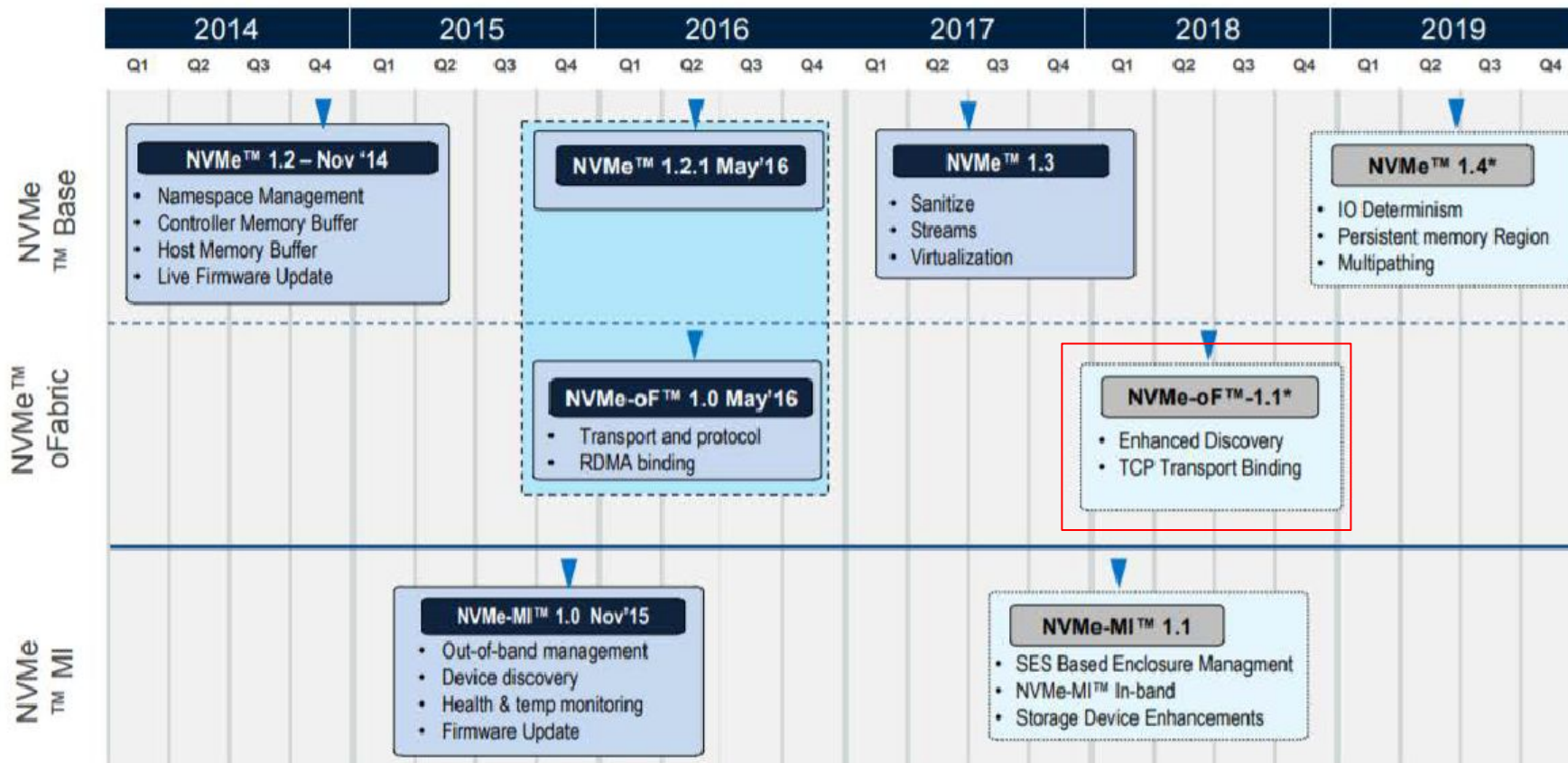


NVMe JBODs Emerge

- Today's JBODs are x86 servers
 - Eg: Toshiba KumoScale
 - High flexibility
 - High cost
- NVMeoF ASICs
 - Vastly reduce costs
 - Sampling from
 - SolarFlare Xilinx
 - Kazan Networks
 - Attala Systems
 - Mellanox



NVMe™ Feature Roadmap



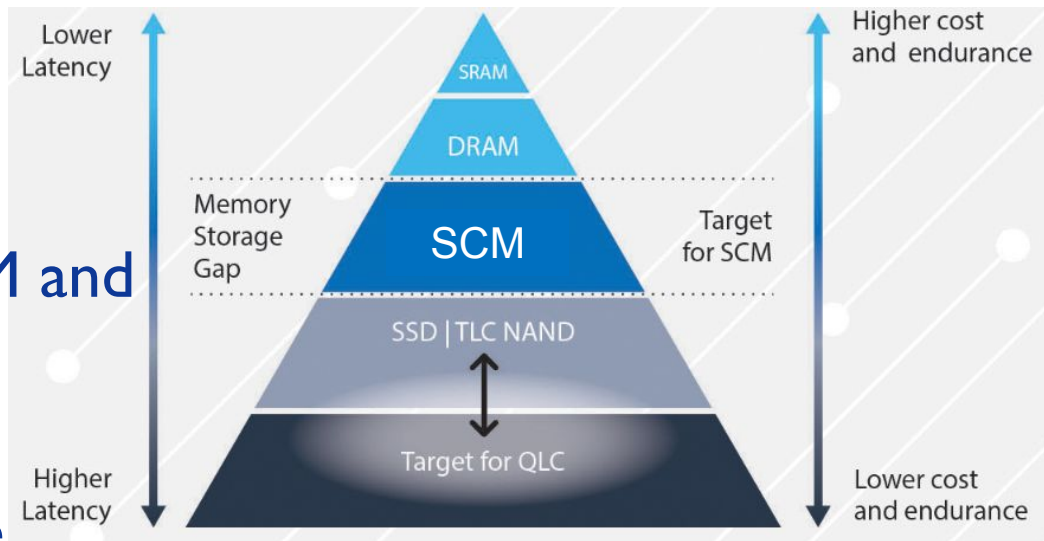
Released NVMe™ specification
 Planned release

* Subject to change



Storage Class Memory

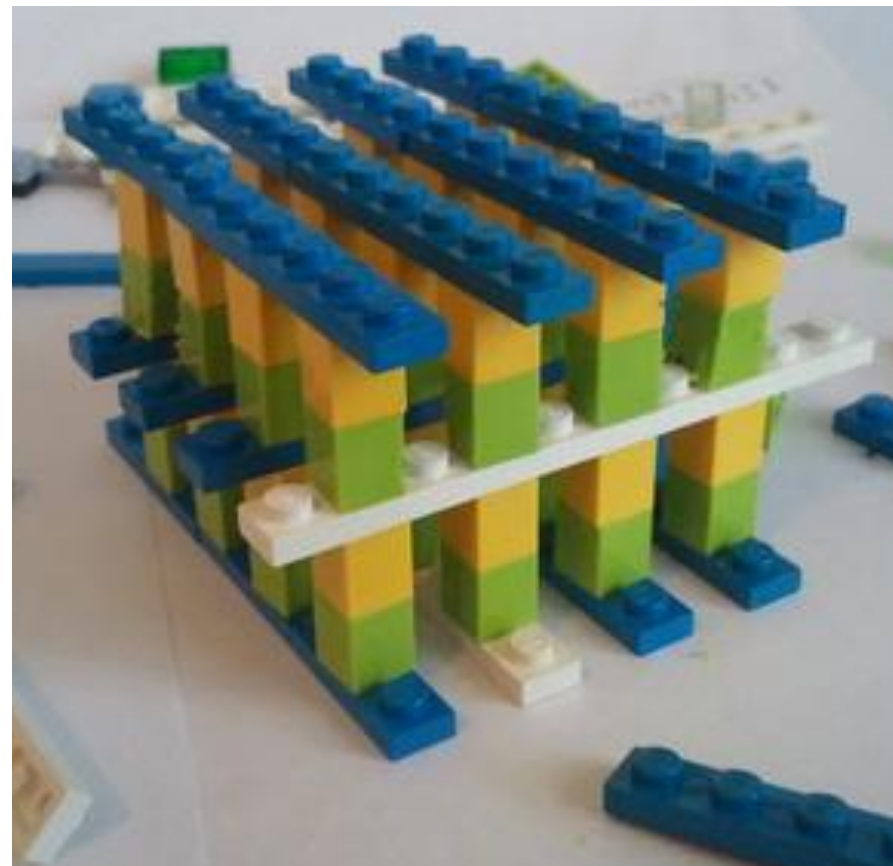
- A controversial term
 - As well defined as Software Defined
- For me:
 - Inherently persistent
 - Latency between DRAM and NAND Flash
 - Bit addressable
 - Both material and usage





Storage Class Memories Today

- 3D Xpoint
 - SSDs not a huge success
 - So far
 - DIMMs show promise
 - 2nd gen still to come (Micron?)
 - Gen I is 3D but only 1 cell deep
- Everspin Spin-transfer Torque MRAM
 - 1Gb/chip @ 28nm
 - NAND 1.33 Tb/chip
 - DRAM replacement on SSDs
- Others SciFi





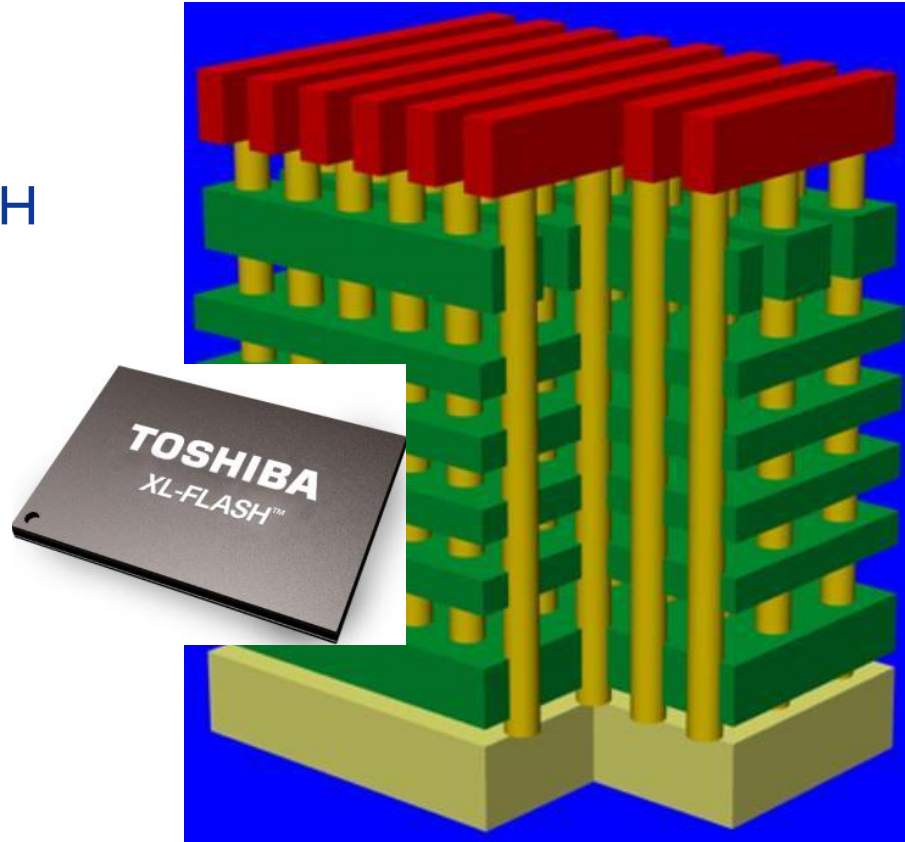
SCM in Enterprise Storage

- HPE
 - Optane AIC in controller
 - 3PAR and Nimble for cache
 - Back-ends still SAS
- Dell EMC PowerMax
 - Optane D4800X (dual port)
 - Tier of storage
- Mostly HCI/SDS Optane SSDs



SLC Returns

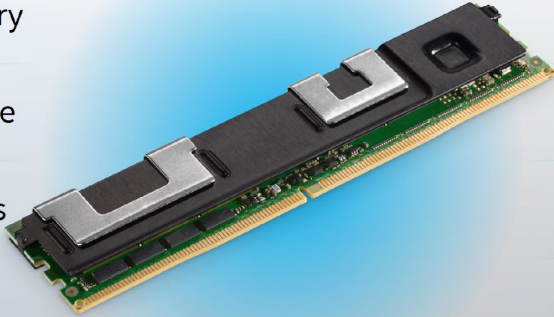
- Samsung Z-NAND
- Kioxia (AKA Toshiba) XL-FLASH
 - Multi-plane for parallelism
 - 4 KB page
 - 128 KB in 1Tb QLC
 - 128 Gb/die
 - X μ sec read latency
- Still flash w/write asymmetry
- SSDs today
 - Flash DIMMs seem passe





Optane DIMMs

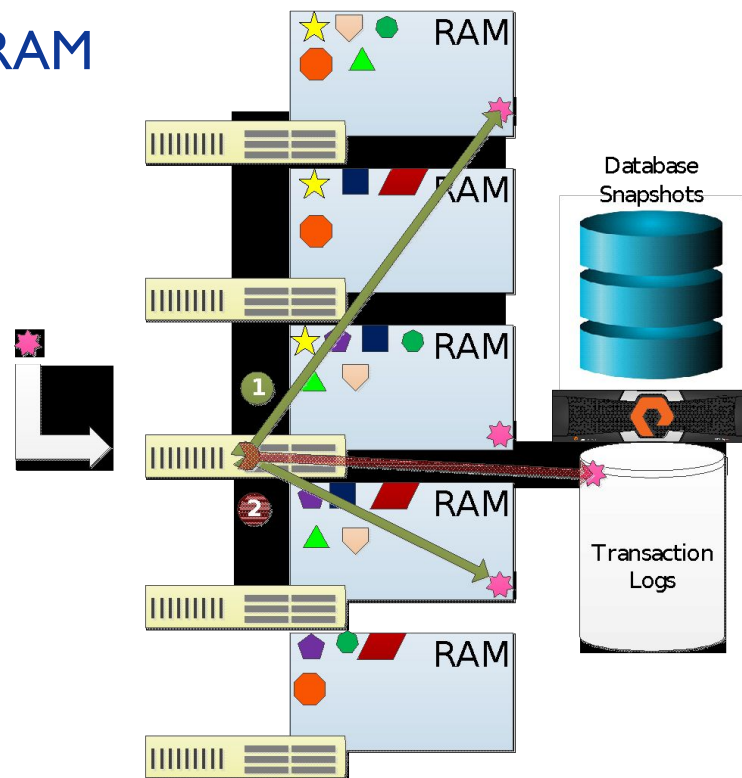
- Require latest Xeons
 - Special models for large memory addresses
- OS/Hypervisor support as PMEM
- Complex programming models

Big and Affordable Memory		128, 256, 512GB
High Performance Storage		DDR4 Pin Compatible
Direct Load/Store Access		Hardware Encryption
Native Persistence		High Reliability



In Memory Databases Today

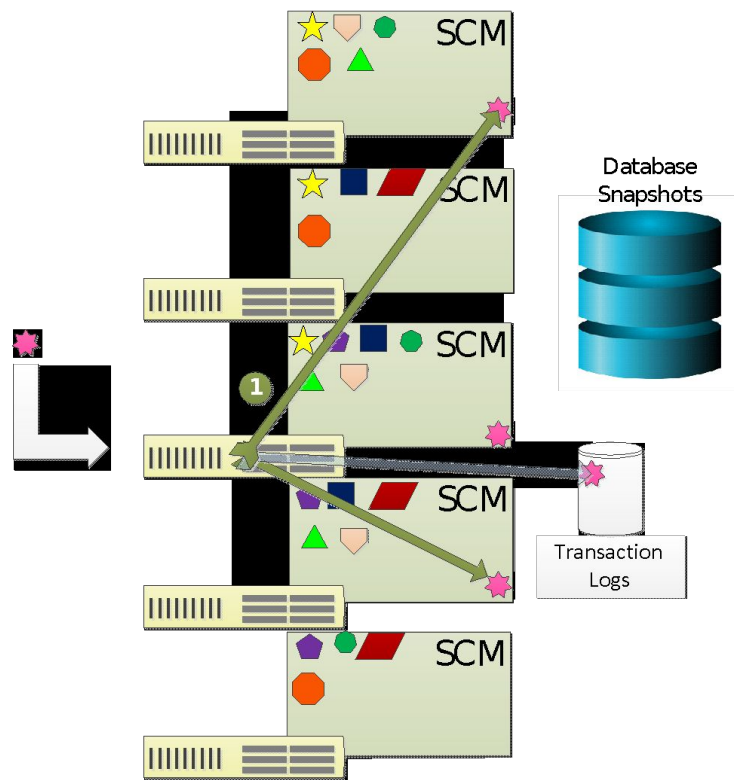
- All database operations performed in RAM
- Data replicated across nodes (x86)
- AFA/HCI back end for persistence
 - Snapshots
 - Transaction Logs
 - Playback in case
- On write:
 1. Replicate to 1-n nodes
 2. Write to persistent log (typically AFA)
 3. ACK





In Memory Database with SCM

- Much larger capacity/node
 - 512GB vs 64GB/DIMM
 - 10X latency (SWAG)
- Lower cost /GB
 - 2-10X we guess
 - More vs 128GB LRDIMMs
 - 3X cost of 64GB
- ACK after n-node write
 - Can be RDMA write
 - Data now persistent
 - Log writes can be aggregated, async





Flash Memory Summit

SAP HANA

SAP HANA Native Support for Persistent Memory Officially Supported in SAP HANA 2.3 (April 2018)

Larger memory capacity with high performance (vs. DRAM & lower tier storage)

Lower TCO data storage hierarchy

Faster start time delivers less downtime

Co-innovation with Intel® leads to first fully optimized major DBMS platform

Early Adoption Program with key partners/customers ongoing

Persistent Memory
non-volatile

Data Reliability
faster starts



Higher Capacity
than DRAM

Transforming
the memory hierarchy

Intel® Optane™ DC persistent memory available in 1H 2019



Benefit

Process more data in real-time at a lower TCO with improved business continuity

> 3 TB

Increased total memory capacity per CPU

12.5x

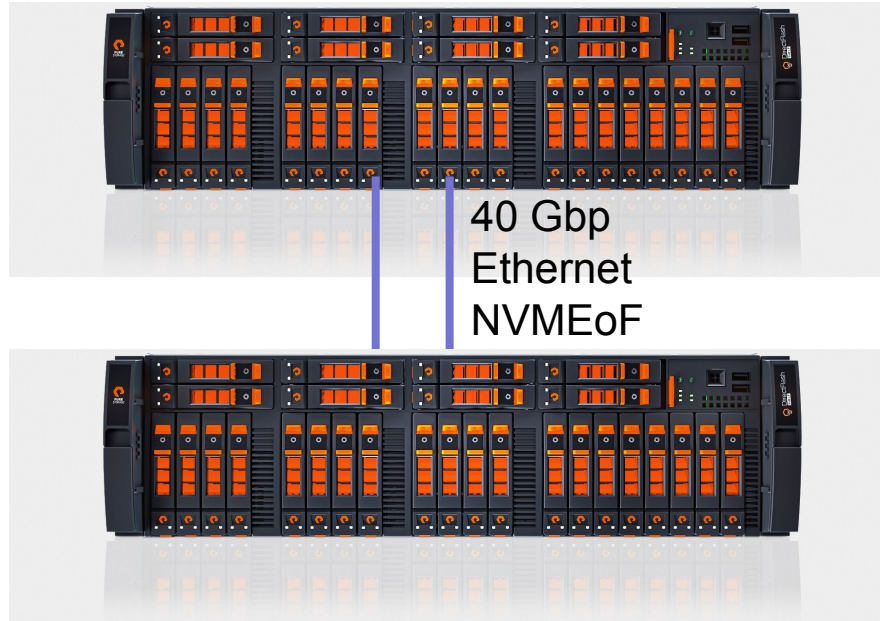
Improvement in startup time*

First major DBMS vendor to officially support Intel Optane DC persistent memory!

sap.com/persistent-memory



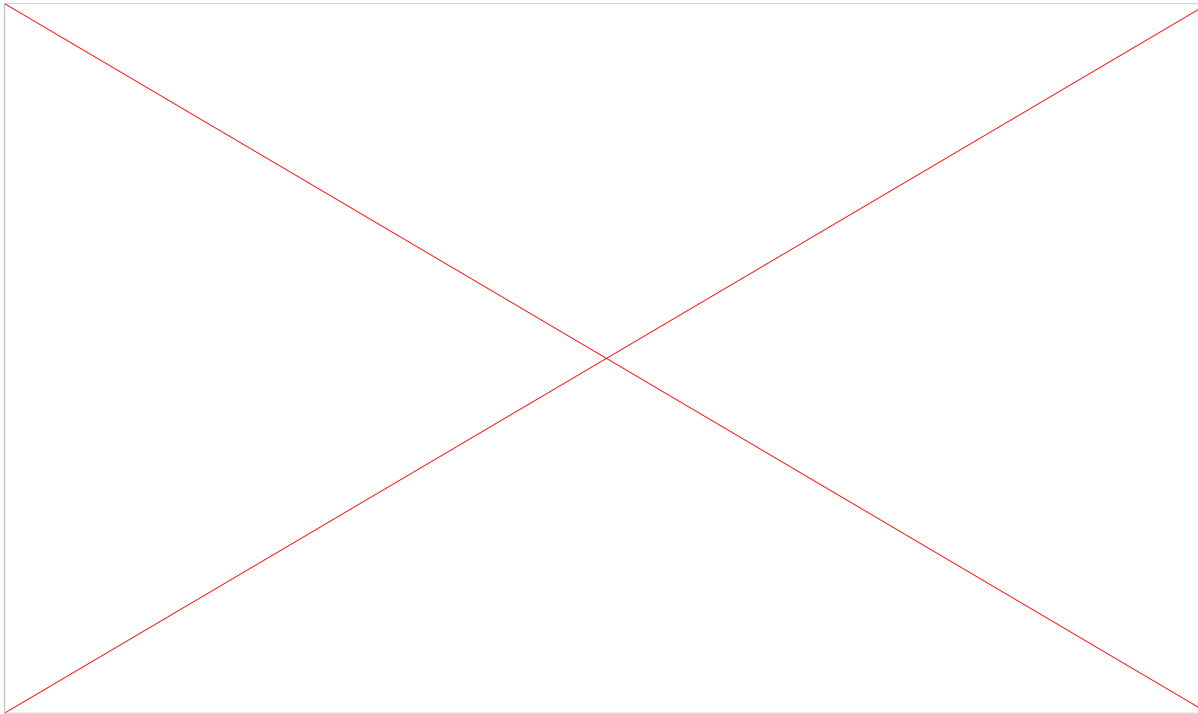
Pure FlashArray//x



- Replaces //m
 - SAS SSDs
 - Expansion via SAS or NVMeoF JBOF
- NVMeoF target on 40Gbps Ethernet
- Full services



Kaminario K2 Composable

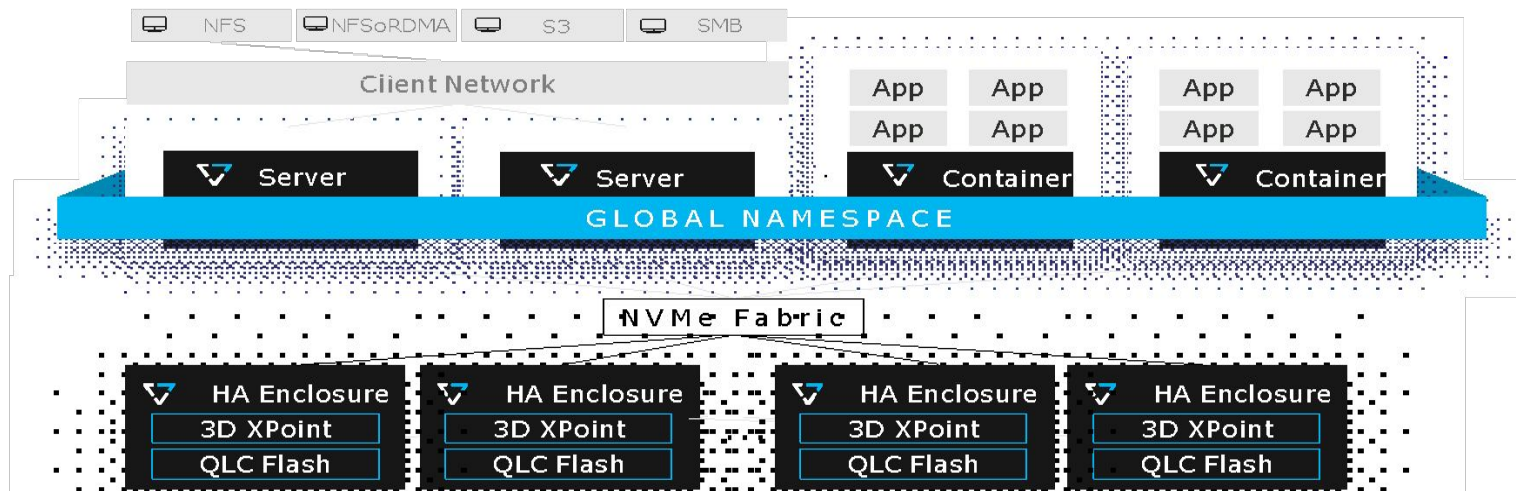


- NVMEoF
 - Controller to JBOF
 - Host to array (opt)
- Dynamically assign controllers and flash to virt array



VAST Data Universal Storage System

- 3D XPoint and QLC Flash in HA NVMe-oF JBOFs
- Storage services via stateless servers (metadata in XPoint)
 - File and object Global Name Space
 - Data reduction Erasure coding
- Global FTL





Flash Memory Summit

