NVMe-oF™ JBOFs

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

JBOF Track Speakers

Bryan Cowger

Nishant Lodha

Peter Onufryk

Fazil Osman

Sujoy Sen

JBOF Session Agenda

- Market Overview
- Composable Infrastructure
- PCIe (direct-attached) JBOF
- Fabric-attached FBOF
- Management Options
- Remaining Challenges
- Q & A

Market Overview

Nishant Lodha
Marvell Semiconductor

Storage Trends from all around!

WW Enterprise Storage spend growing (~$42B(2016) → ~$47B(2020))

- Scale up → Scale Out (Hyperscale – public cloud driven by 3rd platform – mobile, social, cloud, analytics)
- ECB revenues stay flat ($25B) – Flash driving enterprise storage @ 26.2% CAGR; HDD declining @ 14.5% CAGR

Traditional storage deployment models being disrupted!

- Proprietary/siloed architectures → Software Defined Storage (SDS)/Hyper Converged (HCI) on commodity HW
- Direct Attach Storage (DAS) → Disaggregated storage (JBOD → JBOF, FBOF)

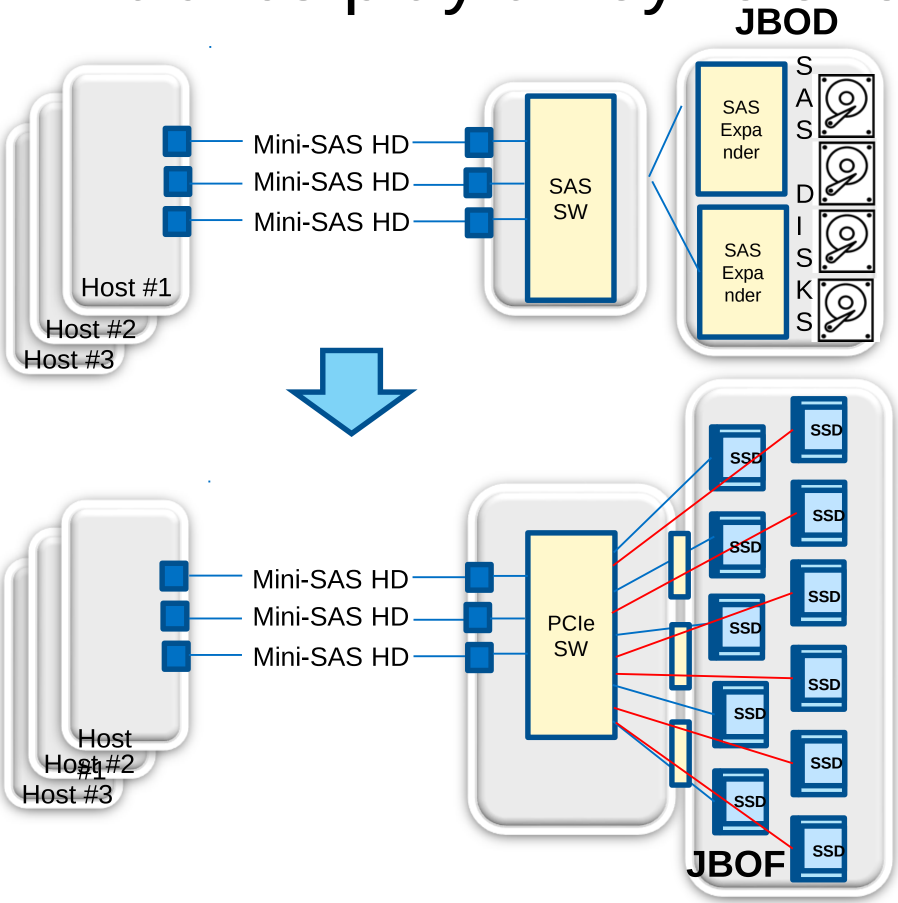
Faster media necessitates new protocol, drives faster interconnects & enables new use cases

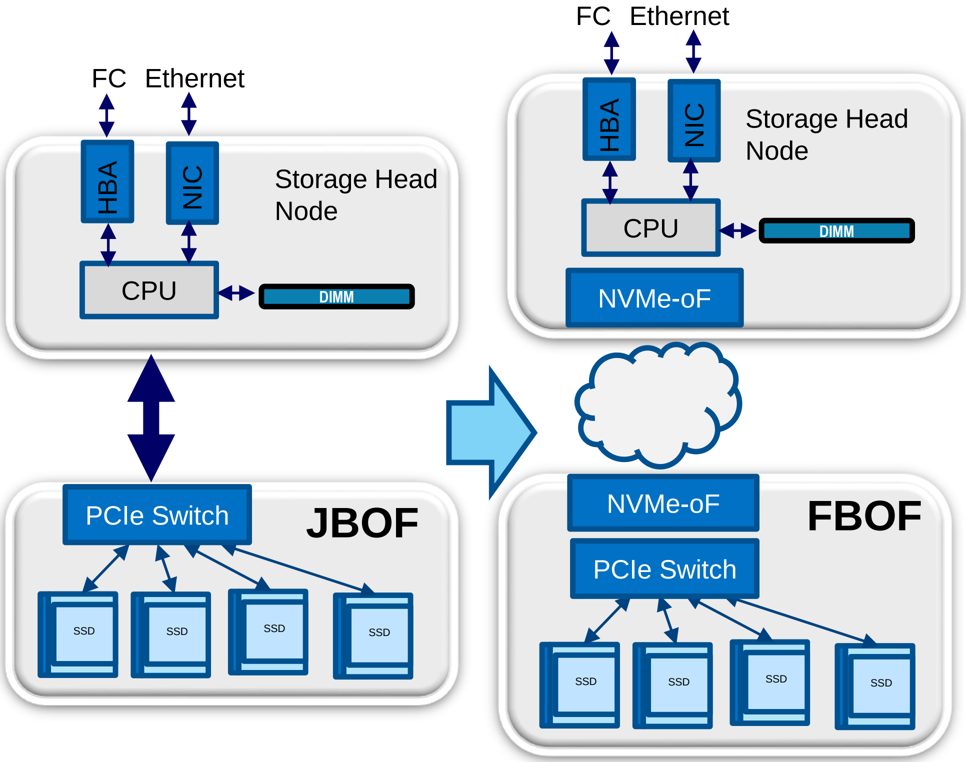
- NVMe™ will displace SCSI as the dominant block storage protocol by 2020 for AFA/CI/Scale-out
- Shared NVMe storage over a variety of Fabrics with NVMe-oF (RDMA (Eth, IB), FC, *TCP*)
- Emerging 3D Xpoint enables storage class memory (SCM)/persistent memory (PMEM)

Cloud storage for Enterprise customers iffy!

- Cost savings questionable; Data security concerns
- Hard to migrate legacy storage; Public cloud SaaS for email/collaboration

Fabrics play a key role for JBOFs -> FBOFs





Scaling our NVMe™ Requires a (Real) Network

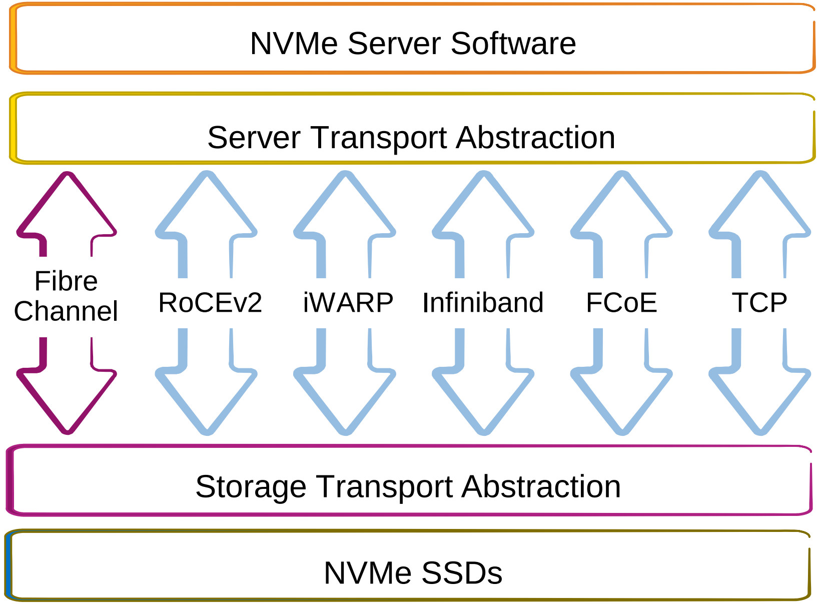
- Many options, plenty of confusion, conversation beyond PCIe®
- Fibre Channel is the transport for the vast majority of today's all flash arrays

  FC-NVMe Standardized in Mid-2017

- RoCEv2, iWARP and InfiniBand are RDMA-based but not compatible with each other

  NVMe-oF RDMA Standardized in 2016

- FCoE is a fabric is a option
- NVMe over TCP - making it way through the standards



Flash Memory Summit

nvm EXPRESS®

RDMA is Most "Considered", Challenges Remain

Infrastructure and Skillset change required!

Congestion	Skillset Requirements	Backward Compatibility
Not Automatic **Not Precise**	Keeping the network '**lossless**' RDMA/**OEFD** expertise	RNIC **Upgrade Required** **RDMA Camps**

Flash Memory Summit

nvm EXPRESS®

New This Year! NVMe-oF™/TCP

Defines a TCP Transport Binding layer for NVMe-oF

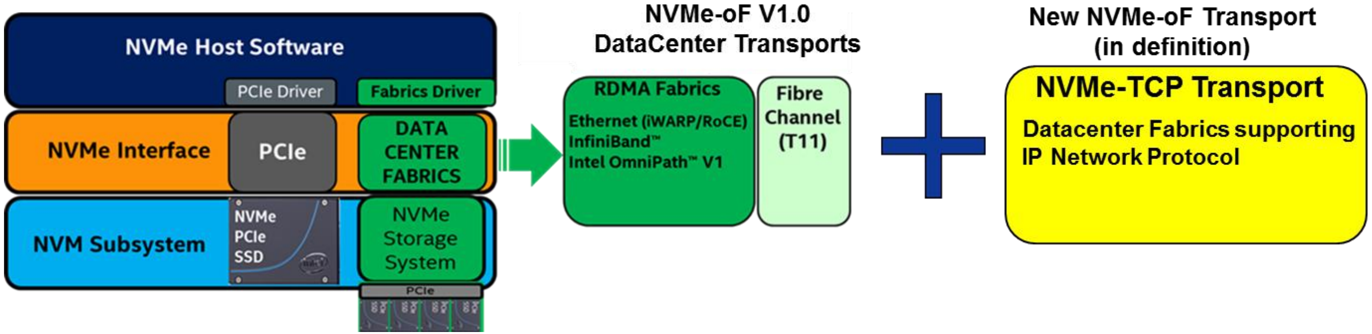
Promoted by Facebook, Google, DELL EMC, Intel, Others. Sweet spots for JBOF/FBOFs

Not RDMA-based

Not yet part of the NVMe-oF standard, Likely in 2018/19

Enables adoption of NVMe-oF into existing datacenter IP network environments that are not RDMA-enabled

TCP offload required to leverage Flash potential



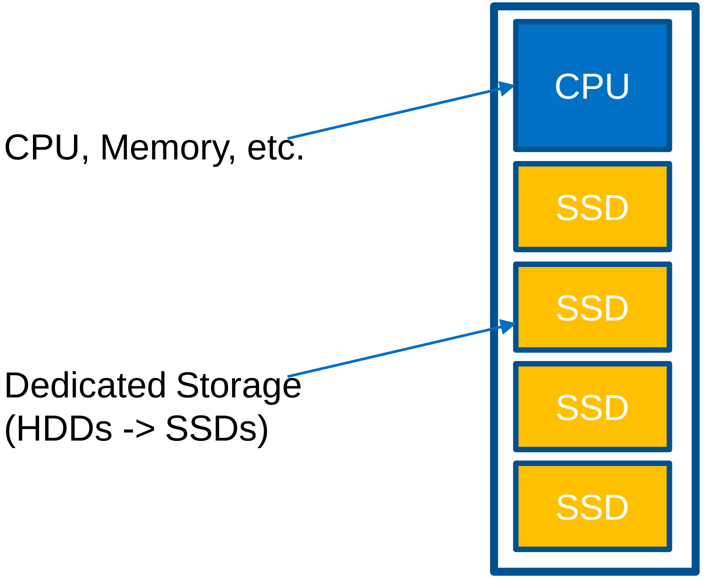
Flash Memory Summit

Composable Infrastructure

Bryan Cowger

Kazan Networks

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

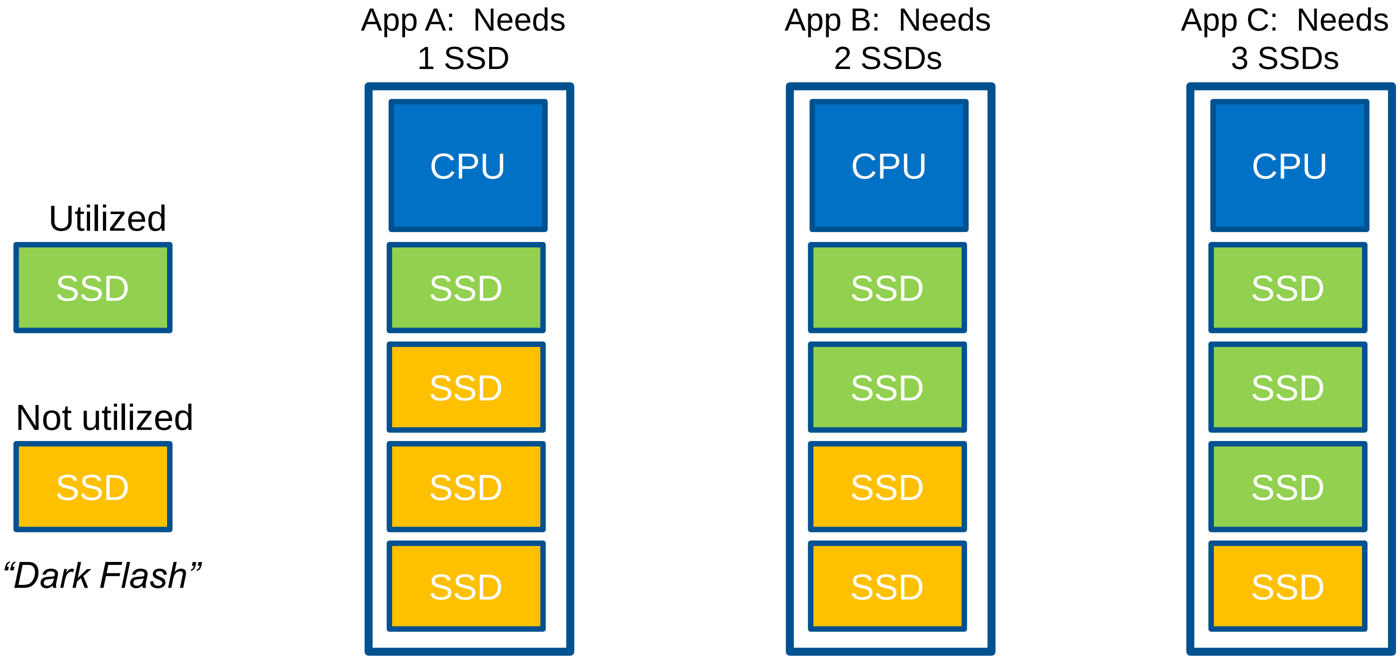


Challenges:
- Forces the up-front decision of how much storage to devote to each server.

- Locks in the compute:storage ratio.

Shared Nothing Model

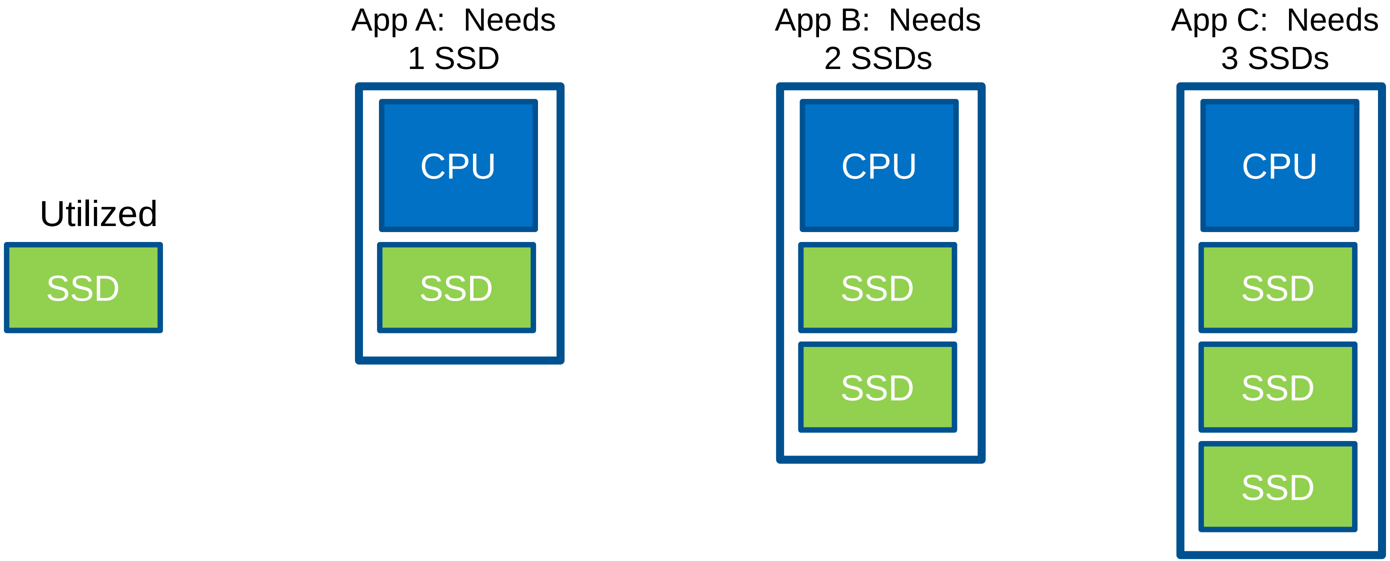
Option A: One Model Serves All Apps



Net utilization: 6 SSDs out of 12 = 50%

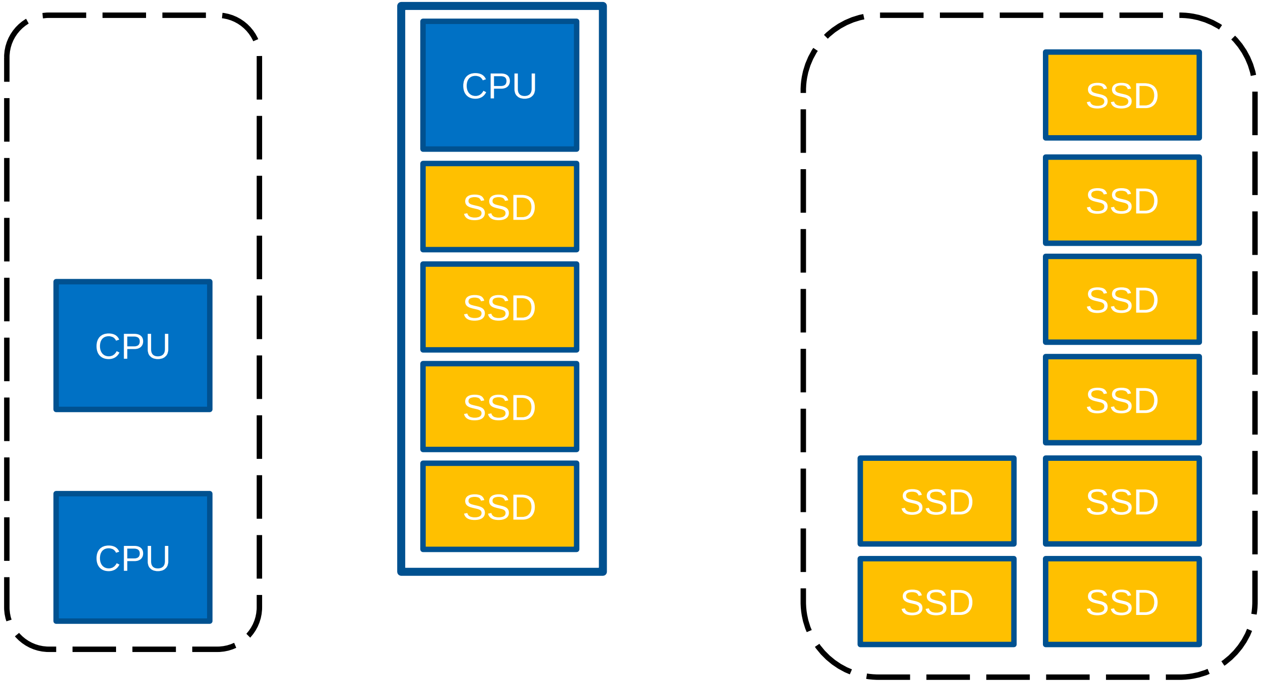
Shared Nothing Model

Option B: Specialized Server Configurations



Dark Flash eliminated, but limits agility and future app deployments

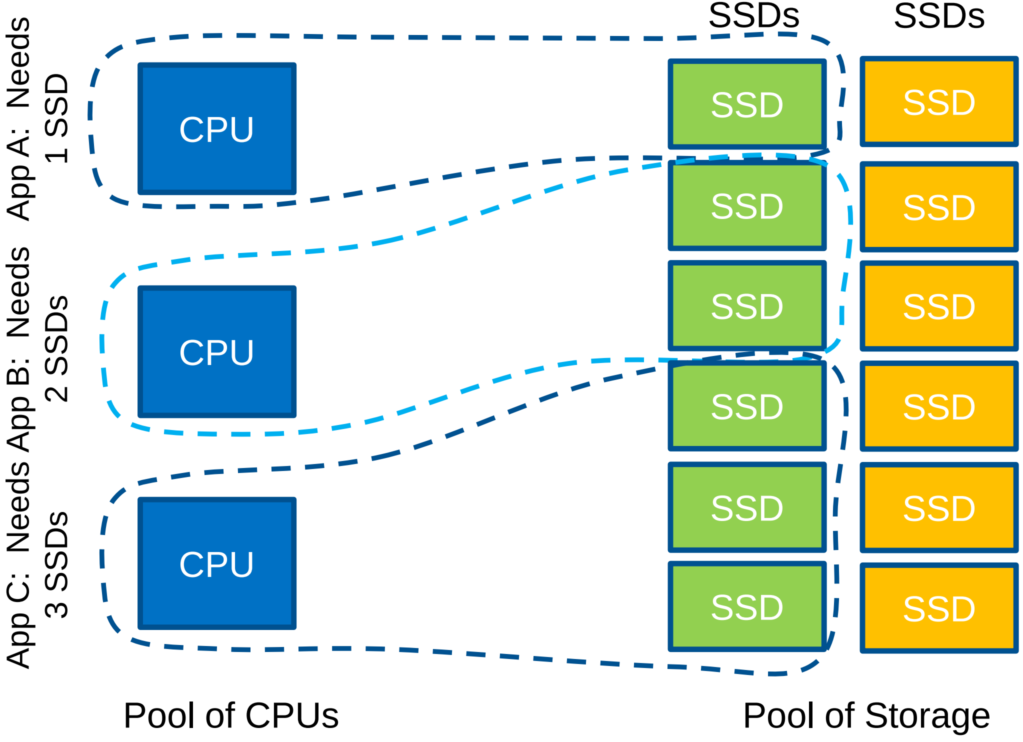
Disaggregated Datacenter



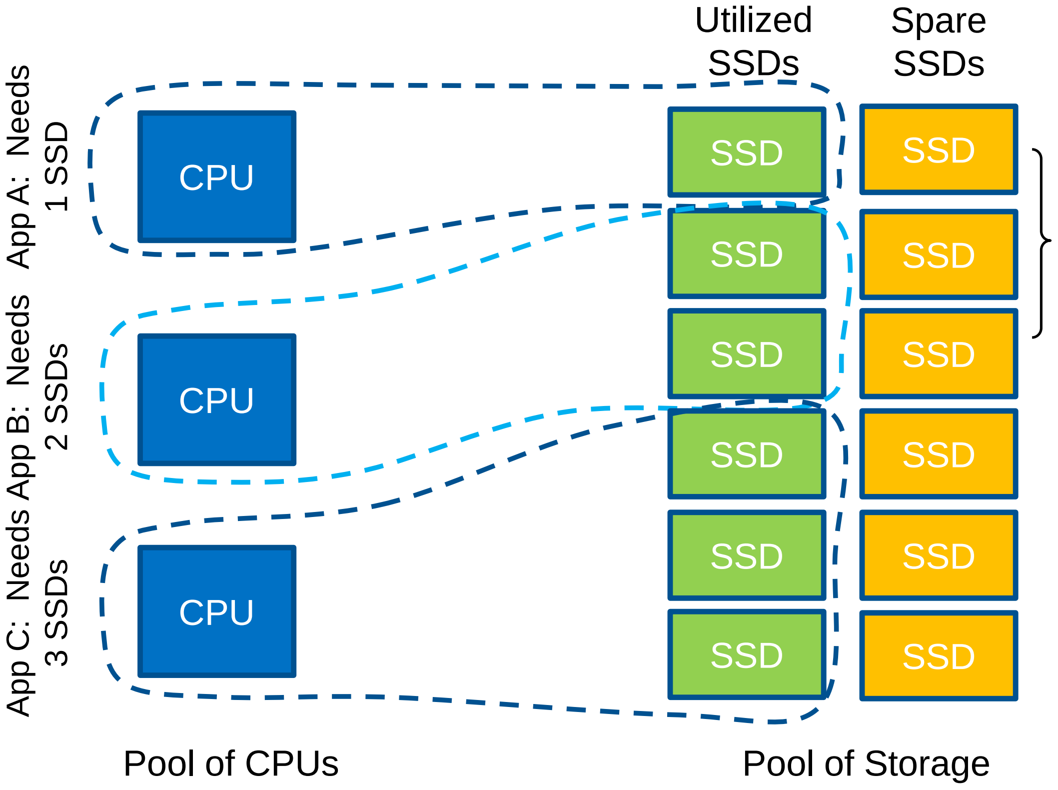
Pool of Compute

Pool of Storage – JBOF/FBOF

The Composable Datacenter



The Composable Datacenter



Spares / Expansion Pool

- Minimize *Dark Flash*!
- Buy them only as needed
- Power them only as needed

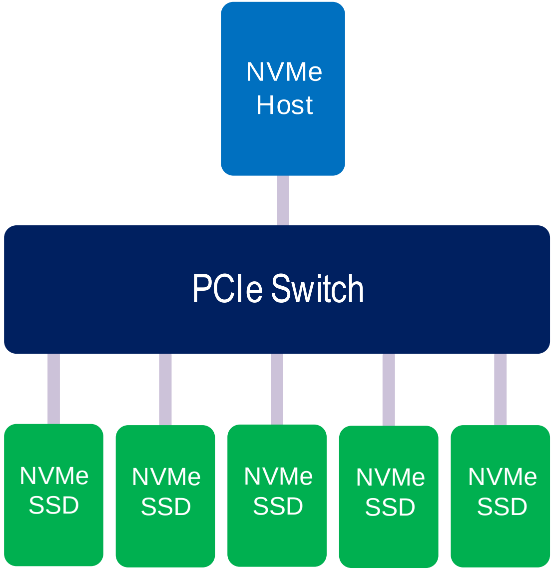
Other benefits

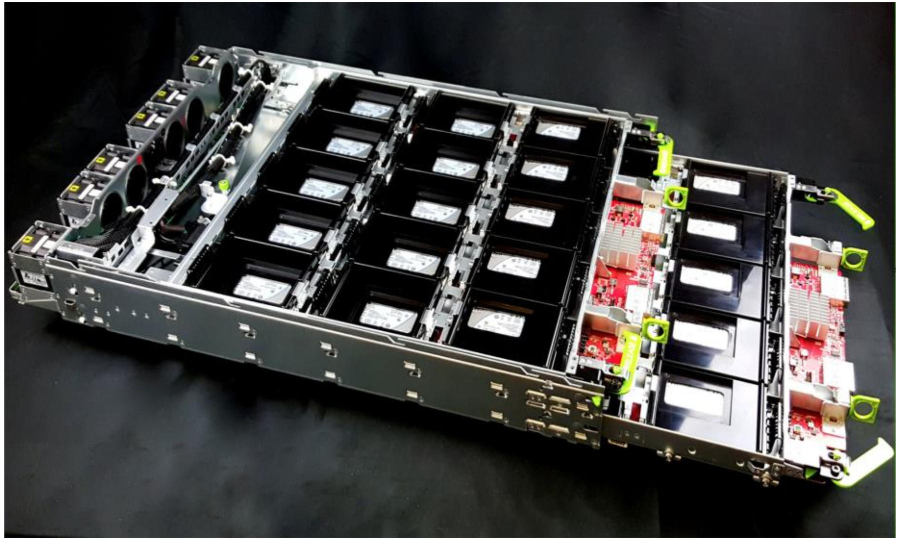
- Dynamically allocate more or less storage
- Return SSDs to Pool as apps are retired
- Upgrade SSDs independently

JBOF Session Agenda

- Market Overview
- Composable Infrastructure
- PCIe (direct-attached) JBOF
- Fabric-attached FBOF
- Management Options
- Remaining Challenges
- Q & A

PCIe® NVMe™ JBOF

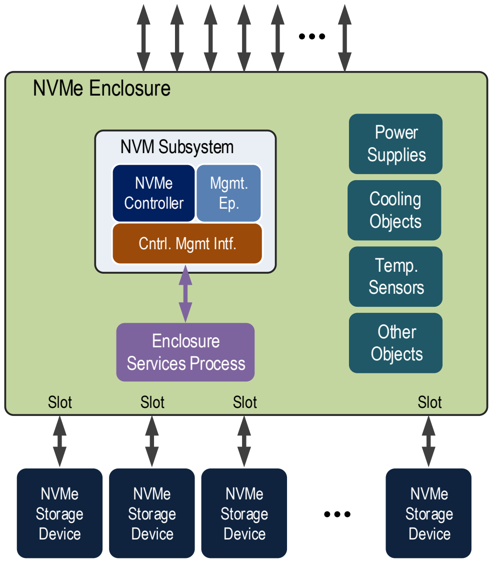




Facebook Lightning PCIe NVMe JBOF

PCIe® JBOF Enclosure Management

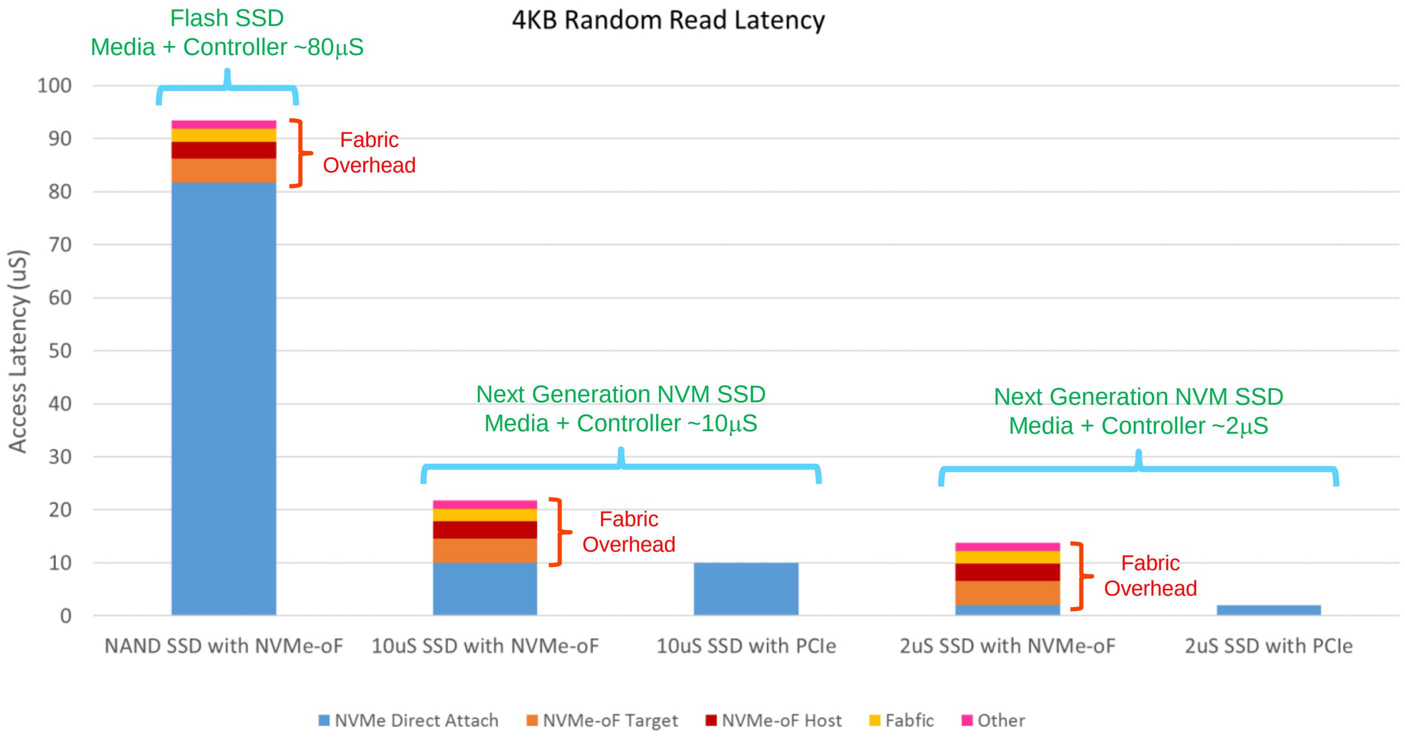
- Native PCIe Enclosure Management (NPEM)
  - Submitted to the PCI-SIG® Protocol Workgroup (PWG) on behalf of the NVMe™ Management Interface (NVMe-MI™) Workgroup
  - Approved by PCI-SIG on August 10th, 2017
  - Transport specific basic enclosure management
- SCSI Enclosure Services (SES) Based Enclosure Management
  - Technical proposal developed in the NVMe-MI workgroup
  - While the NVMe and SCSI architectures differ, the elements of an enclosure and capabilities to manage them are the same
    - Example enclosure elements: power supplies, fans, display or indicators, locks, temperature sensors, current sensors, voltage sensors, and ports
  - Comprehensive enclosure management for NVMe that leverages (SES), a standard developed by T10 for management of enclosures using the SCSI architecture



Flash Memory Summit

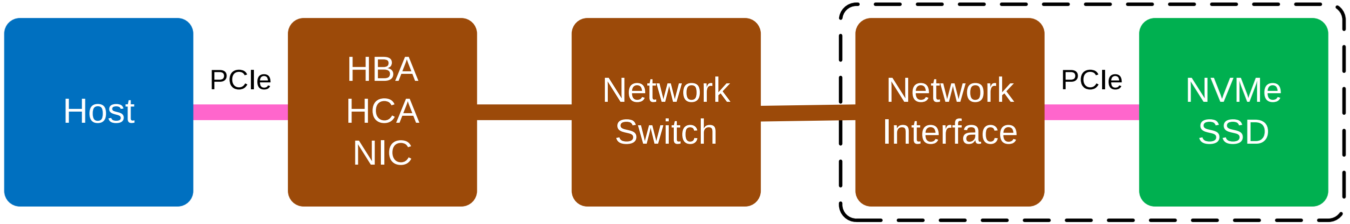
nvm EXPRESS®

The PCIe® Latency Advantage



Latency data from Z. Guz et al., “NVMe-over-Fabrics Performance Characterization and the Path to Low-Overhead Flash Disaggregation” in SYSTOR ‘17

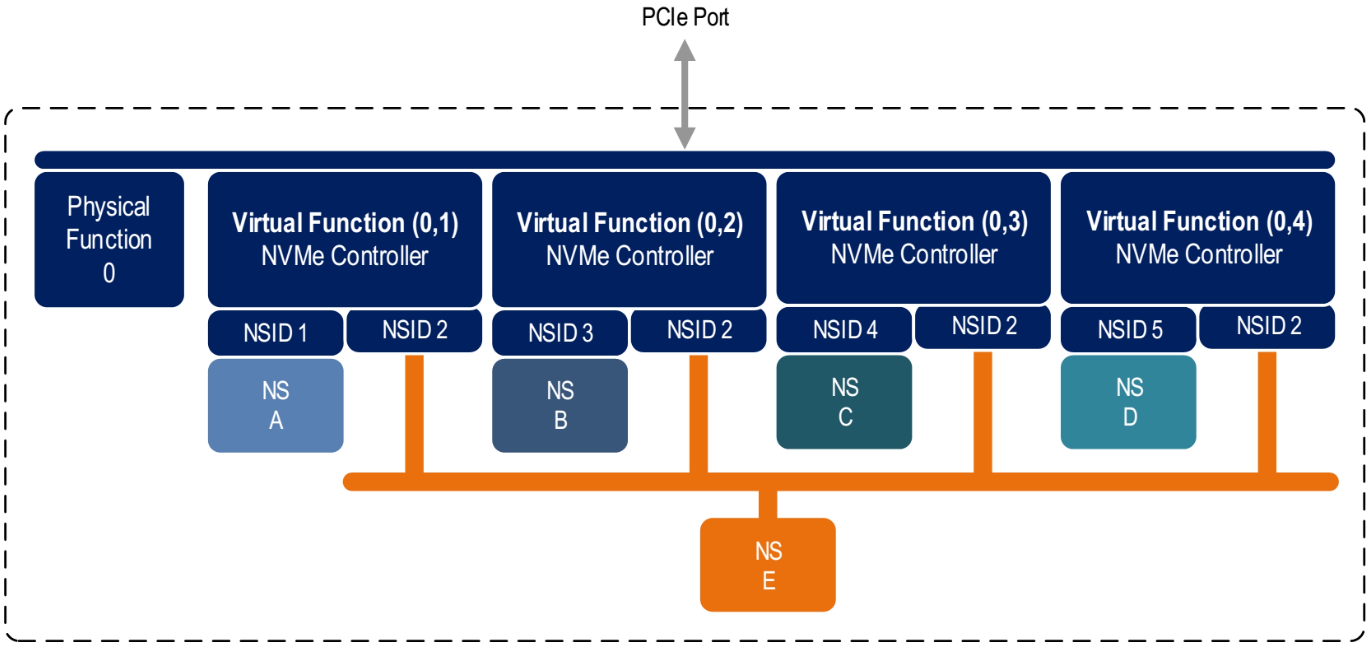
The PCIe® Advantage



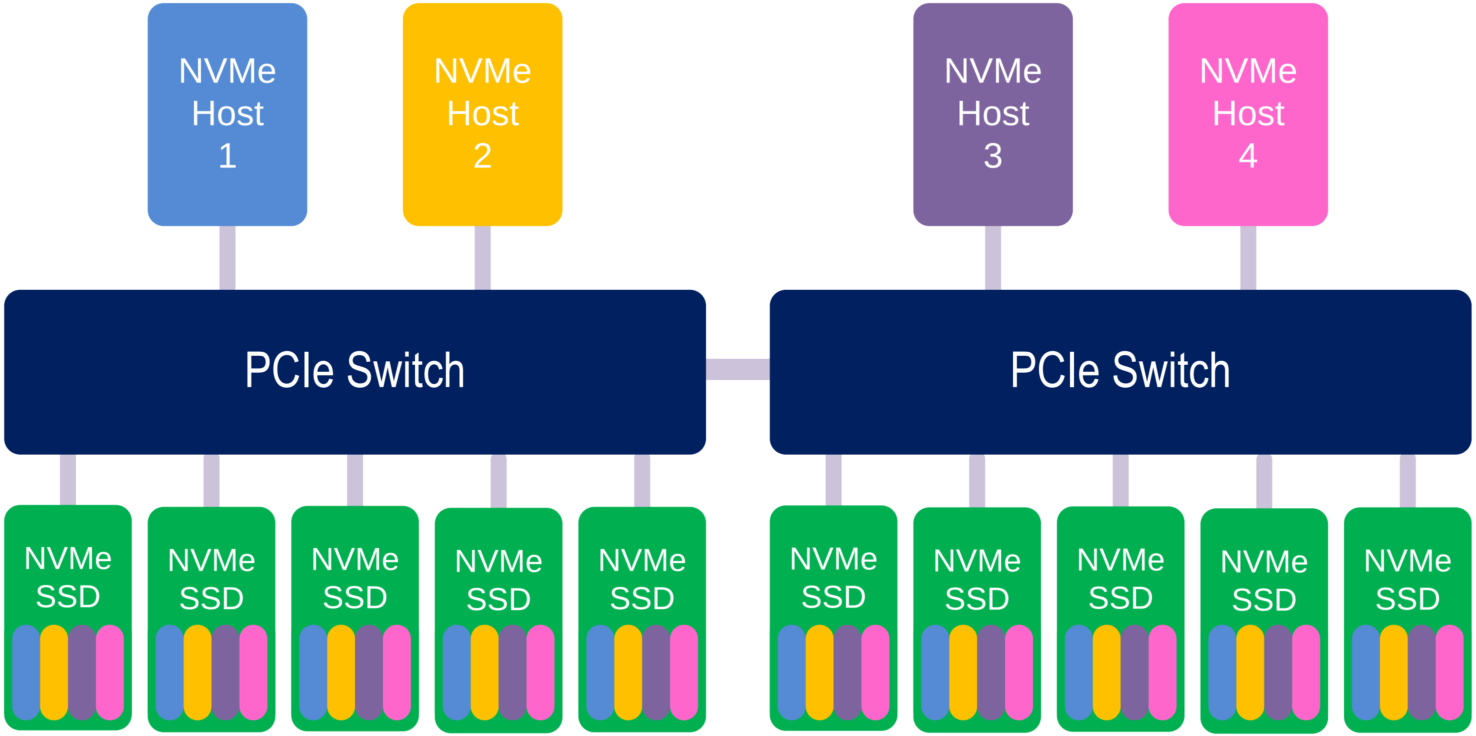
Other Flash Storage Networks

PCIe Fabric

NVMe™ SR-IOV


PCIe Port
Physical Function 0
Virtual Function (0,1) NVMe Controller
Virtual Function (0,2) NVMe Controller
Virtual Function (0,3) NVMe Controller
Virtual Function (0,4) NVMe Controller
NSID 1
NSID 2
NSID 3
NSID 2
NSID 4
NSID 2
NSID 5
NSID 2
NS A
NS B
NS C
NS D
NS E

Multi-Host I/O Sharing



Storage is Not Just About CPU I/O Anymore

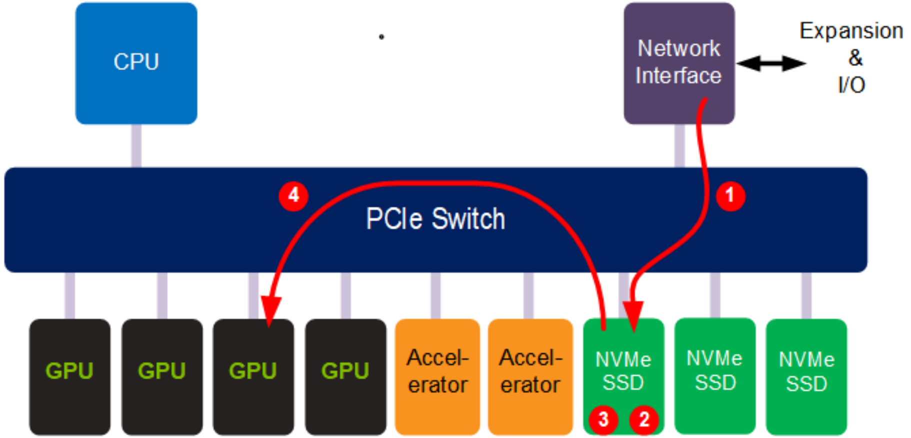
- NVMe™ together with a PCIe® fabric allow direct network to storage and accelerator to storage communications

Example:

1. Data transferred from network to NVMe CMB
2. NVMe block write operation imitated from CMB to NVM

… sometime later …

3. NVMe block read operation initiated from NVM to CMB
4. GPU/Accelerator transfers data from NVMe CMB for processing



Flash Memory Summit

nvm EXPRESS®
Architected for Performance

FBOF Architecture

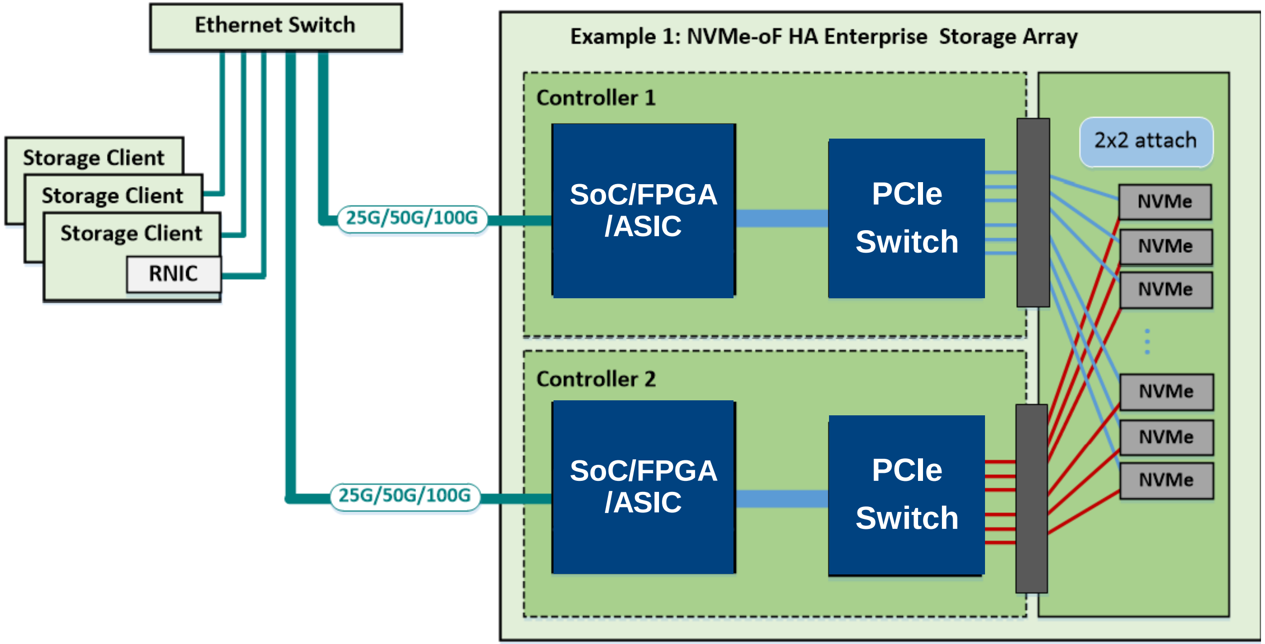
Fazil Osman, Broadcom

NVMe-oF™ Market

SAS Replacement	Composible
High performance Low latency Better scalability than PCIe® Solution for traditional Enterprise iSCSI, cluster architectures etc.	TCP IO Determinism Data Integrity Application Offload Cloud Scale Out

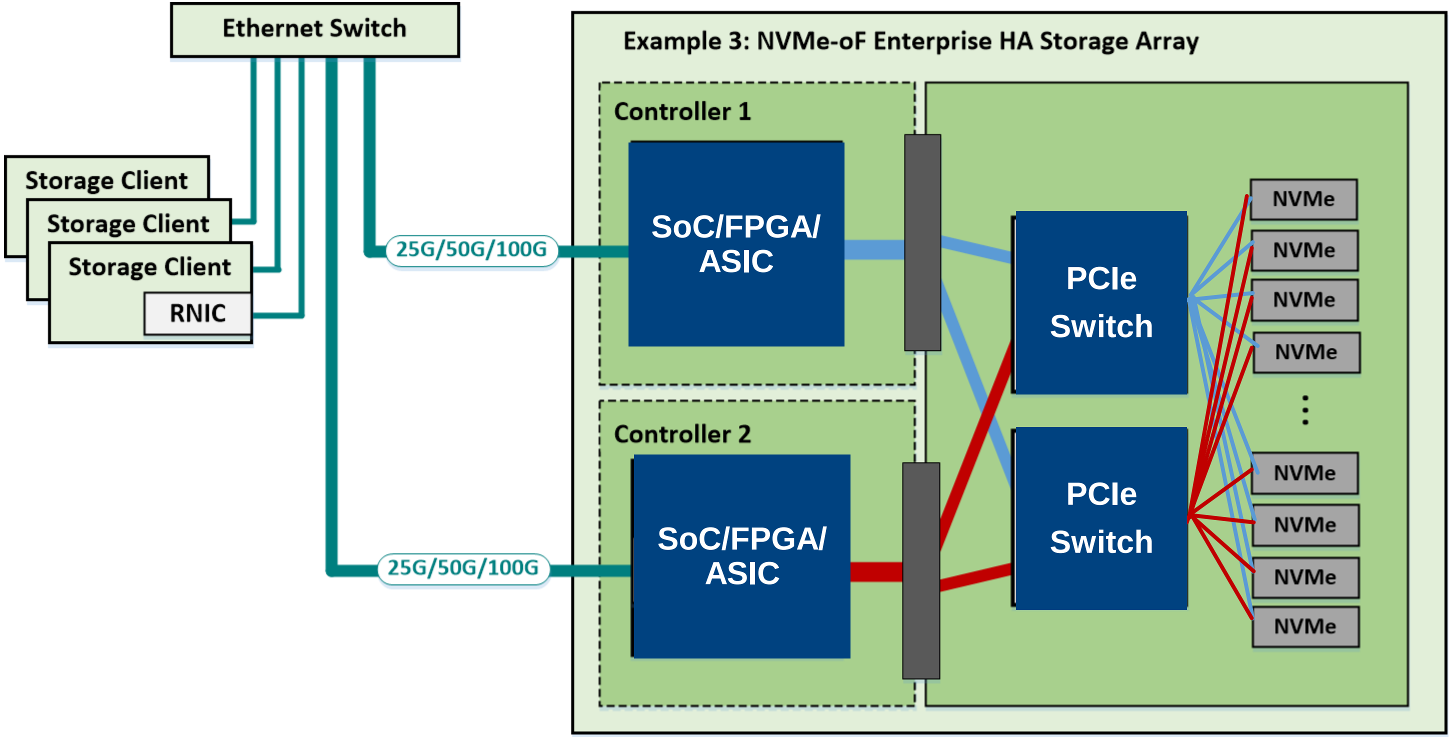
Form Factor		
Today	**24 x U.2**	**30/60 x M.2**
Future	**Ruler (16/32 x 1U)** **EDSFF, NF1**	**Modular w/Ethernet** **EDSFF Derivative**

FBOF architecture examples



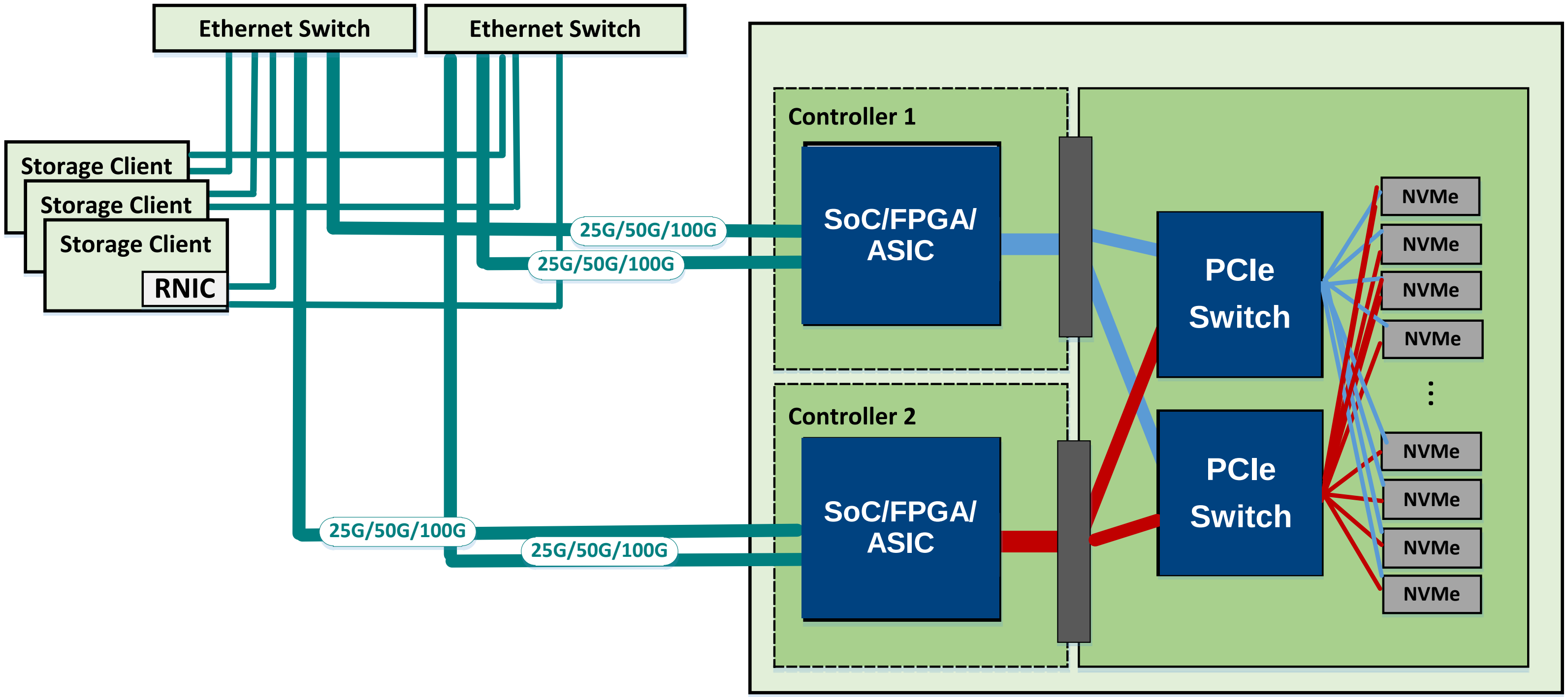
High Availability option 1

HA FBOF architecture



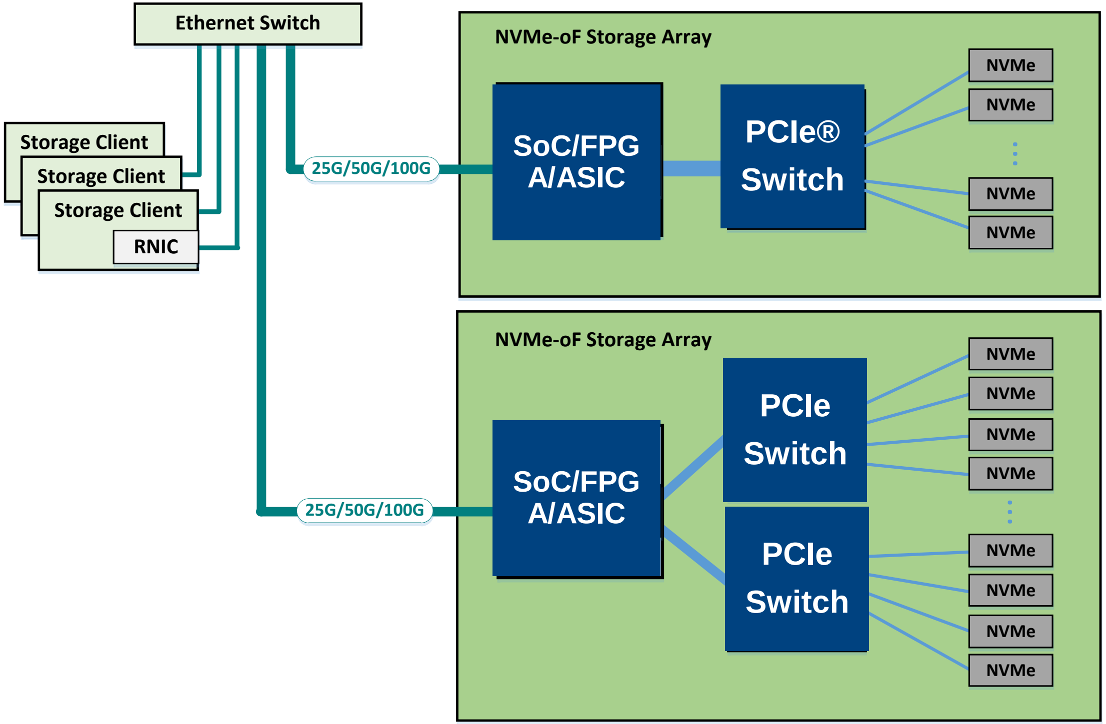
High Availability option 2

HA FBOF architecture with redundant switches



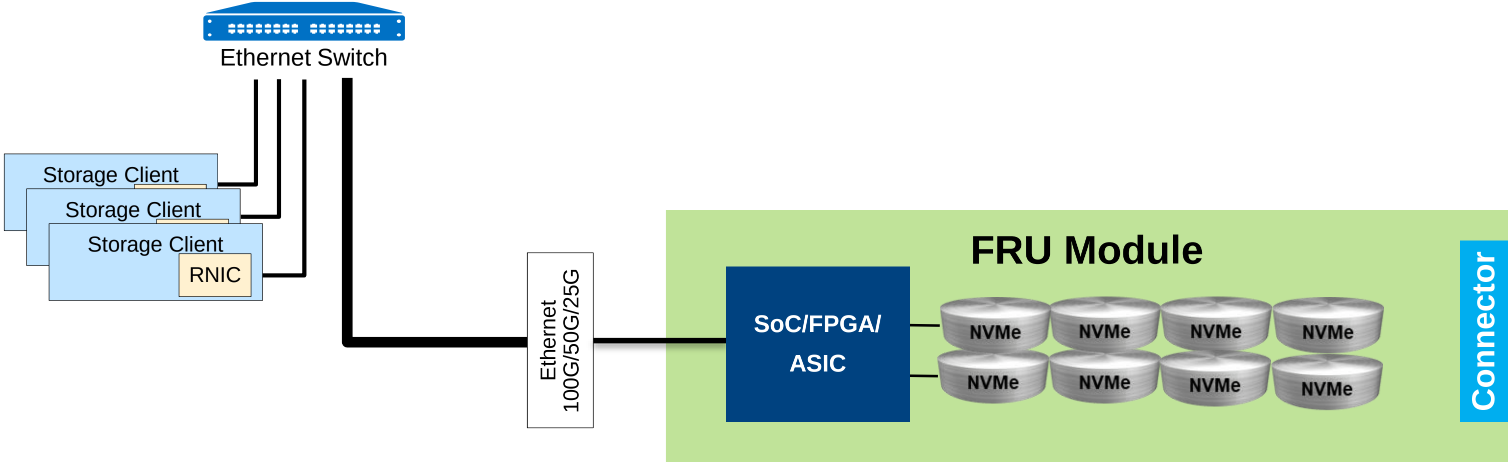
High Availability option 3

FBOF high fanout architecture



High Fanout

Scale Out Cloud Architecture



1U ruler based designs on PCIe® attach being introduced into the market

– i.e. White River Glacier etc., various ODM offerings

Designs provide high density NVMe™ but lack scalability

Goal is to extend concept for cloud scale using NVMeoF™

Gain scalability of fabrics attached

Simplify design by removing PCIe switch

Flash Memory Summit

NVM Express®
Architected for Performance

FBOFs in the Cloud

Sujoy Sen, Intel

Making FBOFs Successful in the Cloud

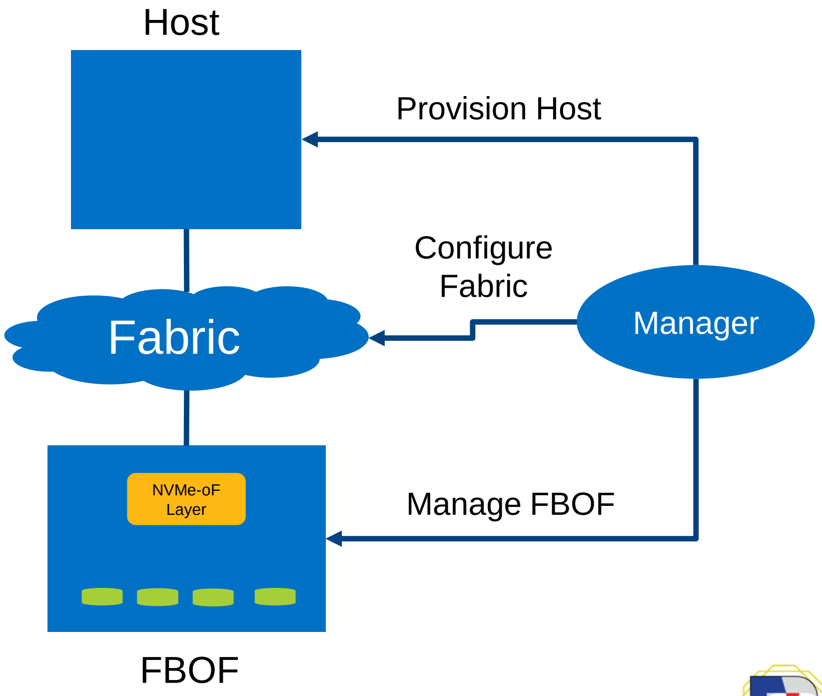
FBOFs in the cloud enable the composable and disaggregated use case

Success will require the following

- Network QoS (especially RDMA@scale)
- Easy to deploy and manage@scale
- Enable Scale-out Distributed Storage architectures

Ease of Use

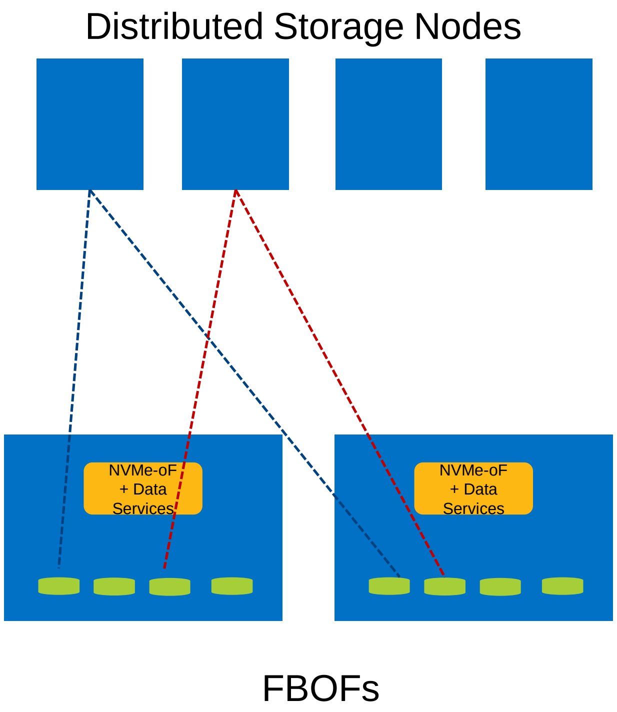
- E2E management
  - Not just FBOFs but the hosts and the network in-between
- Cloud OS Enablement
  - Develop drivers/plug-ins for NVMe-oF™
- Bare Metal
  - Server platform and OS native support for NVMe-oF provisioning



Drive standards-based management eco-system

Scale-Out Distributed Storage

- Blast Radius and Failure Domains
  - Soft vs hard error handling
  - Single Point-of-Failure avoidance
- Partitioning of Data Services between storage node and FBOF, e.g.
  - Data Layout and Media Management
  - Replication/HA
  - Data Compression and Security
- Distributed storage-aware NVMe-oF™
  - Cluster-aware protocol enhancements



Key Takeaways

- JBOF / FBOF represents a key building block for NVMe™ based datacenters
- Two options:
  - PCIe® Direct Connect JBOFs
    - Lowest Latency
    - Limited Scale / Distance
  - Fabric Attached FBOFs
    - Scale at the levels of FC or Ethernet
    - Additional latency, networking / fabric bandwidth
- Manageability represents new opportunities and challenges

Contact Information

For more information please contact the following:

Fazil Osman fazil.osman@broadcom.com

Nishant Lodha Nishant.Lodha@cavium.com

Sujoy Sen sujoy.sen@intel.com

Bryan Cowger bryan.cowger@kazan-networks.com

Peter Onufryk Peter.Onufryk@microchip.com

nvm EXPRESS®

Architected for Performance