Leveraging NVMe-oF™ for Existing and New Applications

Sponsored by NVM Express™ organization, the owner of NVMe™, NVMe-oF™ and NVMe-MI™ standards

Speakers

Marcus Thordal

BROADCOM

Bryan Cowger

KAZAN NETWORKS

Erez Scop

Mellanox TECHNOLOGIES

Nishant Lodha

MARVELL

Agenda

In this panel session we will discuss application and use case examples leveraging NVMe-oF™

Which improvements should you expect with NVMe-oF and how does this apply to different types of applications.

Learn from early adopters implementations of NVMe-oF, what you need to consider when planning to migrate existing applications from SCSI to NVMe™ or deploying new applications with NVMe-oF

We will complete the session with a peak into the future of how NVMe-oF can enable new application use cases.

Application Requirements & NVMe™ Fabric Selection

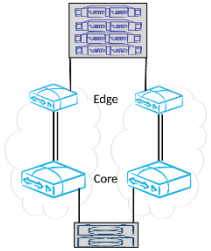
Enterprise Applications

Rigid application storage architecture

Application architedcture cannot be changed.

Fabric requirements:

- ✓ High Reliability
- ✓ Low Latency
- ✓ Deterministic Performance
- ✓ Scale



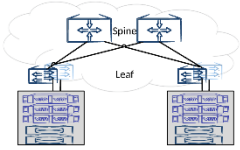
Hyper-Scale Applications

Full control of application storage architecture

Application can be architected and composed to be less dependent on Fabric Properties

Fabric requirements:

- ✓ Reasonable Reliability
- ✓ Low Latency
- ✓ Deterministic Performance?
- ✓ Scale needs can be confined to rack locations



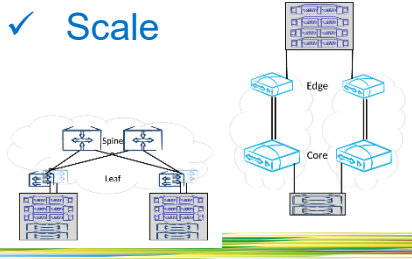
Analytics, ML & AI

Architecture traditionally DAS Changing to centralized shared storage

Science project or Enterprise dependent Application?

Fabric requirements:

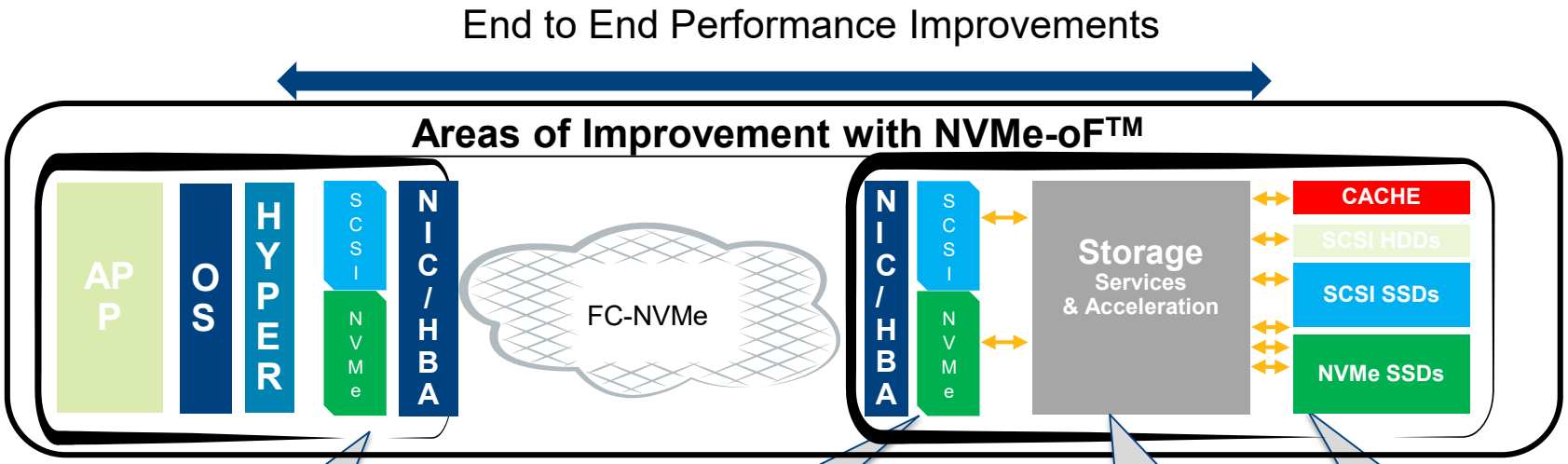
- ✓ Reasonable/High Reliability
- ✓ Low Latency
- ✓ Deterministic Performance
- ✓ Scale



Flash Memory Summit

nvm EXPRESS®

Areas of Performance Improvement with NVMe™ over Fabrics



Server
Performance Improvement with shorter path through the OS storage stack with NVMe™ & NVMe-oF™
-and lower CPU utilization

Front side of Storage Array
Performance Improvement is a shorter path through the target's host port stack

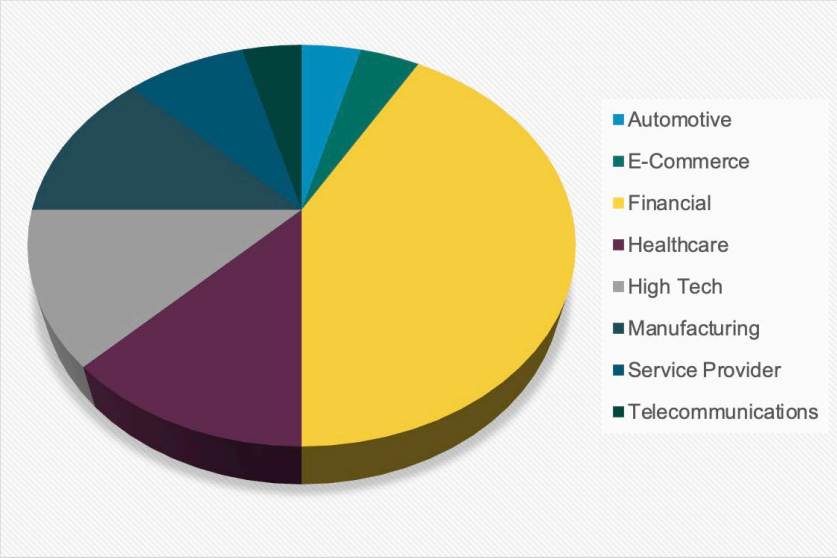
Storage Array Architecture
Performance Improvement is an optimized path through the controller

Back side of Storage Array
Performance improvement by moving from SAS/SATA drives to NVMe SSDs

NVMe/FC – Adoption

Slide from Webinar:
Real World Performance Advantages with NVMe over Fibre Channel

Customers deployed, testing, or planning to test NVMe/FC



Use cases:

- Accelerate business critical application
- Accelerate Oracle and SQL application
- Future proofing – Investment protection

Market Dynamic: Business critical apps run on SAN

- >70% of installed SANs are FC
- Market research suggests continuing reliance and dominance of FC SAN

NetApp

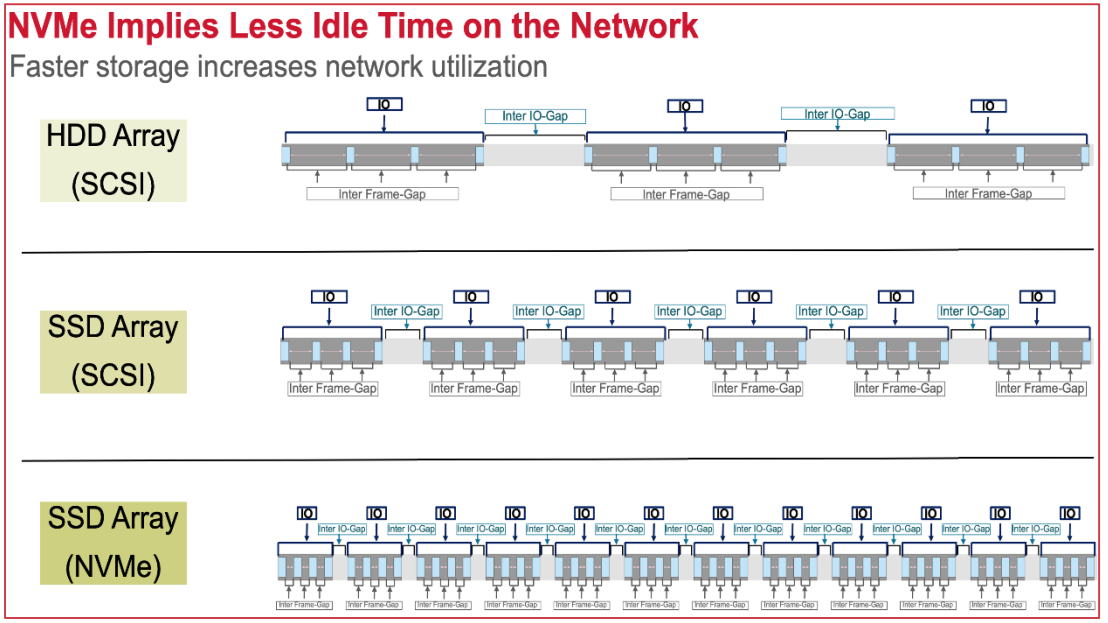
Flash Memory Summit

nvm EXPRESS®

The Future with NVMe-oF™

NVMe-oF™ will change how storage is consumed

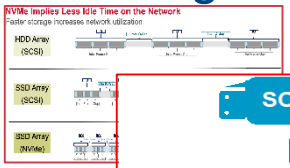
- *Storage Fabric Impact*



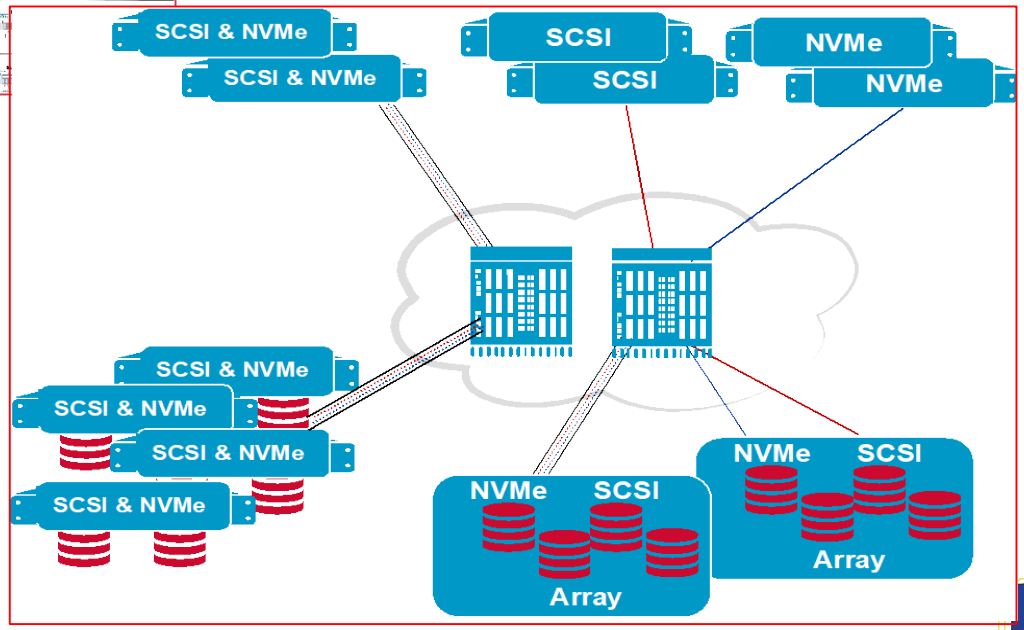
The Future with NVMe-oF™

NVMe-oF™ will change how storage is consumed

- *Storage Fabric Impact*



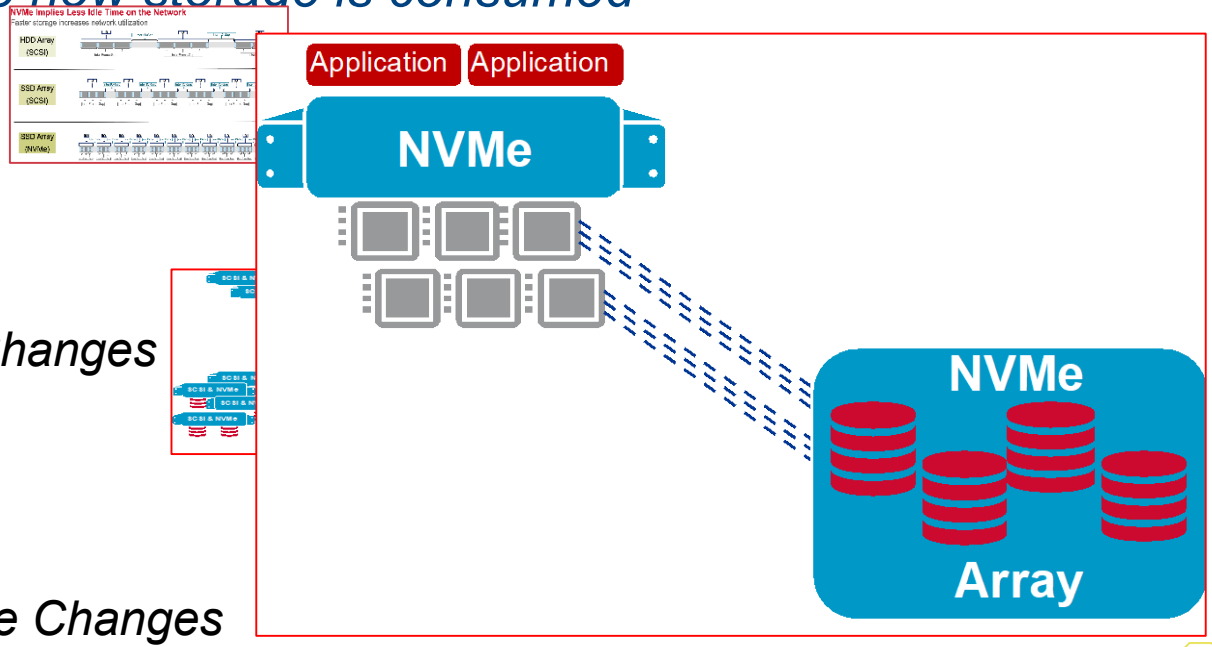
- *Storage Architecture Changes*



The Future with NVMe-oF™

NVMe-oF™ will change how storage is consumed

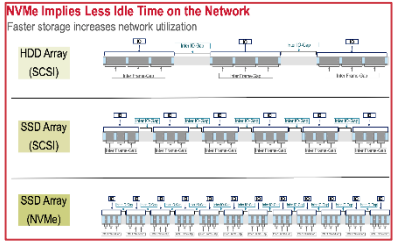
- *Storage Fabric Impact*
- *Storage Architecture Changes*
- *Application Architecture Changes*



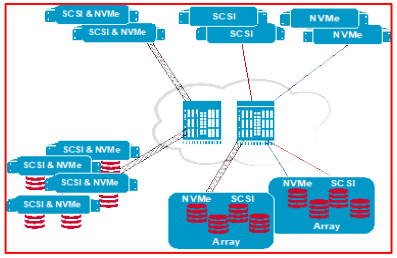
The Future with NVMe-oF™

NVMe-oF™ will change how storage is consumed

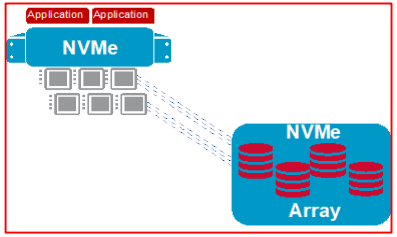
- *Storage Fabric Impact*



- *Storage Architecture Changes*



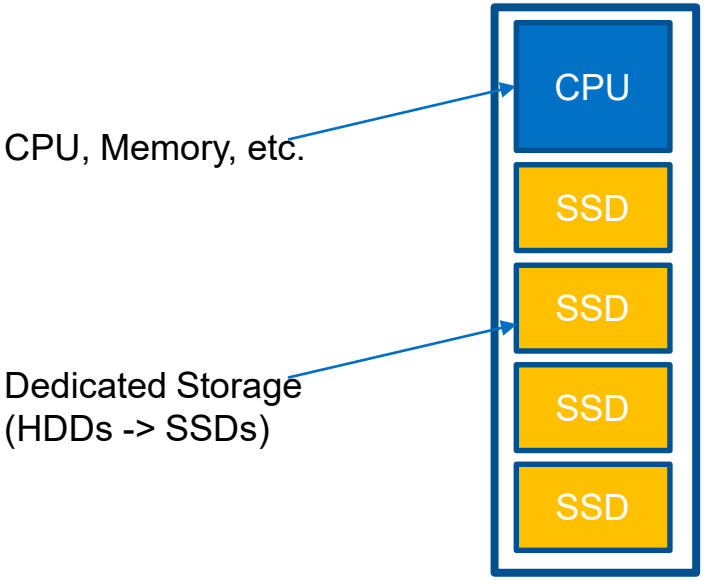
- *Application Architecture Changes*



Agenda

- Composable Infrastructure – enabled by NVMe-oF™
- Various composable storage options
- Hardware-centric architecture
- TCP vs. RDMA transport options
- Performance results using HW-centric designs

Today's "Shared Nothing" Model
a.k.a. DAS

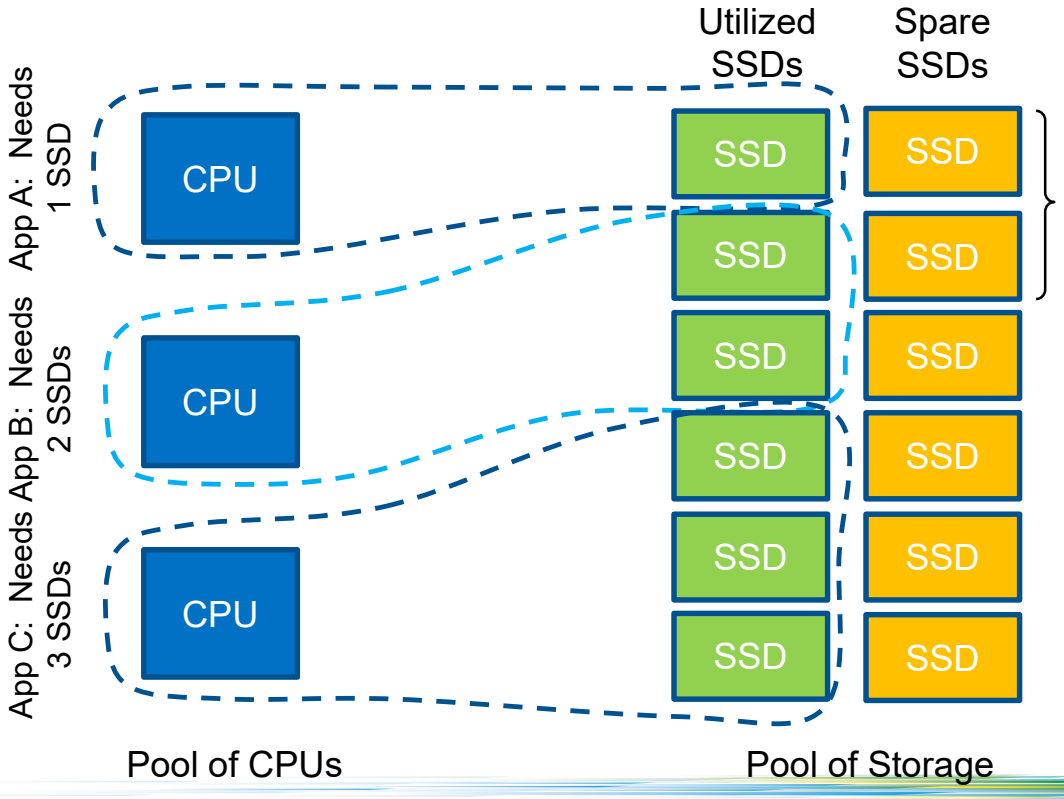


Challenges:

- Forces the up-front decision of how much storage to devote to each server.

- Locks in the compute:storage ratio.

The Composable Datacenter

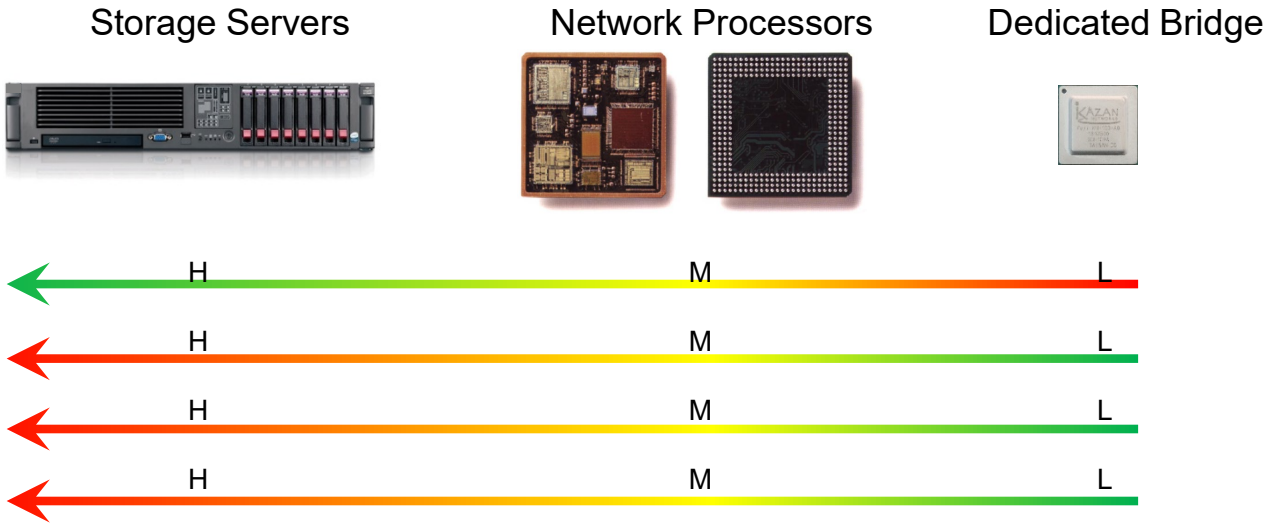


Spares / Expansion Pool

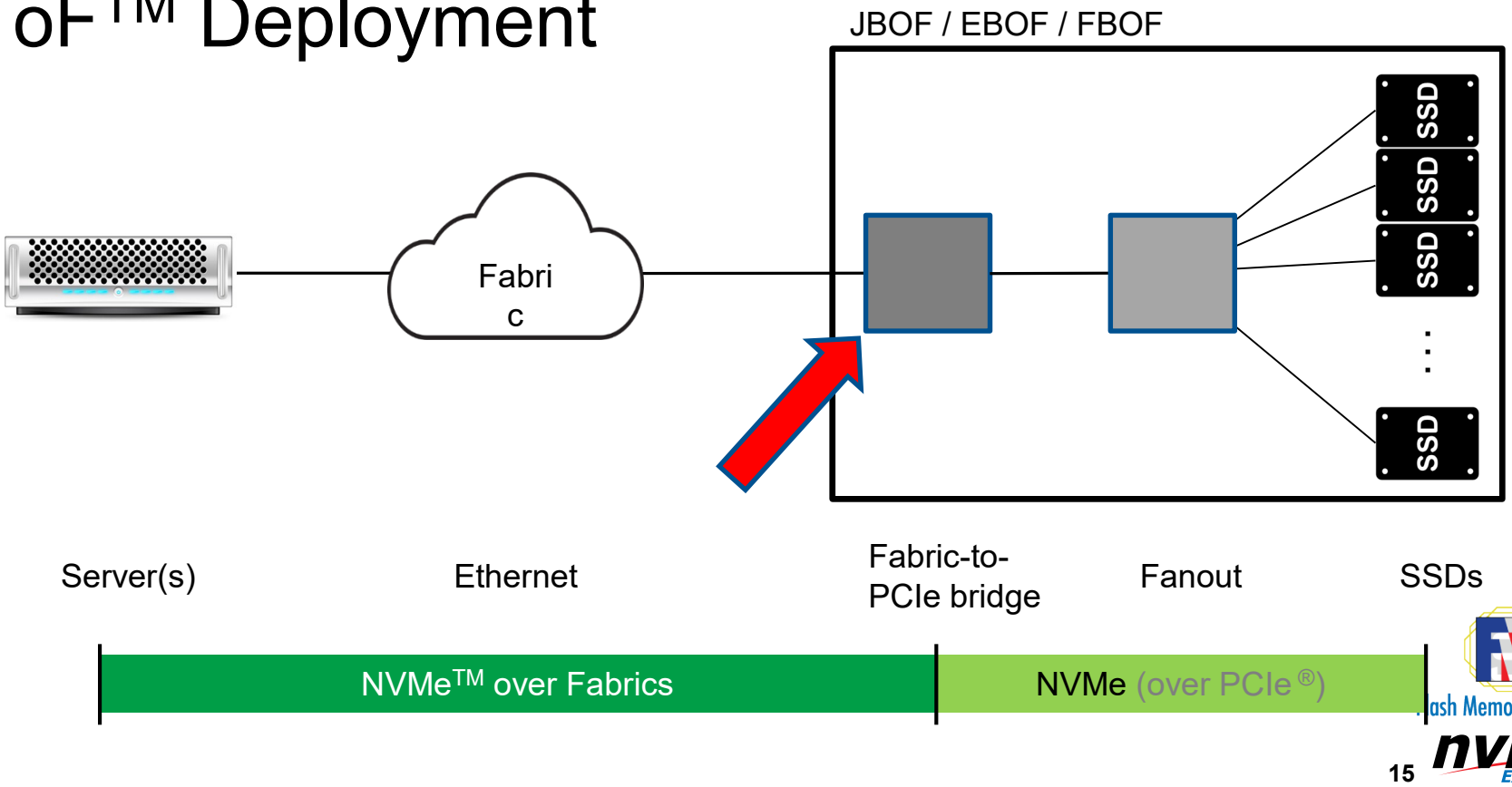
- Minimize Dark Flash
- Buy storage only as needed
- Power SSDs only as needed

A Spectrum of Options

Myriad choices for NVMe-oF™ / Composable Infrastructure target deployments

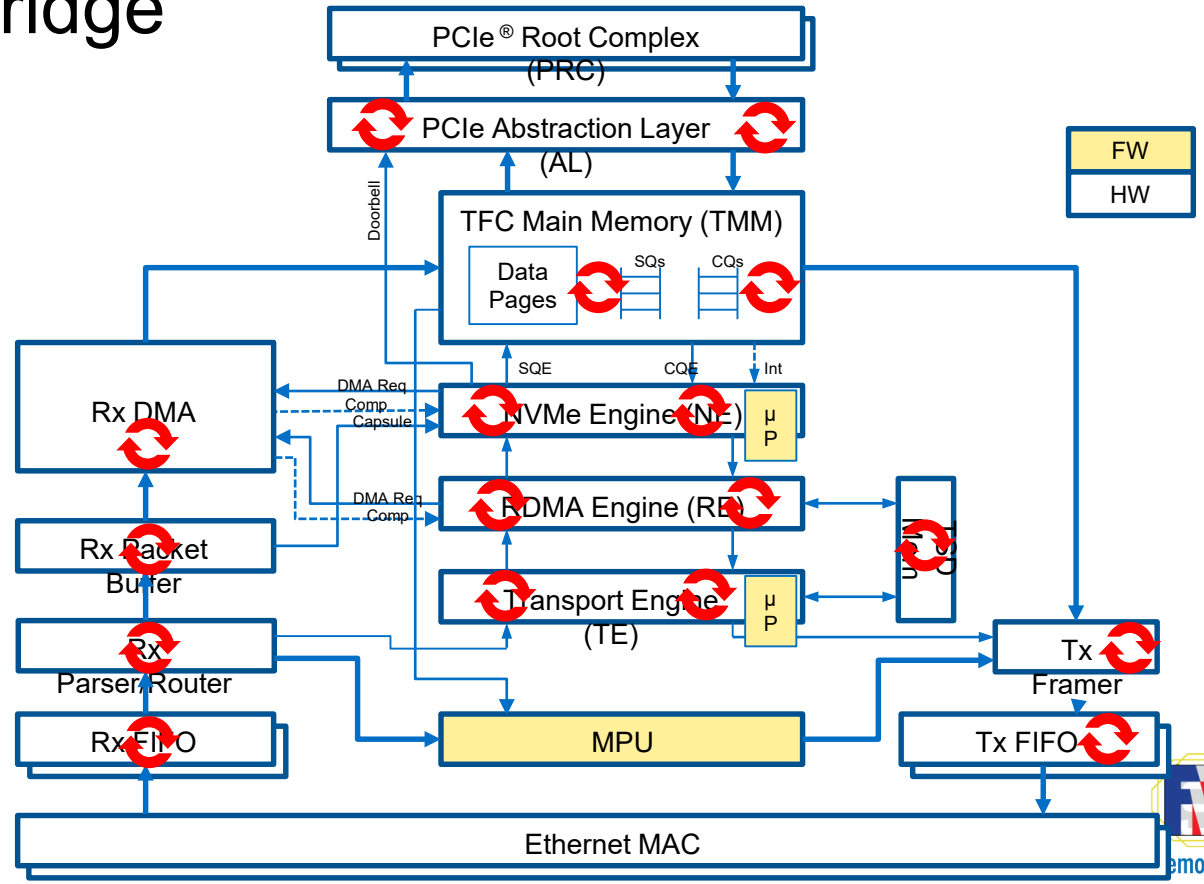


Typical NVMe-oF™ Deployment



NVMe-oF™ Bridge Architecture

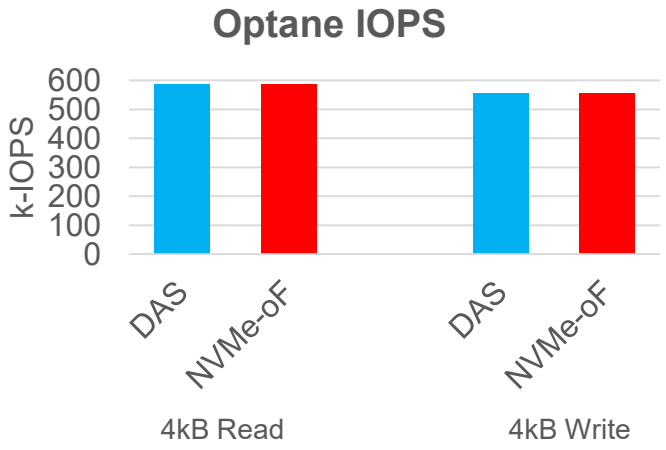
- Approximately 150 FSMs running in parallel
- Some simple, e.g. buffer management
- Some quite complex:
  - RoCE v1, v2
  - iWARP /TCP
  - TCP
  - NVMe™
  - NVMe-oF™
- Slow-path implemented in FW

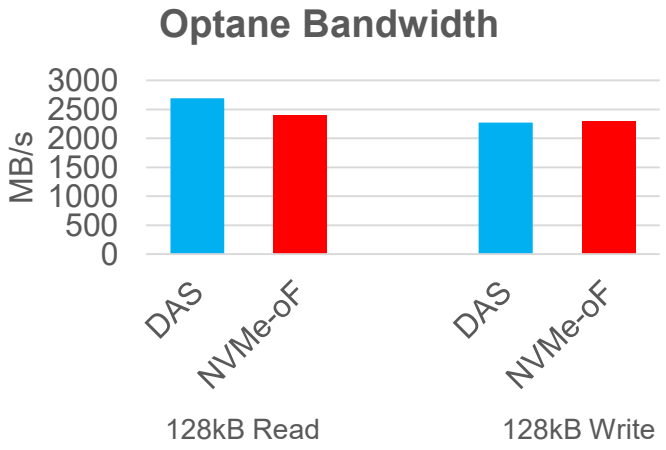


Performance
Across Various Transport Protocols

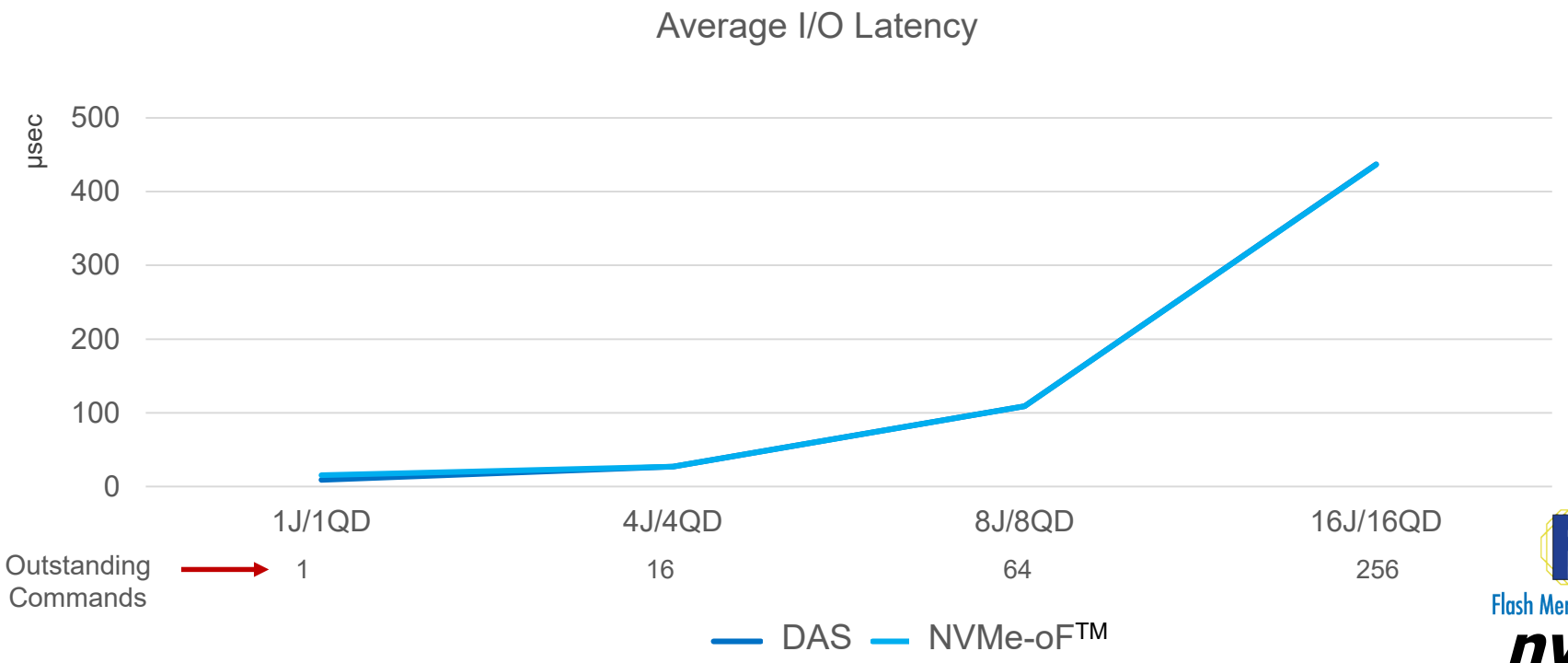
Configuration	Transport Protocol	4kB Read (M IOPS)	4kB Write (M IOPS)	128kB Read (GB/s)	128kB Write (GB/s)
Fuji Config: Ethernet: 2x50Gb PCIe: 2x8 gen3 Enclosure: Kazan "K8" 4 SSDs per PCIe port SSDs: 8 WDC 1.6TB "Gallant Fox"	RoCE v1	2.76	1.08	11.8	10.3
	RoCE v2	2.79	1.10	11.5	9.9
	iWARP	2.77	1.17	11.8	9.7
	TCP	2.65	1.0	11.2	9.6

Single SSD Performance
Optane™

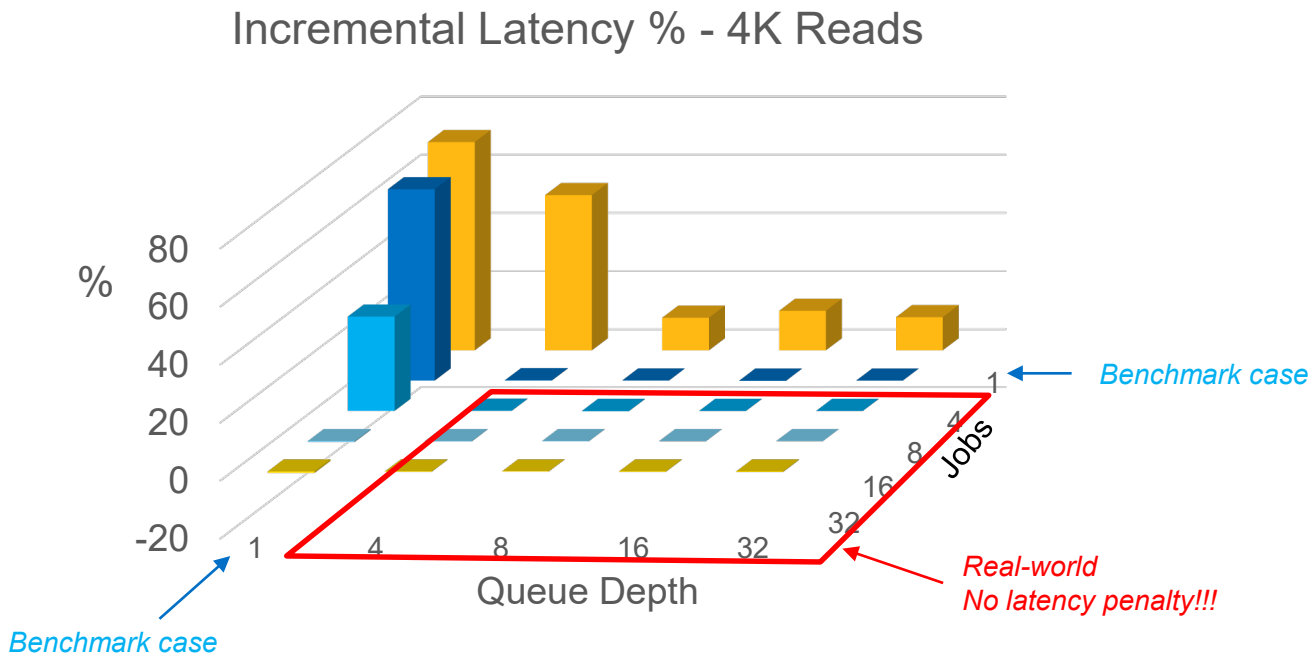




Loaded Latency – Optane™
vs. Outstanding Commands



Incremental Latency
Optane™ SSD Target



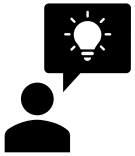
Summary / Takeaways

- Lots of options now hitting the market to enable Composable Infrastructure
- Very low-cost, low-power options for JBOFs
- No loss in performance compared to DAS
  - IOPS, bandwidth, and latency all equivalent

Mellanox NVMe™ SNAP Use Case

Erez Scop, Mellanox Technologies

Storage Disaggregation



Motivation

Grow storage and/or compute independently

- No local disks needed (disk-less)
- Move local NVMe drives to centralized location
- Higher performance per node
- Immediate CAPX saving
- Lower MTBF

Problem

Requires software changes

- RDMA software stack
- NVMe-oF™ drivers – limited OS support
- Different management

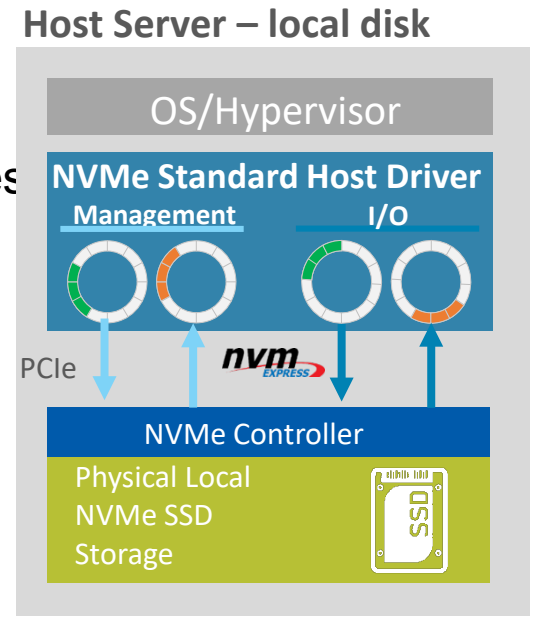
Solution

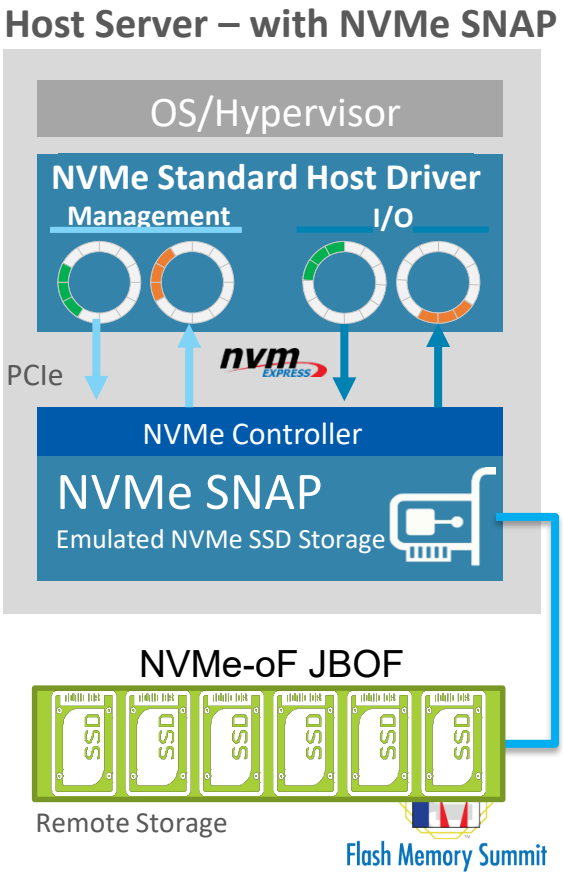
NVMe™ SNAP

- Compute nodes see NVMe local drives (Emulated in-hardware)
- Zero software changes
- Supported on all OSs
- Latency as local NVMe drive
- Bandwidth up to network available (100Gbps and above)

NVMe™ SNAP

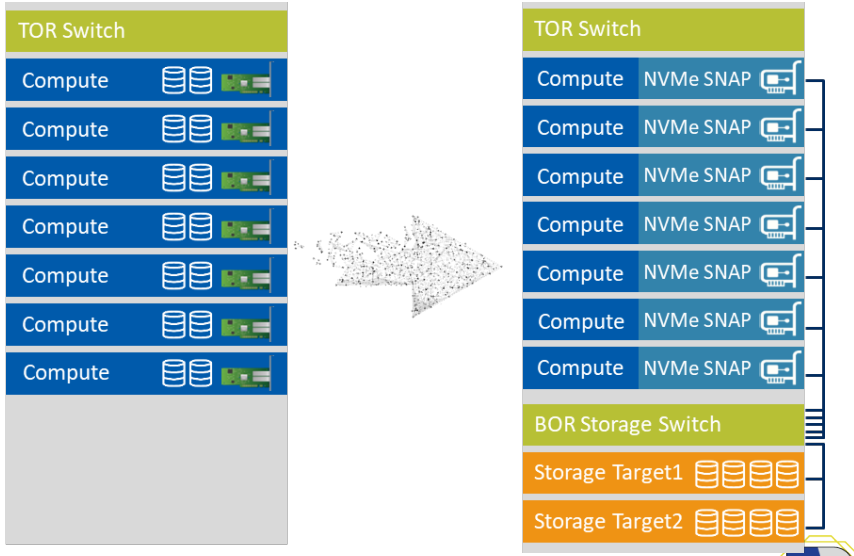
- Emulated NVMe™ PCI drives
- OS agnostic
- Software defined
- Hardware accelerated
- Bootable
- NVMe SRIOV support





Solution – Disaggregation with NVMe™ SNAP

- Utilizing NVMe-oF™ latency and throughput for ‘local feel’
- Scale Storage independently
- Scale Compute independently
- Save $$ on Data Center Storage
- Improved MTBF

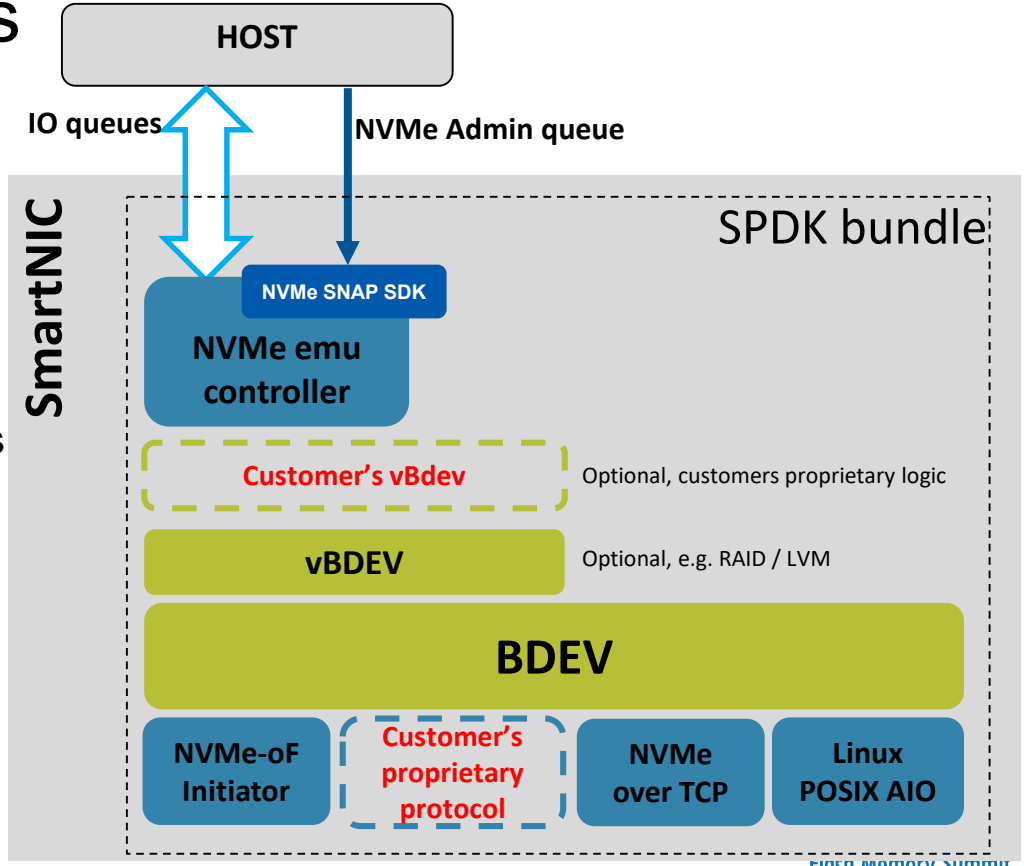


NVMe™ SNAP Internals

SPDK advantages
- Efficient memory management
- Zero-copy all the way
- Full polling
- Multi queues, multi threads, lockless
- Well defined APIs: vBdev, Bdev drivers

NVMe SNAP emulation SDK
- Handle NVMe™ registers and admin

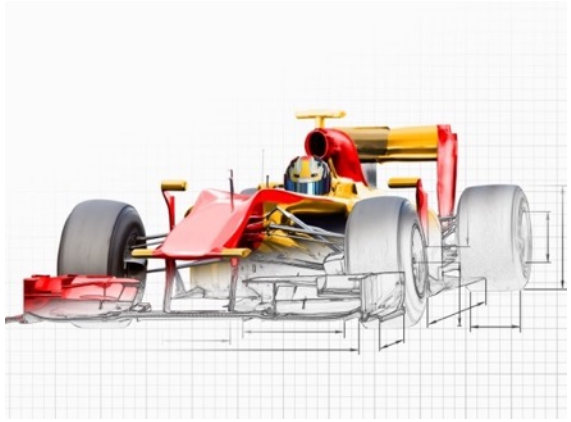
Customer's proprietary code
- BDEV: for proprietary storage network protocols
- vBDEV: for per-io routing decisions, RAIDs, etc



Use Case driven Fabric Choices – Lessons Learnt

Nishant Lodha, Marvell

Making the right “fabric” choice!



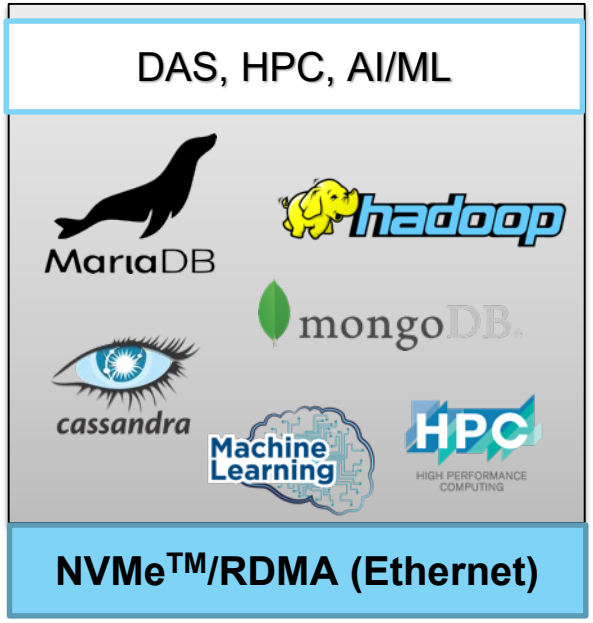
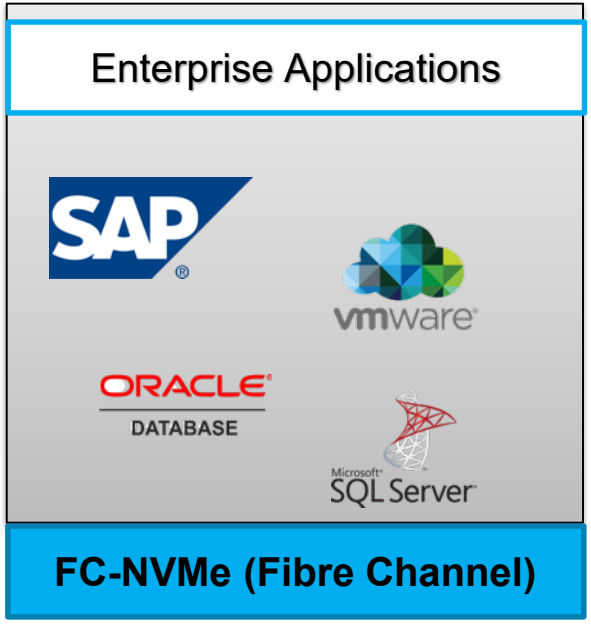
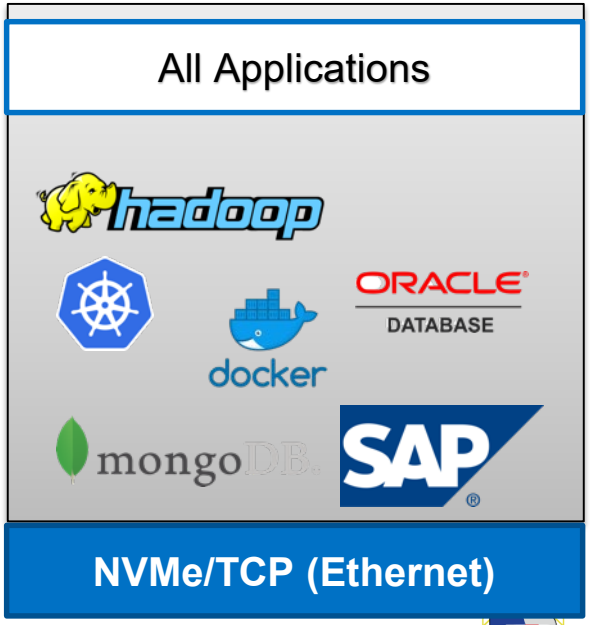
Not “just” about “fabrics” performance

Culture and Install Base

Use Cases

Use Cases by Fabric

No one size fits all!

DAS, HPC, AI/ML	Enterprise Applications	All Applications
		
NVMe™/RDMA (Ethernet)	**FC-NVMe (Fibre Channel)**	**NVMe/TCP (Ethernet)**
Performance at the cost of complexity	Leverage existing infrastructure. Reliability is key	Simplicity is key. Balance of performance and cost

Logos are indicative of workload characteristics only.

New: Hyper Converged meets Dis-aggregated

Challenge

- Scale Storage Independent of Compute for HCI

Solution

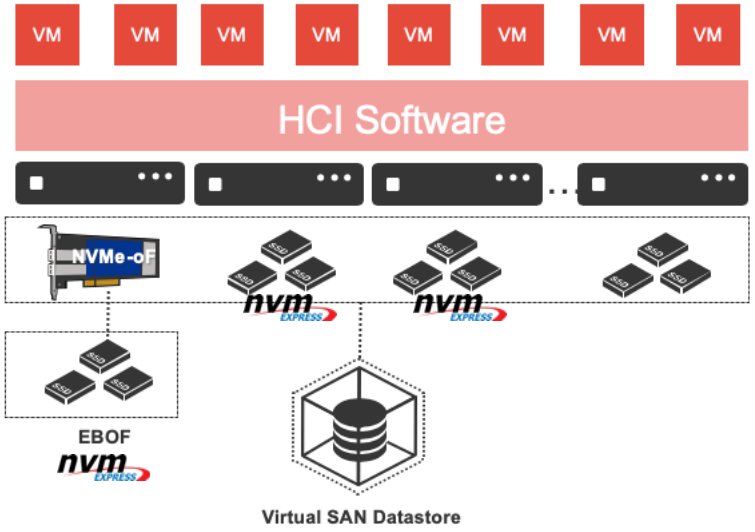
- Pool local and remote NVMe™ by deploying NVMe-oF™ connected EBOF

Benefits:

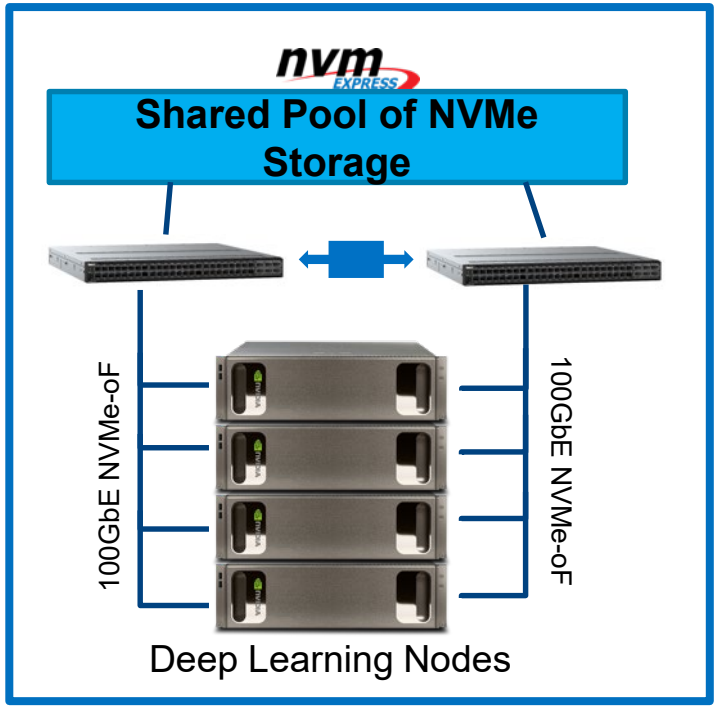
- Retain the simplicity of HCI management and provisioning
- Deliver storage services in software
- Reduce captive storage in a server

When:

- Next generation HCI fabrics being enabled to consume external EBOF



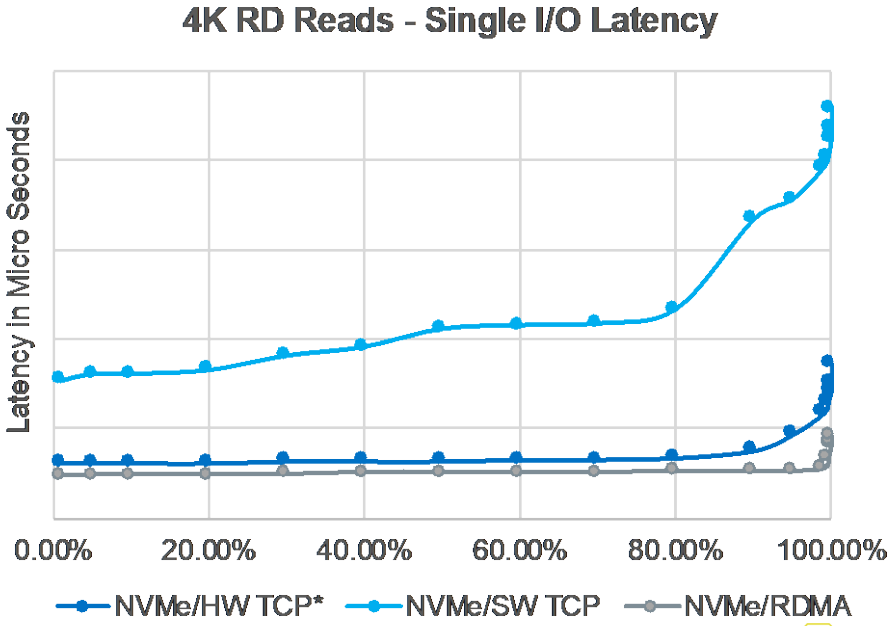
New: External Storage for Deep Learning Clusters



- Deep Learning architectures require access to low latency and high capacity storage
- Storage Pools vs. Captive per node storage
- A 25/100GbE NVMe-oF™ fabric provides excellent scalability
- Delivers a high-performance data platform for deep learning, with performance on par with locally resident datasets

Web 2.0 Use Cases – Average does not cut it!

- Web 2.0 Traffic require “consistent” response times from underlying storage and networking infrastructures.
- Fabric decisions based on “Average” NVMe-oF™ latencies are just not ”ok”
- Tail latency measurements indicate how well the outliers perform
- I/O cost to CPU also helps predict latency variances in cases of heavy load



Source: Marvell Internal test labs.

32

Questions?


nvm
EXPRESS®
Architected for Performance