Are Ethernet Attached SSDs Happening?

NVMF-302B-1

Organizer/Chair: Rob Davis, Mellanox

Presenters:

Ilker Cebeli, Samsung

John Kloeppner, NetApp

Balaji Venkateshwaran, Toshiba

Khurram Milak, Netronome

Woo Suk Chung, SK Hynix

Session Agenda

- <u>Ilker, Samsung – 15 minutes</u>
- **John - NetApp, Balaji -Toshiba, Khurram - Marvell – 40 minutes**
- **Woo, SK-Hynix – 15 minutes**
- **Q&A – 10 minutes**

Are Ethernet Attached SSDs Happening?

Disaggregated NVMe-oF Storage

Ilker Cebeli
Sr. Director of Planning
Samsung

Disclaimer

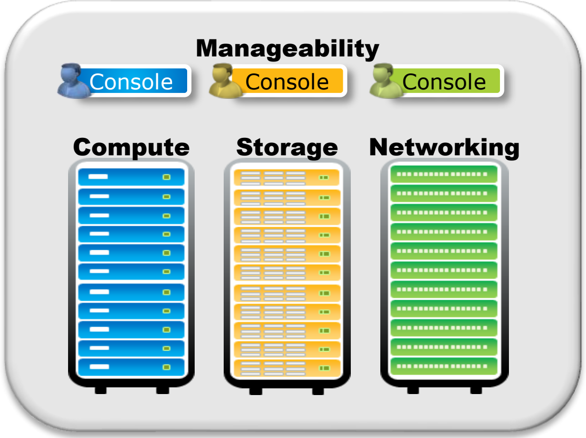
Flash Memory Summit

This presentation and/or accompanying oral statements by Samsung representatives collectively, the "Presentation") is intended to provide information concerning the SSD and memory industry and Samsung Electronics Co., Ltd. and certain affiliates (collectively, "Samsung"). While Samsung strives to provide information that is accurate and up-to-date, this Presentation may nonetheless contain inaccuracies or omissions. As a consequence, Samsung does not in any way guarantee the accuracy or completeness of the information provided in this Presentation.

This Presentation may include forward-looking statements, including, but not limited to, statements about any matter that is not a historical fact; statements regarding Samsung's intentions, beliefs or current expectations concerning, among other things, market prospects, technological developments, growth, strategies, and the industry in which Samsung operates; and statements regarding products or features that are still in development. By their nature, forward-looking statements involve risks and uncertainties, because they relate to events and depend on circumstances that may or may not occur in the future. Samsung cautions you that forward looking statements are not guarantees of future performance and that the actual developments of Samsung, the market, or industry in which Samsung operates may differ materially from those made or suggested by the forward-looking statements in this Presentation. In addition, even if such forward-looking statements are shown to be accurate, those developments may not be indicative of developments in future periods.

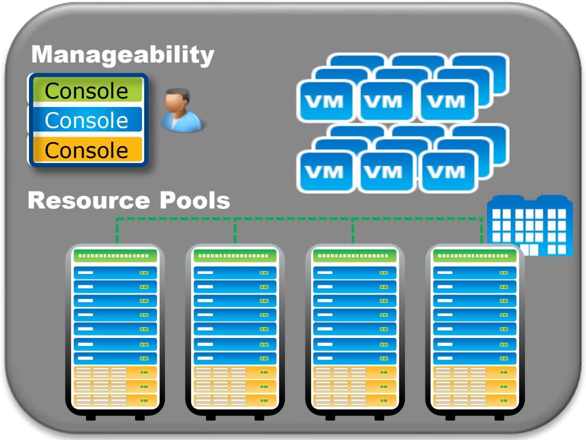
Data Center Evolution

Traditional Data Center



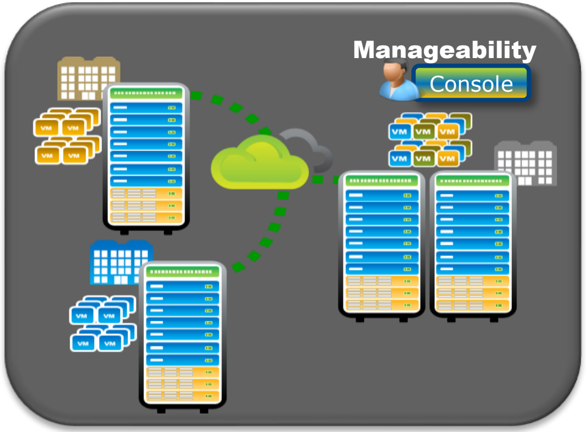
Stand Alone Component
Suited for Enterprise Applications
1GbE Networking

Hyper-converged Virtualized



Converged Management
Virtualized Computing/ Networking
10GbE Networking

Software-Defined Composable



Rack Scale Software Defined
Disaggregated Compute and Storage
Composable
25-100GbE Networking

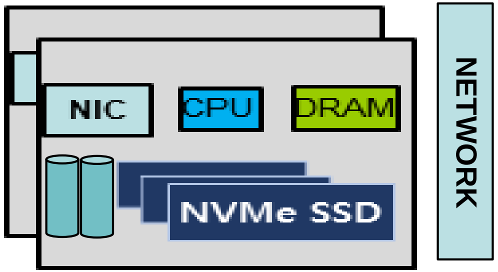
Evolution

Flash Memory Summit

Why Disaggregation?

Converged

Compute + Storage

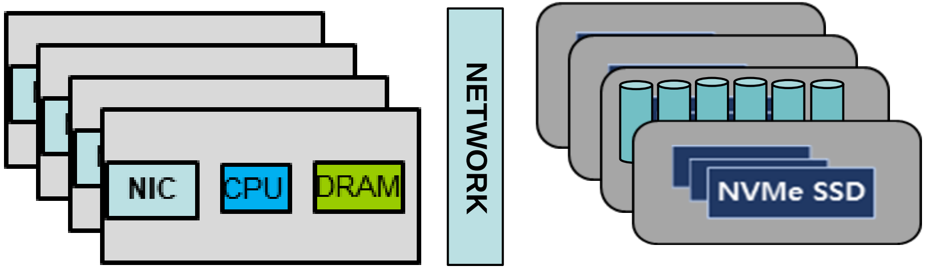


- Pros:
  - ✓ Scale Compute and Storage linearly
  - ✓ Managed resources and storage services
- Cons
  - ➢ Resources under-utilization
  - ➢ Storage and Compute on the same network

Disaggregated Compute & Storage

Compute

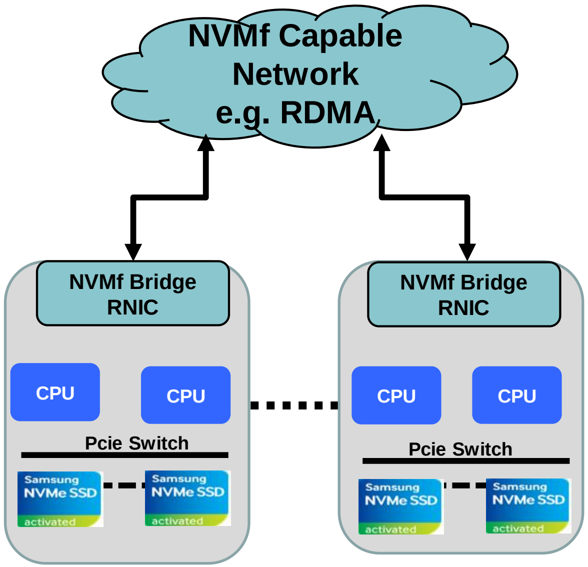
Storage



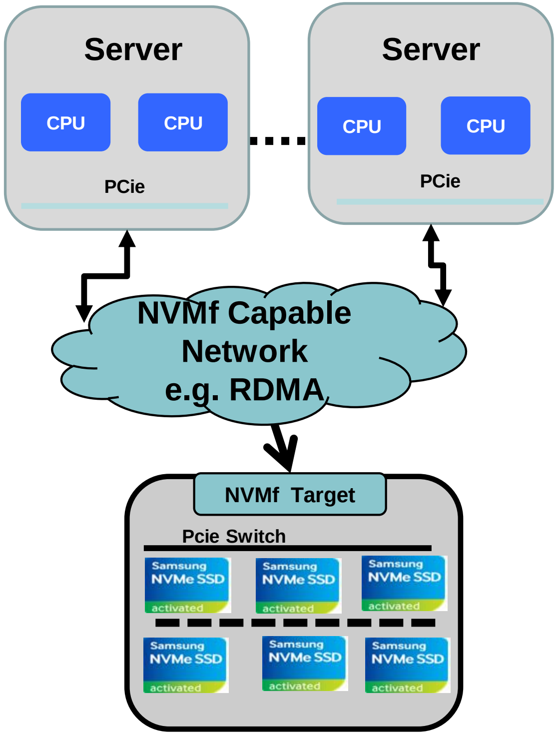
- Pros:
  - ✓ Compute and Storage scale independently
  - ✓ Shared resources
  - ✓ Improved utilization
  - ✓ Grow as you go model based on workload demand
  - ✓ Centralized storage services
- Cons
  - ➢ Requires efficient storage protocols and latency
  - ➢ Low latency and high bandwidth networking

Some of the Use Cases for NVMe Over Fabrics

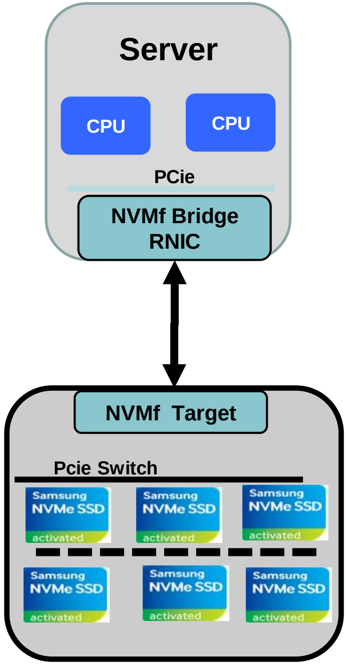
Flash Memory Summit



Hyper-Converged
Most Common

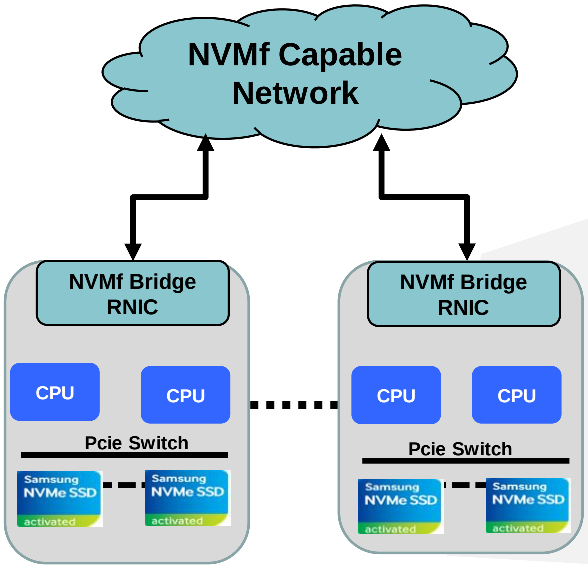


Disaggregated
JBOF Storage



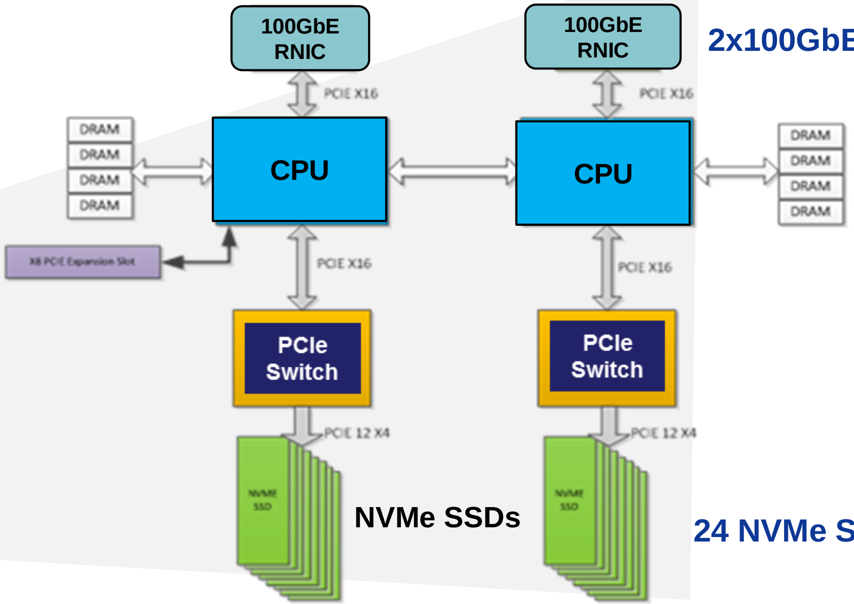
Direct Attached JBOF
SAS DAS Replacement

NVMe-oF JBOF

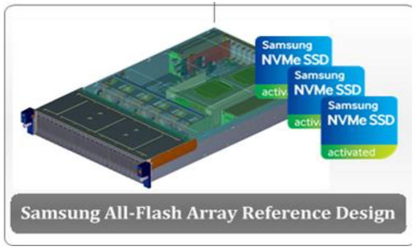


Hyper-Converged
Most Common

2015 NVMe-oF JBOF



Typical NVMe-oF JBOF

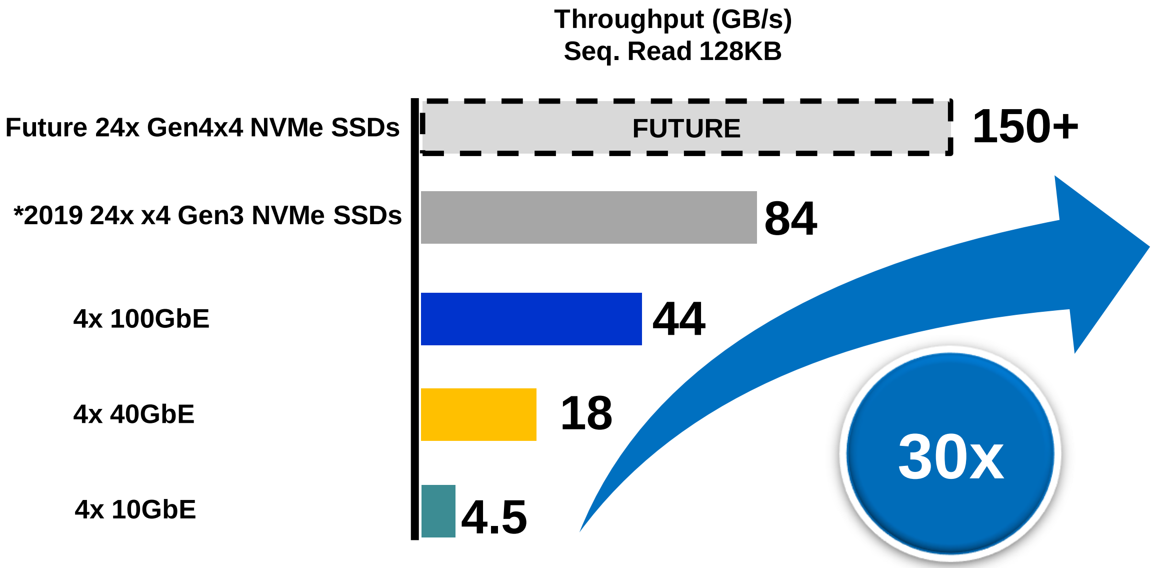


2x100GbE ➔ ~24GB/s

24 NVMe SSDs ➔ ~24GB/s

2015 Platform Balanced Bandwidth between IO and NVMe

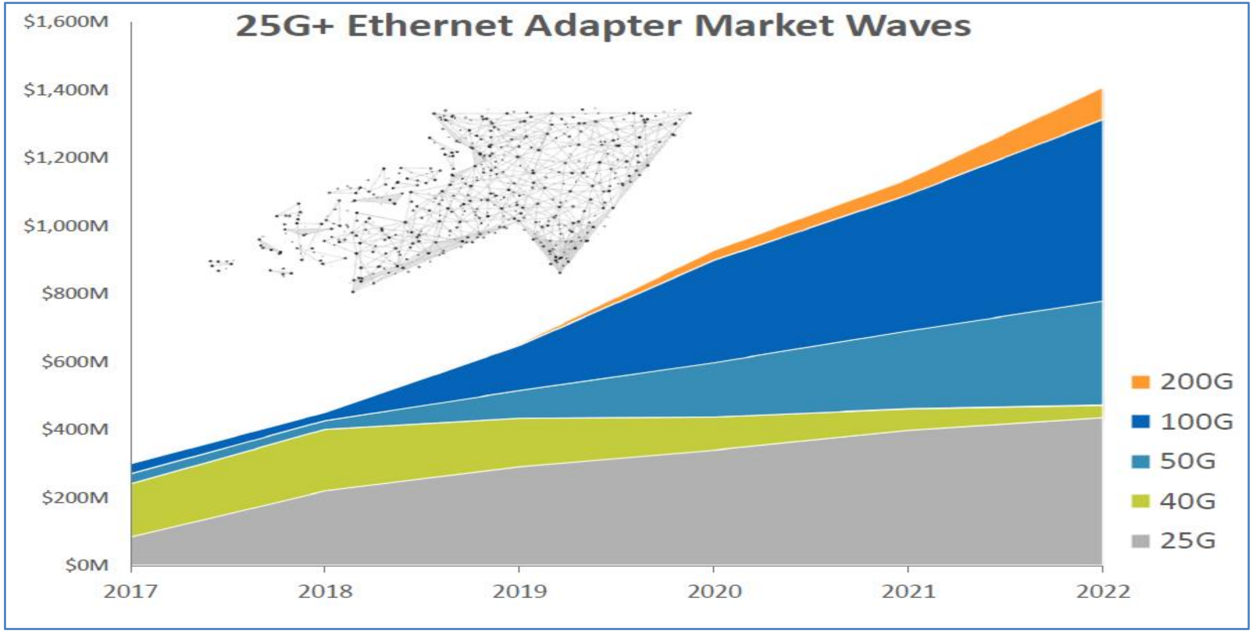
Future NVMe Bandwidth



* 24x Samsung PM1725b NVMe SSDs (3.5GB/s throughput each)

Network links could throttle the storage throughput performance

Evolution of Networking Speeds and 25Gb/s and Above

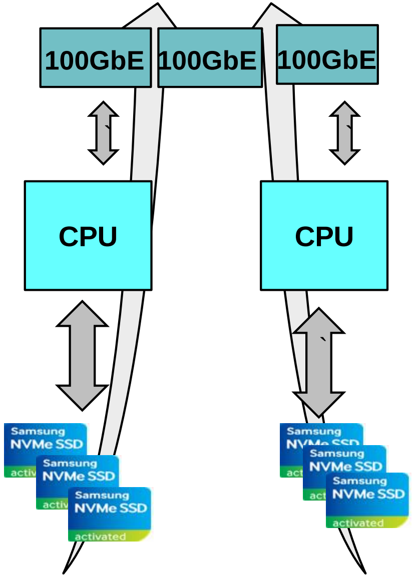


Source: Crehan Long-range Forecast - Ethernet Adapter forecast, January 2019 via Mellanox Q2'2019

IO Bottleneck

2019

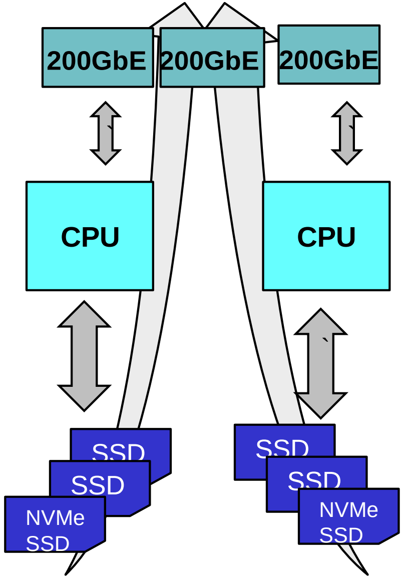
3 X 100GbE ➔ ~36 GB/s



24 NVMe SSDs ➔ ~48 GB/s

Future

3 X 200GbE ➔ ~75 GB/s



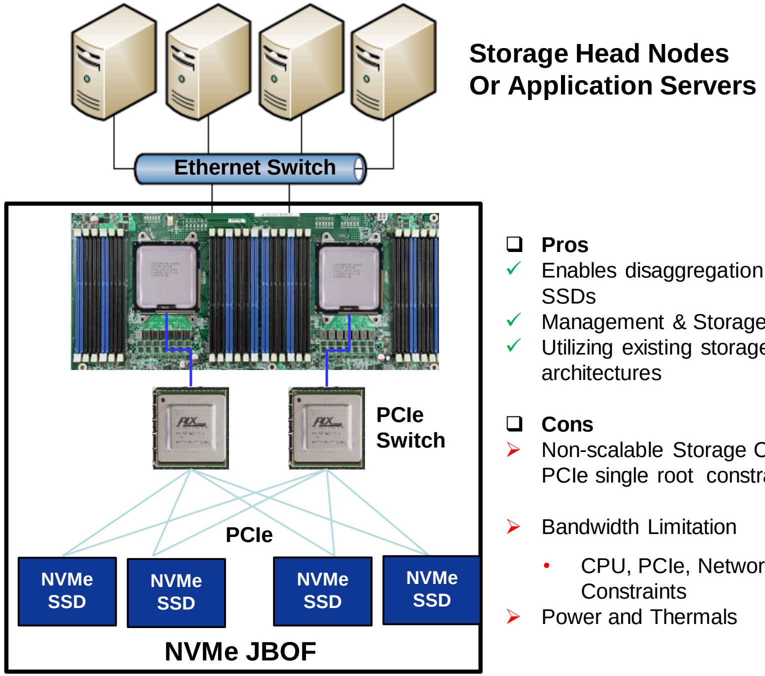
24 NVMe SSDs ➔ ~150+ GB/s

CPU and IO bottleneck for storage throughput performance

NVMe-oF SSD based EBOF

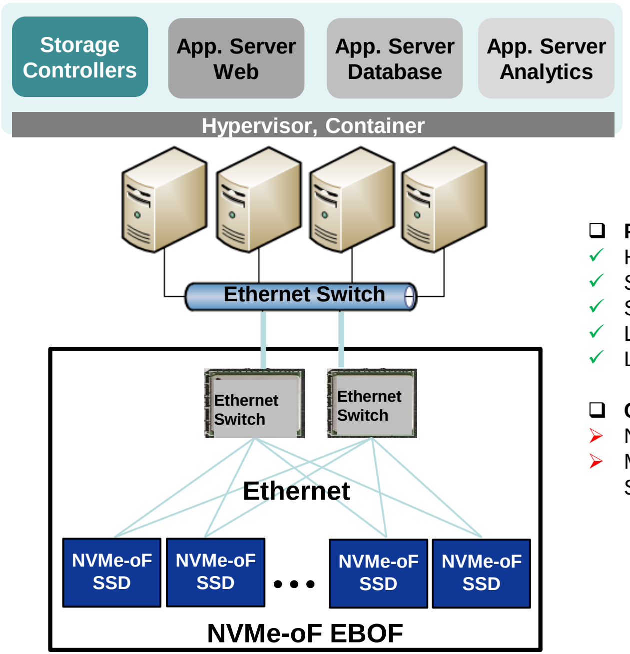
Flash Memory Summit

Conventional NVMe JBOF



- **Pros**
  - ✓ Enables disaggregation of NVMe SSDs
  - ✓ Management & Storage Services
  - ✓ Utilizing existing storage & server architectures
- **Cons**
  - ➢ Non-scalable Storage Controller - PCIe single root constraint
  - ➢ Bandwidth Limitation
    - CPU, PCIe, Networking Constraints
  - ➢ Power and Thermals

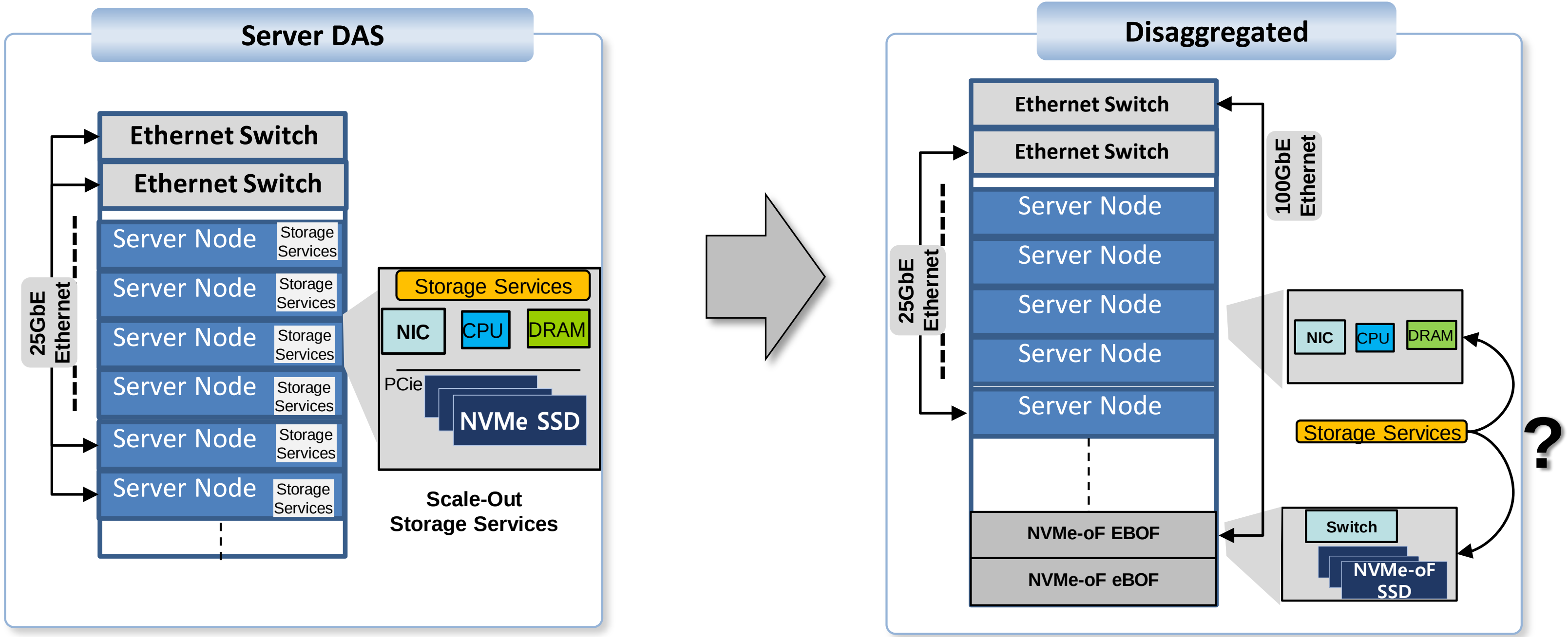
NVMe-oF EBOF



- **Pros**
  - ✓ High Bandwidth
  - ✓ Scaled Linearly (Ethernet)
  - ✓ Sharable via NVMe-oF
  - ✓ Less power
  - ✓ Lower latency
- **Cons**
  - ➢ New platform architecture
  - ➢ Management of Storage Services & Network Devices

NVMe-oF EBOF can address bandwidth, scalability, and flexibility

Example Datacenter Storage Disaggregation



Where Storage Services and Network Devices managed

ilker.cebeli@samsung.com

Thank You

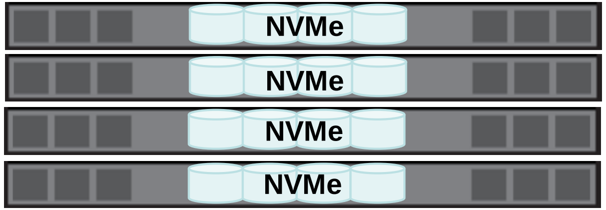
Session Agenda

- **Ilker, Samsung – 15 minutes**
- ***John - NetApp*, Balaji -Toshiba, Khurram - Marvell – 40/3 minutes**
- **Woo, SK-Hynix – 15 minutes**
- **Q&A – 10 minutes**

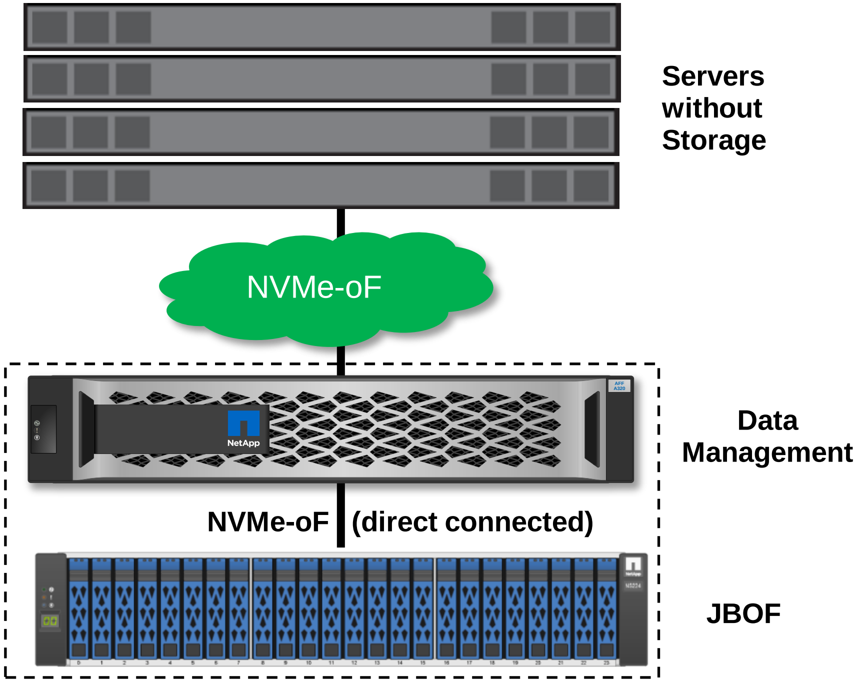
NVMe -> NVMe over Fabrics

Flash Memory Summit

Servers with embedded NVMe Storage



- Local high performance / low latency access
- Isolated Storage
- Under-utilized SSD Performance and Capacity

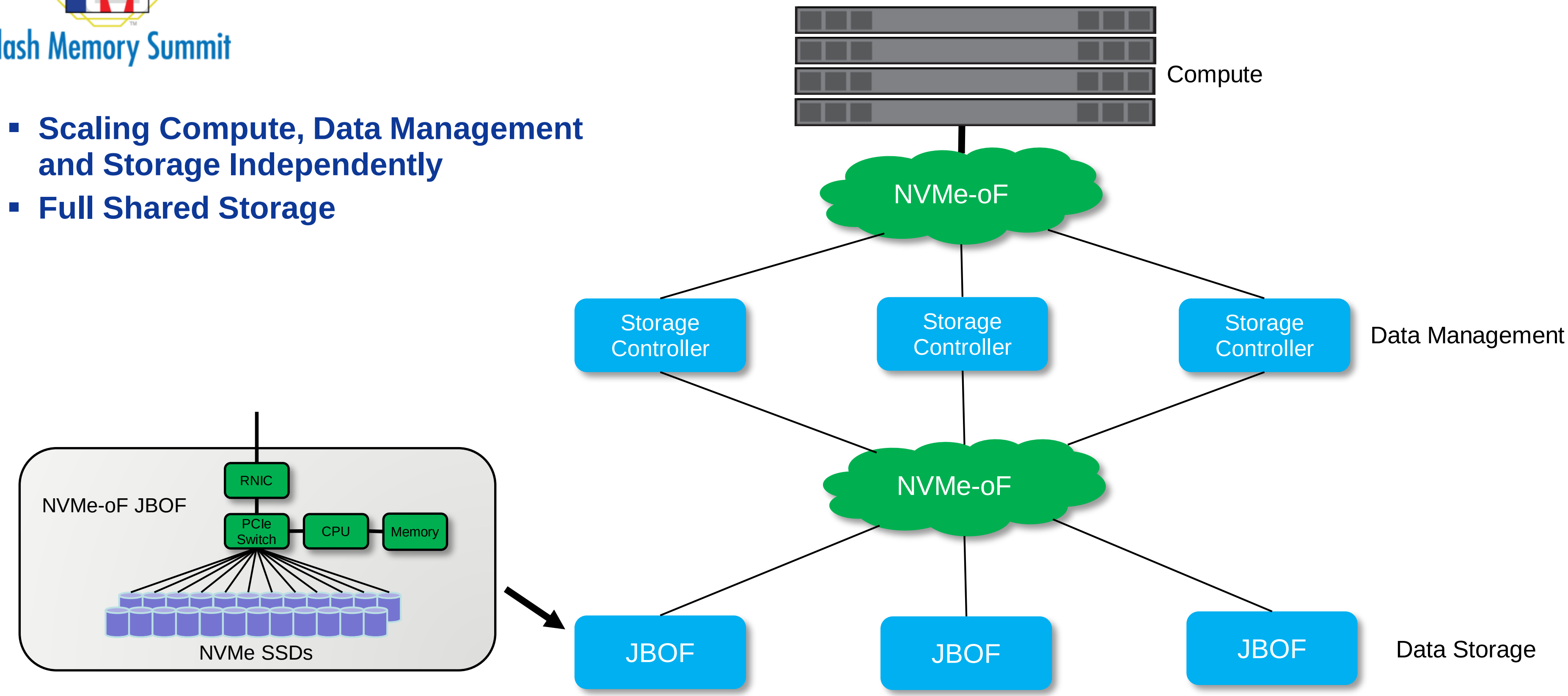


- Shared Storage, better utilization of storage
- Similar NVMe Performance

NetApp

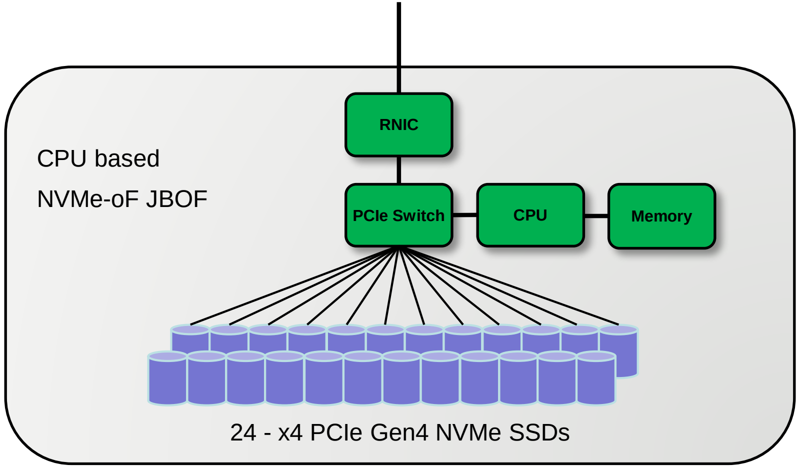
Disaggregated Compute/ Data Management / Storage

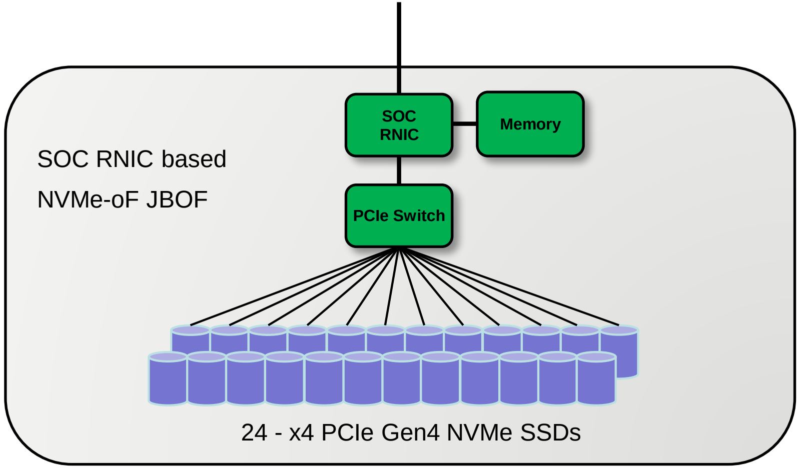
- **Scaling Compute, Data Management and Storage Independently**
- **Full Shared Storage**



NVMe-oF JBOF Limitations

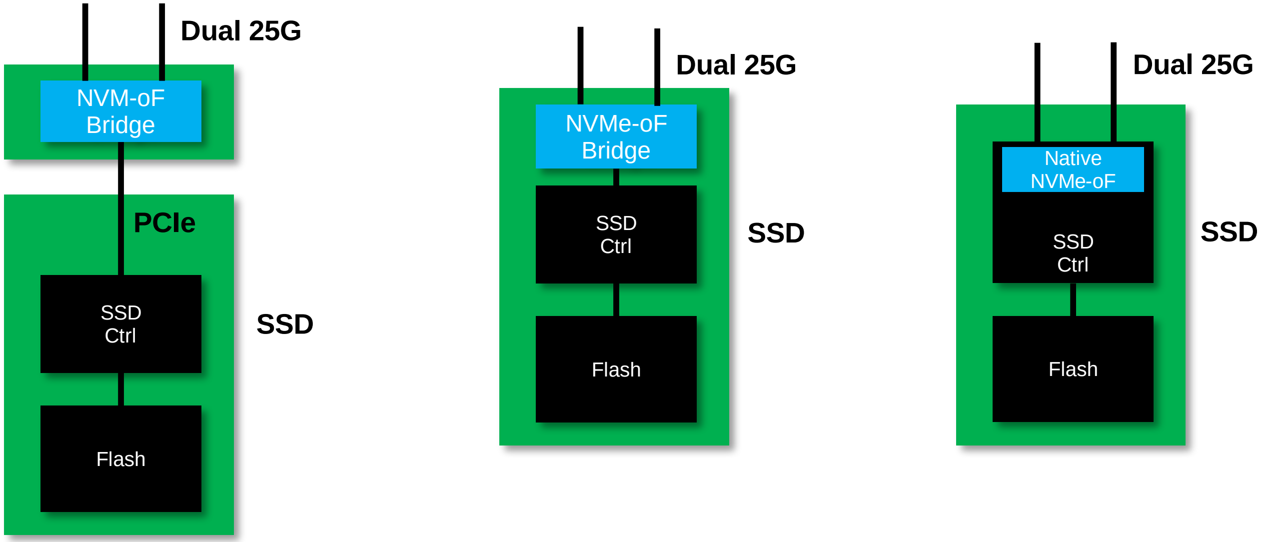
- **Performance**
  - **Throughput - PCIe Gen3 -> PCIe Gen4 -> PCIe Gen5, SCM, limit by existing infrastructure**
  - **Latency - Store and Forward architecture**
- **Cost – CPU, SOC/RNICs, Switches, Mem don’t scale well to match increasing SSD performance**





Native Ethernet / NVMe-oF SSDs

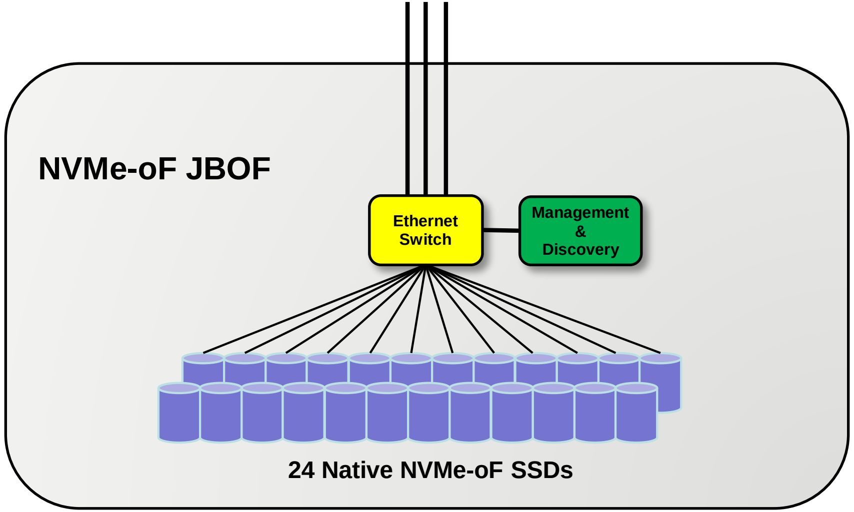
- Optimize NVMe-oF performance at SSD
- Options for NVMe-oF SSDs

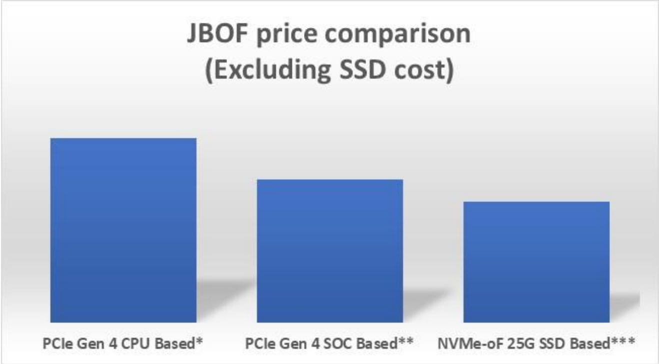


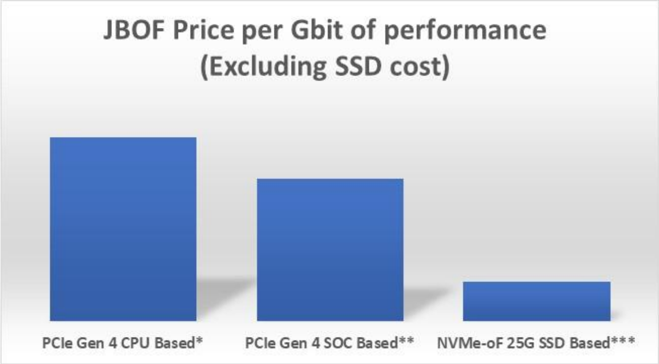
NetApp

Solution with Native NVMe-oF SSDs

- **Lower Latency, Higher Throughput**
- **Lower Cost and overall TCO**







*** Supports one 2x200G RNIC connected with x16 PCIe Gen4**

**** Supports one 2x200G SOC RNIC connected with x16 PCIe Gen4**

***** Supports three 200G Host connected Ethernet ports**

Additional Benefits

- Additional Benefits
  - Performance/cost scales with SSDs
  - Lower Power, reduced TCO
  - Including Ethernet switching within JBOF … potential to reduce networking cost, footprint, cabling

Other Activities

- Industry Standardization / Enablement
  - Standardization – Work underway in SNIA to define Form Factor, Pinout, Management – Toshiba will cover
  - Enablement – Fabrico Interposer – Marvell will cover

Thanks!

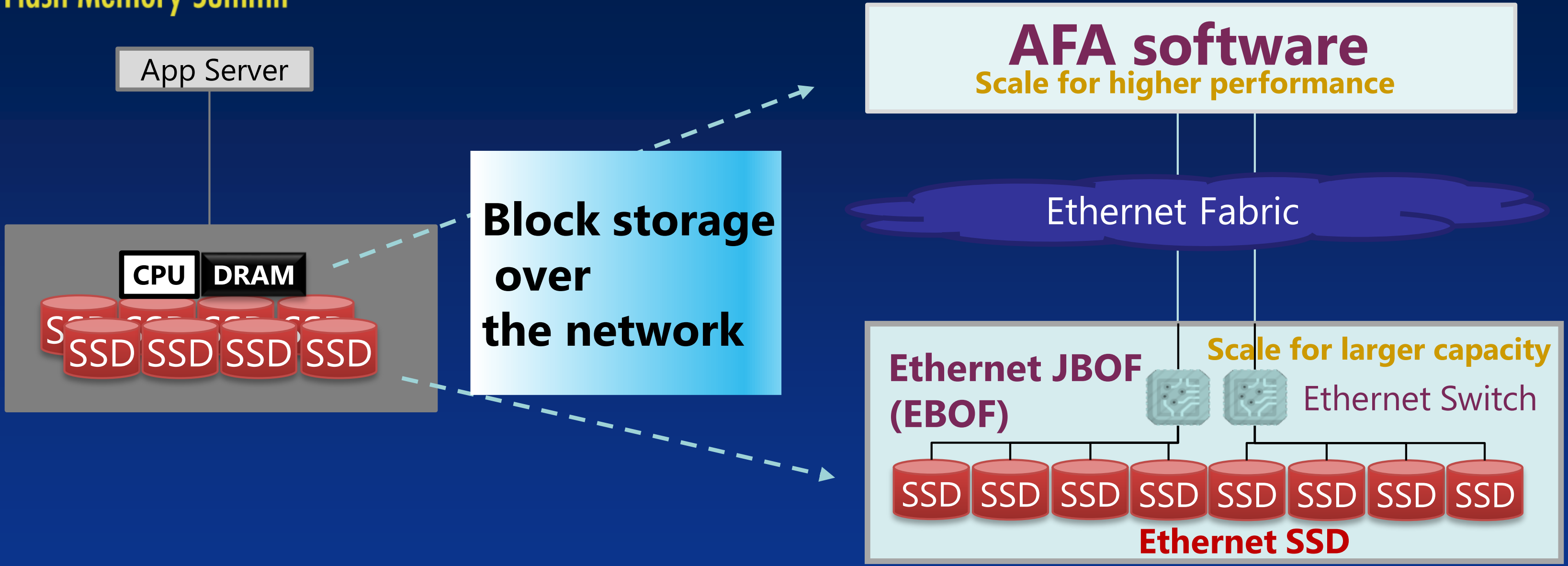
Session Agenda

- **Ilker, Samsung – 15 minutes**
- **John - NetApp, *Balaji -Toshiba*, Khurram - Marvell – 40/3 minutes**
- **Woo, SK-Hynix – 15 minutes**
- **Q&A – 10 minutes**

Enabling Native NVMe-oF™ SSDs (Ethernet SSDs)

August 2019

Ethernet SSD-based Storage Platforms



Advantages

- Independent scaling between performance (controller node) and capacity (JBOF) for optimal HW deployment in large scale systems
- Manage NVMe™ -based pools for separate storage/caching tiers

Flash Memory Summit

Enabling NVMe-oF™ Functionality in SSDs

- Connector
  - SFF 8639 connector predominant for NVMe™ -based systems
  - SFF-TA-1002 (EDSFF) specification a future-proof option
  - Standardizing Ethernet pinout in the connector a must for industry adoption

- Management Framework
  - NVMe™ devices attached to a system get enumerated using OS resources
  - Ethernet-attached device enumeration needs equivalent network functionality
  - Potential candidates for easier manageability:
    - NVMe-MI – from a BMC (not network)
    - RedFish – works for scalability in a Datacenter Network
    - RSD – uses RedFish

Considerations in Connector Standardization

- Ethernet-based pinout should ensure:
  - SSDs of different types can be interchanged without electrical damage
    - First look in the VPD via SMBus, then apply power and signals
  - Forward-compatible
    - Connector of choice should support 25G → 50G → 100G transitions
    - Multi-lane for dual-port connectivity
  - Backwards-compatible
    - Ethernet pinout-based SSD should share midplane with SAS/SATA/PCIe pinouts
- Discovery of SSD:
  - Use standardized discovery mechanisms to obtain IP address, slot location
  - Discover and manage through RedFish
- Partnering to solve these challenges
  - Comprehensive standard specification in development in SNIA

Management Frameworks for Ethernet SSDs

- Some administration will be done in-band via NVMe™ Admin commands once attached to a host

- But allocation and attachment needs to happen first at scale
  - Drive parameters and health monitoring
  - Encryption / Decryption key management
  - Host usage Authentication and Authorization
  - Logical assignment of drive resources on demand to multiple hosts

- NVMe™ functionality being mapped to RedFish management schema for these purposes

Other Advanced NVMe™ Features

- Data Path Functionality
  - Zoned Name Space Support
  - Key Value namespaces
  - Endurance Group / NVM Set / Namespace Management
  - Future Computational Storage platform for FPGA, Accelerators, etc.
- Part of a Composable Infrastructure
  - Storage “stack” assembled on demand tailored to application needs
  - Drawn from pools of Ethernet Drives, then returned to the pool when finished

Flash Memory Summit

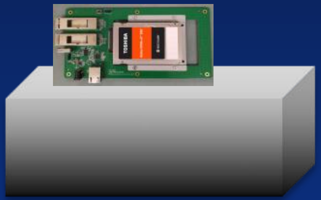
World's First True Ethernet NVMe-oF™ SSD

In-Form Factor Native NVMe-oF™ SSD (Ethernet SSD)

- Standard 2.5" In-Form Factor
- No external components needed
- SFF 8639 / 9639 standardized connector with Ethernet pinout
- Dual-port 25Gbit Ethernet
- RDMA over Converged Ethernet ver. 2 (RoCEv2)
- 675K IOPS @ 4KB Random Read
  - Equivalent performance to PCIe® Gen3x4

Visit the Toshiba Memory FMS Booth #307

Flash Memory Summit



2.5” Ethernet SSD Prototype

Demonstration of 2.5” Ethernet SSD prototype with native NVMe-oF™ support

Example of Ethernet SSD-based AFA architecture

Prototype of a possible AFA platform using EBOF (Ethernet SSD-based)

Flash Memory Summit

Thanks!

Session Agenda

- **Ilker, Samsung – 15 minutes**
- **John - NetApp, Balaji -Toshiba, *Khurram - Marvell* – 40/3 minutes**
- **Woo, SK-Hynix – 15 minutes**
- **Q&A – 10 minutes**

Native NVMe-oF SSD

Khurram Malik

Sr. Product Marketing Manager, Marvell

Current Challenges with NVMe-oF

- SSD Industry is Diverging:
  - Different interfaces (SATA, SAS, PCIe)
  - Different protocols/transports (NVMe-oF variants; NVMe; SCSi …)
  - Different form factor (U.2, U.3, EDSFF S, ESDFF L, EDSFF 3”)
- Challenges:
  - **Standards are diverging instead of converging.**
  - **No clear direction which standard will eventually win.**
  - Selecting a right standard and enable NVMe-oF SSD.
  - Managing two different SSDs skews; NVMe and NVME-oF
  - **Managing two different midplanes; PCIe (NVMe) & Ethernet (NVMe-oF)**
  - **Designing a new chassis to use NVMe-oF SSDs.**

Flash Memory Summit

OCP Kinetic and SNIA Ethernet Drive Pins

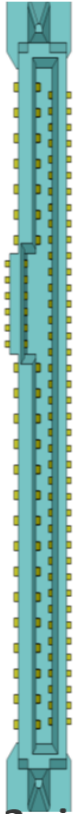
		SATA	SATA Express	SAS	MultiLink SAS	Quad PCIe	USB	OCP Kinetic	SNIA Ethernet Drive
S1	Ground	GND	GND	GROUND	GROUND	Ground	GND	Ground	Ground
S2	Rcvr+	A+	PETp0	PR+	RX0+		SSRX+	RX0+	RX0+
S3	Rcvr-	A-	PETn0	PR-	RX0-		SSRX-	RX0-	RX0-
S4	Ground	GND	GND	GROUND	GROUND	Ground	GND	Ground	Ground
S5	Xmtr-	B-	PERn0	TP-	TX0-		SSTX-	TX0-	TX0-
S6	Xmtr+	B+	PETR0	TP+	TX0+		SSTX+	TX0+	TX0+
S7	Ground	GND	GND	GROUND	GROUND	Ground	GND	Ground	Ground
S8	Ground		GND	GROUND	GROUND	Ground		Ground	Ground
S9	Rcvr+		PETp1	SR+	RX1+			RX1+	RX1+ optional
S10	Rcvr-		PETn1	SR-	RX1-			RX1-	RX1- optional
S11	Ground		GND	GROUND	GROUND	Ground		Ground	Ground
S12	Xmtr-		PERn1	ST+	TX1-			TX1-	TX1- optional
S13	Xmtr+		PERp1	ST-	TX1+			TX1+	TX1+ optional
S14	Ground		GND	GROUND	GROUND	Ground		Ground	Ground

- U2 OCP Kinetic and SNIA Ethernet Drive pin assignments induce crosstalk between adjacent TX and RX pairs, which reduce the max supported channel length. Therefore we recommend different differential pin assignments for 25Gbps PAM2 or 50Gbps PAM4 two Lanes Ethernet application.

U.2 connector pin assignment for Ethernet application

Flash Memory Summit

Name	Pin
GND	S1
S0T+ (A+)	S2
S0T- (A-)	S3
GND	S4
S0R- (B-)	S5
S0R+ (B+)	S6
GND	S7
RefClk1+	E1
RefClk1-	E2
3.3Vaux	E3
ePERst1#	E4
ePERst0#	E5
RSVD	E6
RSVD(Wake#) /SASAct2	P1
sPCIeRst/SAS	P2
RSVD(DevSLP#	P3
IfDet#	P4
	P5
Ground	P6
	P7
	P8
5 V	P9
PRSNT#	P10
Activity	P11
Ground	P12
	P13
	P14
12 V	P15



Pin	Name	SAS & Ethernet Signals proposal1	PCIe & Ethernet Signals proposal2
E7	RefClk0+		
E8	RefClk0-		
E9	GND		
E10	PETp0	TX1+	
E11	PETn0	TX1-	
E12	GND		
E13	PERn0		RX0-
E14	PERp0		RX0+
E15	GND		
E16	RSVD		
S8	GND		
S9	S1T+		
S10	S1T-		
S11	GND		
S12	S1R-	RX1-	
S13	S1R+	RX1+	
S14	GND		
S15	RSVD		
S16	GND		
S17	PETp1/S2T+		TX0+
S18	PETn1/S2T-		TX0-
S19	GND		
S20	PERn1/S2R-	RX0-	
S21	PERp1/S2R+	RX0+	
S22	GND		
S23	PETp2/S3T+		TX1+
S24	PETn2/S3T-		TX1-
S25	GND		
S26	PERn2/S3R-		
S27	PERp2/S3R+		
S28	GND		
E17	PETp3	TX0+	
E18	PETn3	TX0-	
E19	GND		
E20	PERn3		RX1-
E21	PERp3		RX1+
E22	GND		
E23	SMClk		
E24	SMDat		
E25	DualPortEn		

Fig1. U.2 pin assignment

Notes:

Marvell has recommended two high speed signal pin assignment proposals for Ethernet application to minimize connector impacts on the overall Channel Operating Margin(COM).

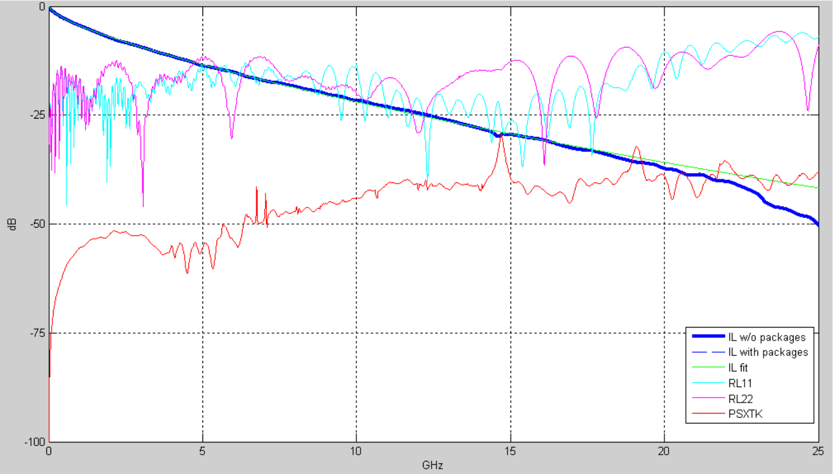
- **Proposal1: Maximize the distance from one differential pair to other signals; (Highlighted as red column)**
- **Proposal2: Based on proposal1 concept, keep pin compatible with PCIe signals. (Highlighted as blue column)**

PCIe Signals
PCIe/SAS Signals
SAS Signals
SAS/SATA Signals

MRVL COM simulation Setup and Results

Based on below long lossy channel, run end to end COM/ERL simulation with two proposed U.2 pin configurations.

- IEEE 802.3by 25GBASE-KR Channel Operating Margin(COM>3dB) without FEC.
- IEEE 802.3bs 50GBASE-KR Channel Operating Margin(COM>3dB,ERL>10dB)

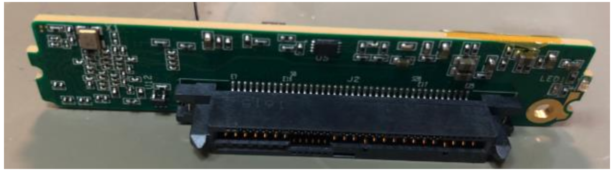


Operation mode	U.2 Pin proposal1 (SAS & Ethernet Signals) Proposal1		U.2 Pin proposal2 (PCIe & Ethernet Signals) Poposal2	
	COM(dB)	ERL(dB)	COM(dB)	ERL(dB)
25Gbps PAM2	3.52	NA	3.65	NA
50Gbps PAM4	3.25	14.08	3.20	14.18

Flash Memory Summit

Convert NVMe SSD to NVMe-oF SSD

NVMe-oF Converter Controller interposer in a carrier



NVMe-oF Converter Controller Interposer (SSD side)

NVMe-oF Converter Controller Interposer (network side)
*(*8639 is used to drive 2x25Gb Ethernet)*

NVMe-oF Converter Controller Interposer (profile)
Connected to U.2 (non-carrier)

Enabling NVMe-oF

Simple, low RBOM, low power backplane

Enabling NVMe-oF

- Marinating NVMe and NVMe-oF support
  - NVMe
  - NVMe-oF : ROCEv2 ; TCP
- NVMe-oF Converter Controller
  - Can fit interposer
  - Can fit inside U.2/EDSFF
  - Can be merged with SSD Controller
- Re use of backplane
  - Re use 8639/9639
  - No changes to mid plane
  - Swap IOM
- No extra enclosure expense (other than IOM)
- Single SSD can work both PCIe and Ethernet (Better inventory management)

Thanks!

Session Agenda

- **Ilker, Samsung – 15 minutes**
- **John - NetApp, Balaji -Toshiba, Khurram - Marvell – 40 minutes**
- **Woo, SK-Hynix – 15 minutes**
- **Q&A – 10 minutes**

- Woo slides

Thanks!

Session Agenda

- **Ilker, Samsung – 15 minutes**
- **John - NetApp, Balaji -Toshiba, Khurram - Marvell – 40 minutes**
- **Woo, SK-Hynix – 15 minutes**
- **Q&A – 10 minutes**

Q/A

Thanks!