SCSI Trade Association

# King of the Hill

Flash Memory Summit

August 11th, 2016

# Panelists

- Mohamad El-Batal – Sr. Director of Architecture Seagate

- Tom Friend – Director of Industry Standards Toshiba America Electronic Components Inc.

- Fred Knight – Principal Engineer, CTO's Office, NetApp

# Fred's New Bike

# NVMe & All-Flash Arrays
# Flash Memory Summit 2016

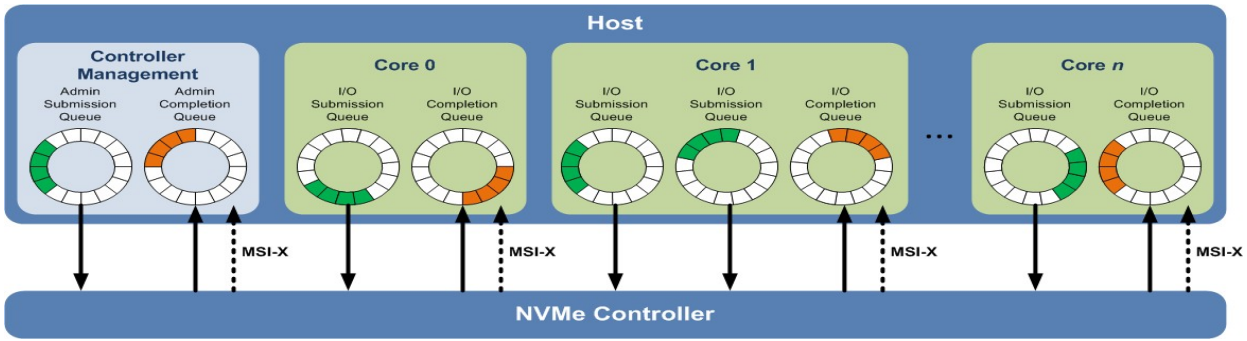
Mohamad El-Batal

# NVMe – Architected for NAND & NVM

- NVMe is a standardized high performance software interface for PCIe SSDs
- Standardized register set, feature set, and command set, where there were only proprietary PCIe solutions before
- Architected from the ground up for NAND and next generation NVM
- Designed to scale from Enterprise to Client systems
- Developed by an open industry consortium with a 13 company Promoter Group, which is now grown in the 100's
- Continuously evolving standard to keep up with new market demand and new technology opportunities
- Latest additions:
  - NVMe 1.3 (forthcoming core release)
  - NVMe MI (Management)
  - NVMeoF (Fabrics)

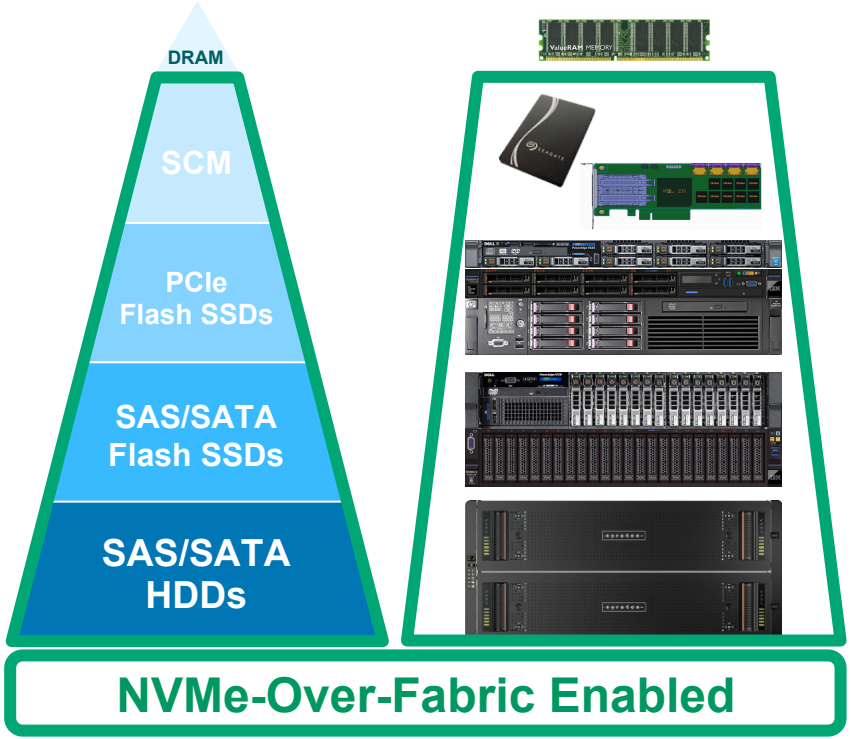
# NVMe Technical Overview

- Supports deep queues (64K commands per queue, up to 64K queues)
- Supports Controller Management through a separate Admin queue pairs
- Supports MSI-X and interrupt steering
- Streamlined & simple command set (13 required commands)
- Designed to scale for next generation NVM, agnostic to NVM type used
- Optional features to address target segment:
  - Data-Center/Enterprise: End-to-End data protection, Multi-Streams, Multi-Namespace, Persistent-Reservations, Dual-port, PRP/SGL, Controller-Memory-Buffer(CMB) … etc.
  - Client: Autonomous power state transitions, Host-Memory-Buffer(HMB) for DRAM-Less support ... etc.



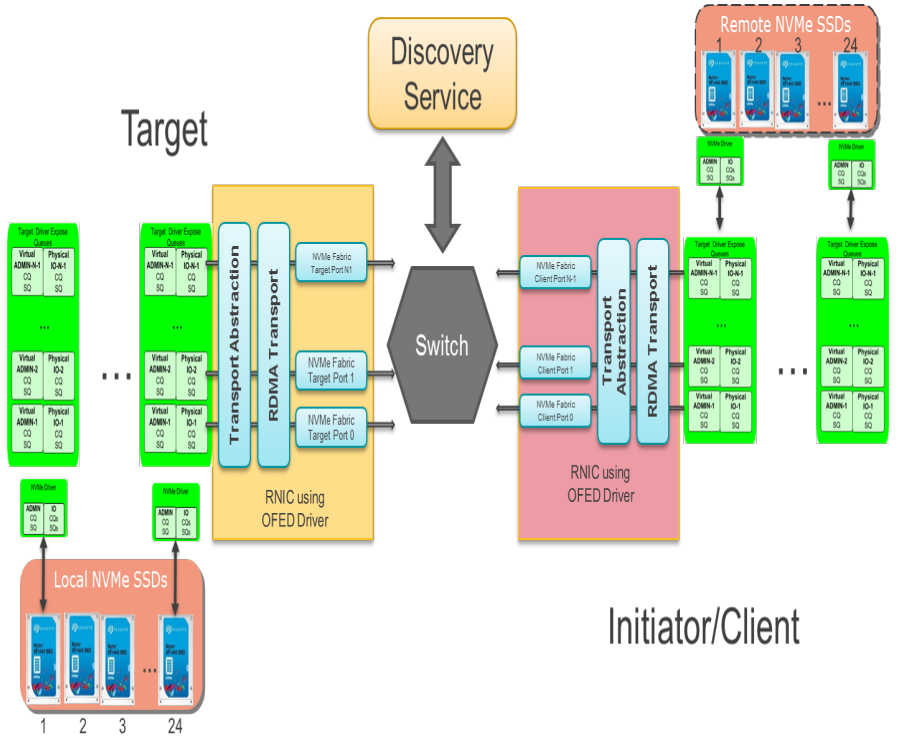
# NVMe-over-Fabric Storage Hierarchy

- Unified Fabric with NVMe protocol enables hardware automated I/O Queues and Interrupts with SGL support
- Inherent parallelism of multiple I/O Queues is exposed
- NVMe command and data-Integrity structures are transferred end-to-end
- Maintained consistency between fabric transports by standardizing definition
- Fabric Agnostic Ethernet, IB, FC …etc, with common NVMe requirements
- Optional Ultra-low-latency NVMe-Direct zero-copy RDMA protocol data transfers with minimal Target CPU overhead



# NVMf w/ NVMe-Direct

- Single/Multiple SSDs shared by many Hosts
- Scaling AFAs to Multi-Million IOPS with DAS like latency
- Establishing the communication path between the host and the NVMe subsystem through discovery and authentication
- Enabling the controller Admin Queue, and I/O Queue(s) to be setup and then utilized
- Reliable in-order delivery of command and response capsules between a host and NVMe subsystem
- Command capsule is created in client side and place in target's memory so that NVMe command is directly processed in NVMe SSD controller
- Optional P2P RDMA on NVMe controller to support Remote-DMA access of NVMe register sets
- Optional Controller-Memory-Buffer(CMB) NVMe SSD support to enable even faster NVMe-Direct command queue placement and local memory execution



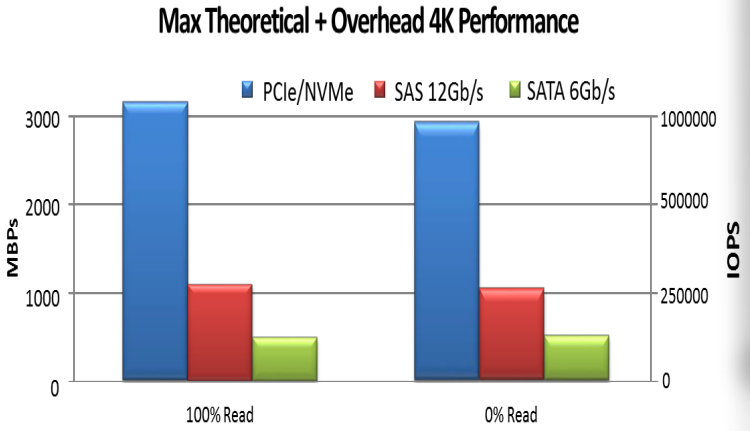
# Why NVMe is Better?

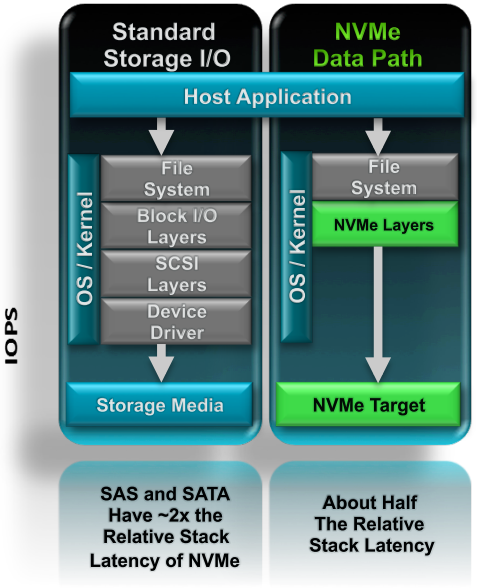
- NVMe delivers lowest latency overhead compared to any standard storage interface
- NVMe x4 PCIe-Gen3 U.2 & M.2 SSD's deliver > 3.2GB/s of Read/Write Performance
- NVMe Read & Write Performance ~ 3X of 12Gb-SAS ~ 6X of 6Gb-SATA
- Fabric Agnostic Ethernet, IB, FC …etc, with common NVMf requirements
- Optional NVMe-Direct RDMA protocol data transfers & CMB for Zero Target CPU/DRAM overhead
- Best in class $/IOPs JBOF & IBOF enclosure designs

Mean Host Turnaround Times (microseconds)

| Kernel 3.10 | | |
|---|---|---|
| | SAS | NVME |
| RQ Affinity = 1 | 85.3 | 11.0 |

| Kernel 4.5 | | |
|---|---|---|
| | SAS | NVME |
| RQ Affinity = 1 | 114.6 | 14.9 |



# Thank You! Questions?

## Visit Seagate Booth #505

Learn about Seagate's ever-expanding portfolio of SSDs, Flash solutions and system level products for every segment

www.seagate.com/flash

TOSHIBA

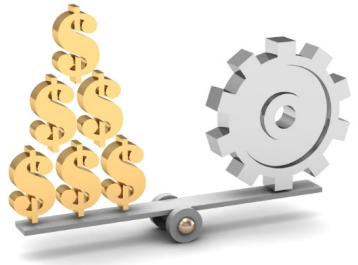
Leading Innovation

# Which interface is King Of The Hill?

## Should You Build Your All-Flash Array with SAS, SATA, or NVMe?

# SATA benefits

- **Lowest cost: (TCO)**
  - Initial (drives) (CAPEX)
  - Initial (interface & cabling) (CAPEX)
  - Power (watts per GB) (OPEX)



- **Standard form factor**
  - Fits into array racks easily (CAPEX)
    - Mechanical and electrical design is easy and modular


- **Standard features**
  - Best known and understood interface

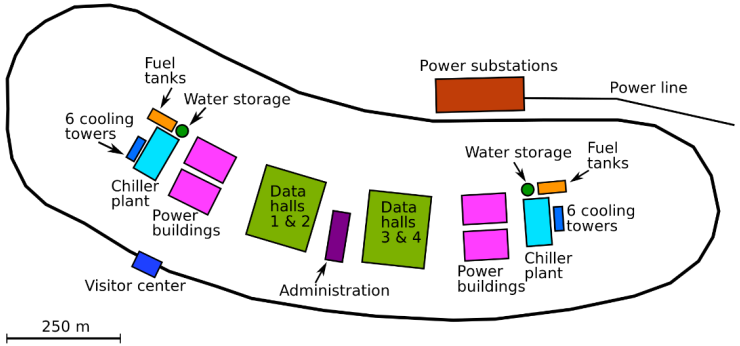
# SATA case study- HighperScale.com

- **Design goals:**
  - Scalable
    - Scale out- not up
    - Tune system for responsiveness
  - Reliable
    - Add more data centers for availability
    - Redundant across locations, networks, power grid
  - Low CAPEX:
    - Standard hardware (COTS)
    - Standard software (Open Source)
  - Low OPEX:
    - Low power and heat
    - No licensing fees

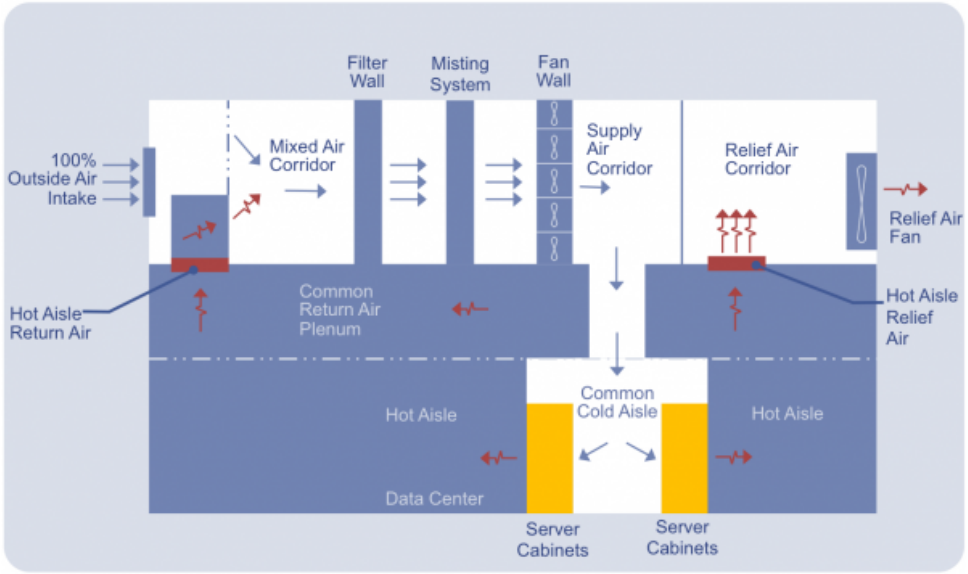
# Physical scale

- **Data Centers are large**

- **Data Centers need power and cooling**



# Physical design




Open 19 or Open 21?

Standard components

OPEN
Compute Project

# Software components



# Storage in the large data centers

- **SATA SSDs for hot data**
- **SATA SSDs or HDDs for warm data (Nearline)**
- **SATA HDDs for cold data (Glacier)**

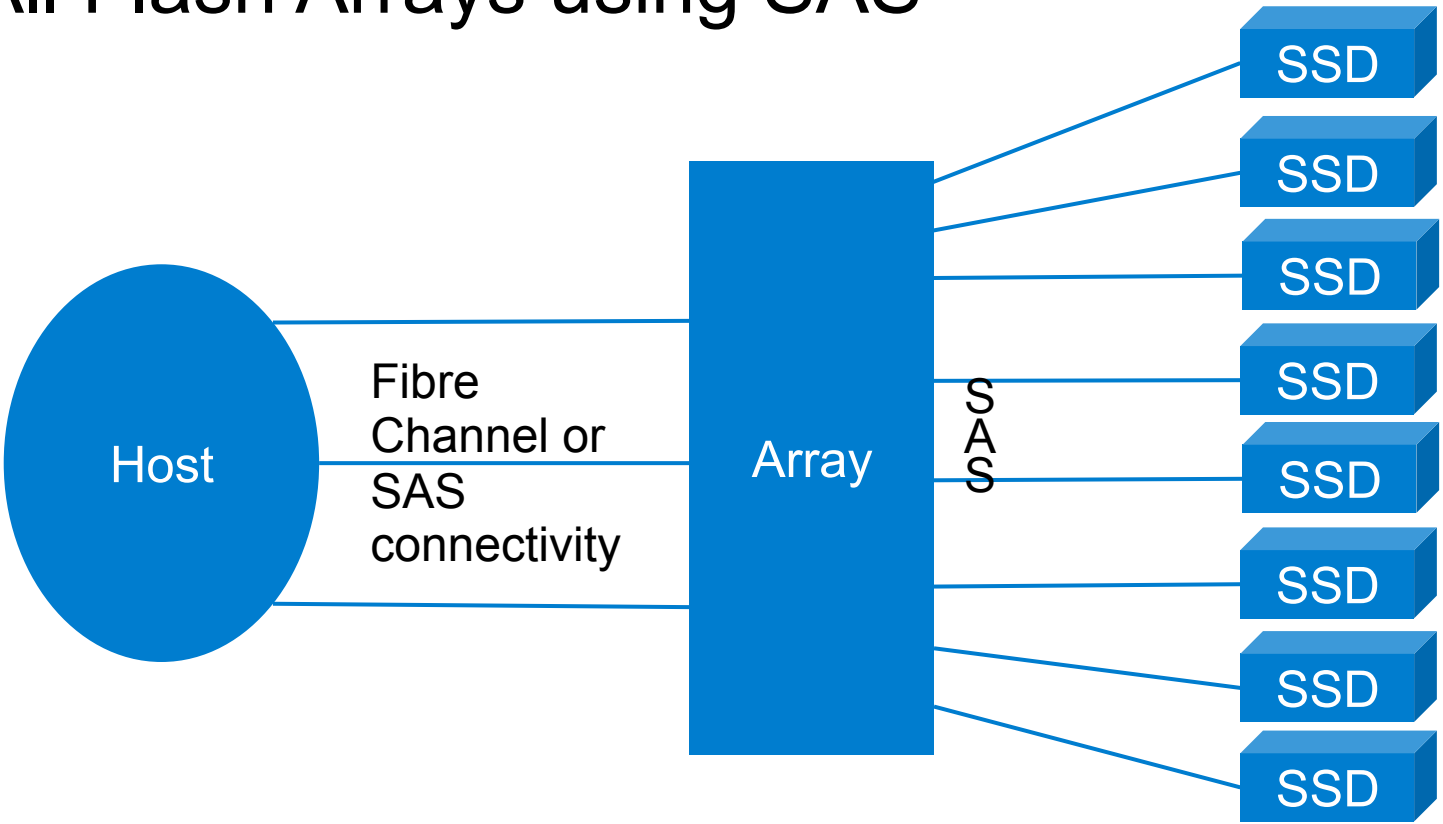
# All Flash Arrays using SAS

Frederick Knight
Principal Standards Technologist
8 Aug 2016

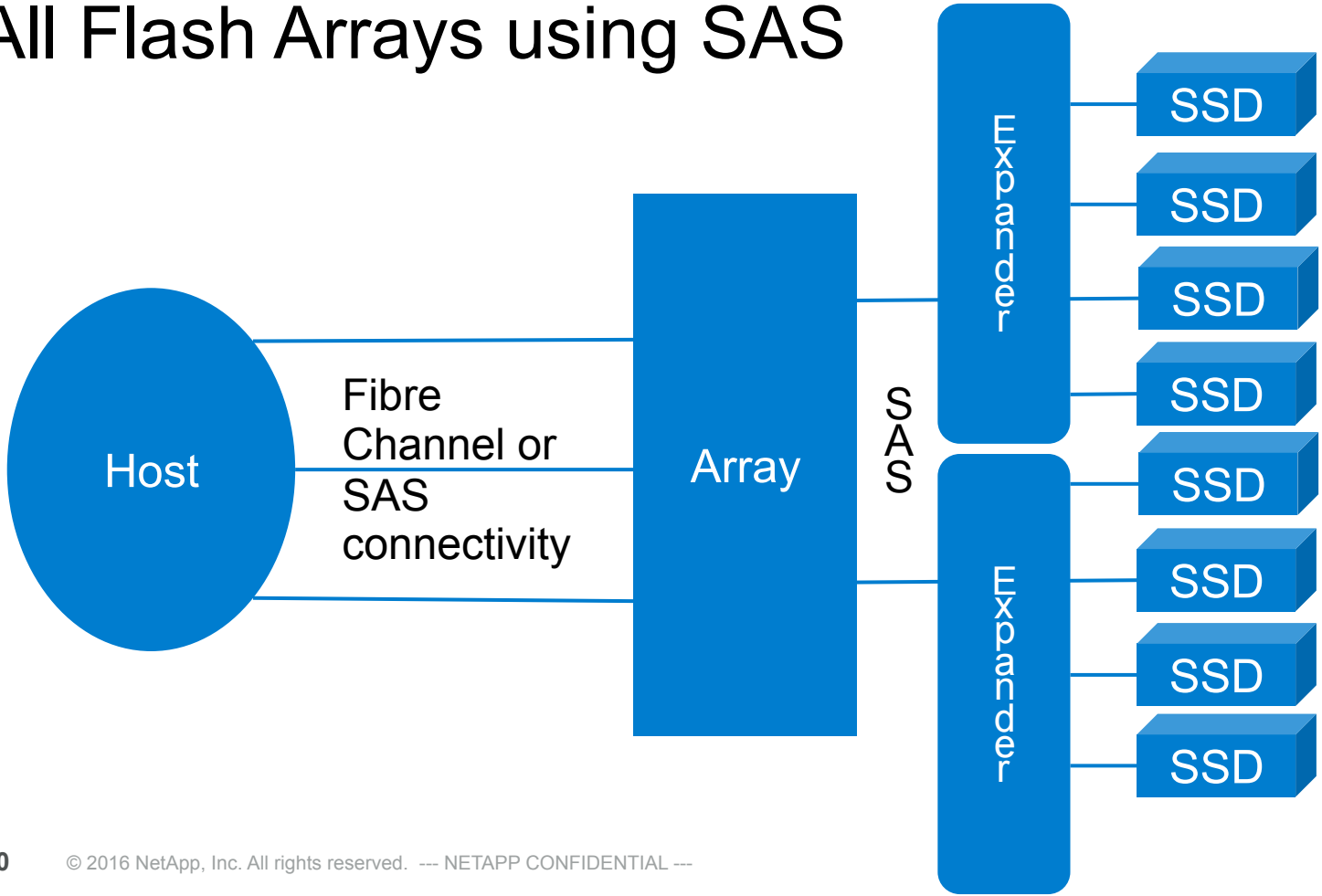


NetApp

# All Flash Arrays using SAS



# All Flash Arrays using SAS



# All Flash Arrays using SAS

- SAS is a modern interconnect
  - Actively being developed
  - New wires + new commands – streams, BG control, hints
  - Highly Scalable (dev count, SAS expanders, 19” BP, 6M copper, 300M optical)
  - 3Gb, 6Gb, 12Gb, 24Gb speeds
- SATA is based on IDE
  - Used on the original IBM PC computers (minimal ongoing development)
  - Frozen at 6Gb
- NVMe is bleeding edge
  - Very active development (lots of rapid changes)
  - Reduced features and capabilities (today)
  - PCIe gen 4 backplane slots (~3” FR4, cables under development)

# All Flash Arrays using SAS

- SAS is:
  - Full featured,
  - Scalable,
  - Stable,
  - Reliable, and
  - High Performance

- The best choice for today.

# SAS Value Proposition

SCSI Trade Association

- Reliability – Error handling, robust Storage optimized channel models
- Performance – Highest performance per lane w/capability for x2 & x4
- Scalability – Scalable to 1,000s of devices
- Serviceability – True hot-swap capabilities to add/remove media & cables
- Manageability – Storage management built into the standard


Serial Attached SCSI

# Questions

STA™
SCSI Trade Association