NVMe™ Lessons Learned

Deploying NVMe™ Flash in Real Systems

Facilitated by Tom Heil, Senior Systems Architect & Distinguished Engineer, Broadcom

Forum A-12: NVMe and PCIe SSDs

Agenda

- Introduction
- Panel Member Presentations
- Q&A

NVM Express® Storage Genesis

- NAND Flash in Mainstream IT

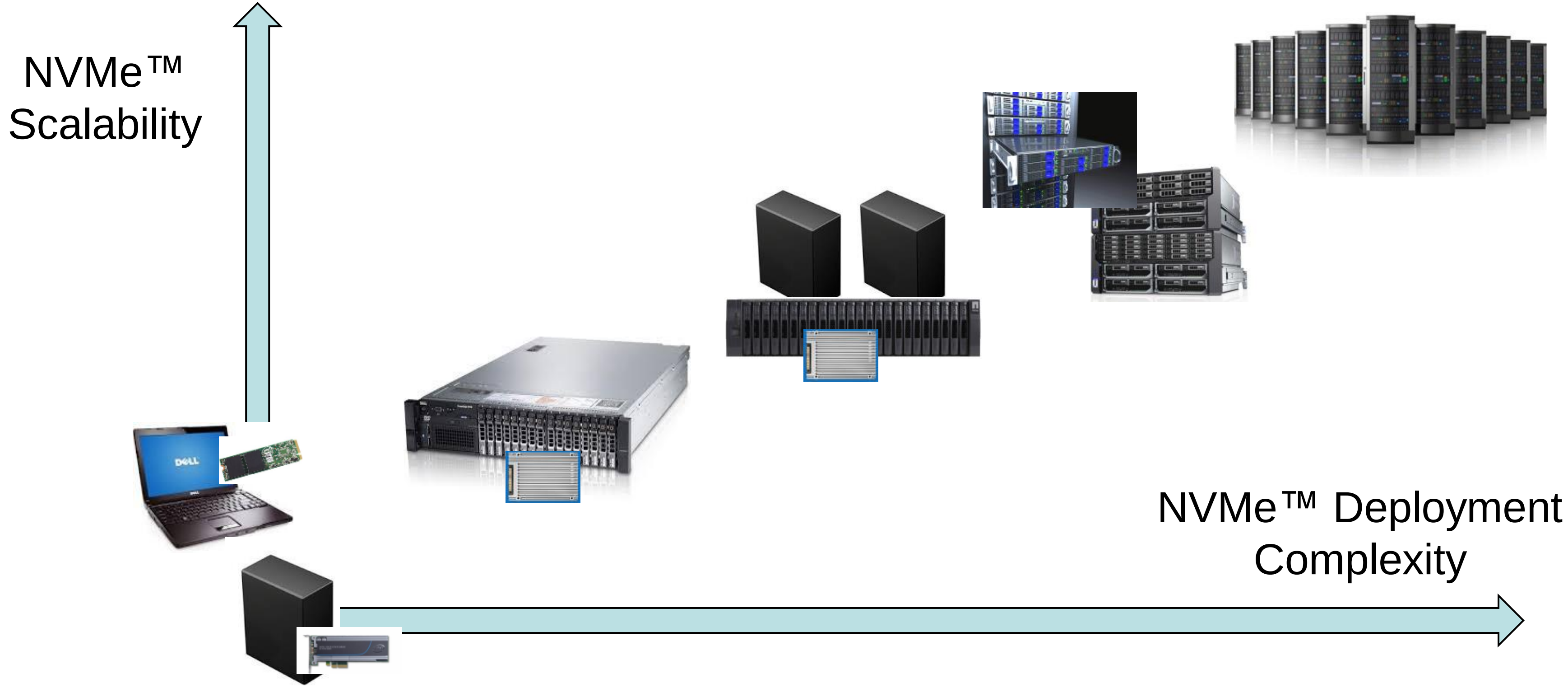
- PCIe Flash Host Adapters

- Standard Storage-optimized PCIe Form Factors: U.2, M.2

- Standard Flash-optimized Storage Protocol: NVMe™

This is just the beginning …

NVMe™ Across the Storage Infrastructure



Challenges Enabling NVMe™ Across the Storage Infrastructure

- PCIe signal routing, power, thermal, clocking
- Cabling: inside the box > box-to-box
- Hot-plug serviceability: synchronous > asynchronous
- Manageability: device > device bay > enclosures
- Performant, SSD-optimized Data Protection: RAID & replication
- High Availability Dual-Domain Topologies
- Storage sharing: sub-rack > rack > data center > cross geography
- Creative M.2 use models

Gary Kotzur

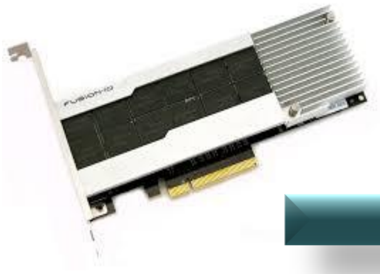
Executive Director / Senior Distinguished Engineer

Platform Solution Considerations

- Industry
- Standards
- Drive Metrics
- Platform

NVMe™ SSD Drive History

Great Performance!



End-User Needs

- Front Access
- Hot-plug ability
- Form-Factor(s)
  - Standard Form-factor
  - Higher power
- Connector
  - Standard
  - Multiple lanes
  - Backward Capability
  - No active backplane devices
- Protocol
  - Industry standard
  - Inbox
  - High Performance
- Management

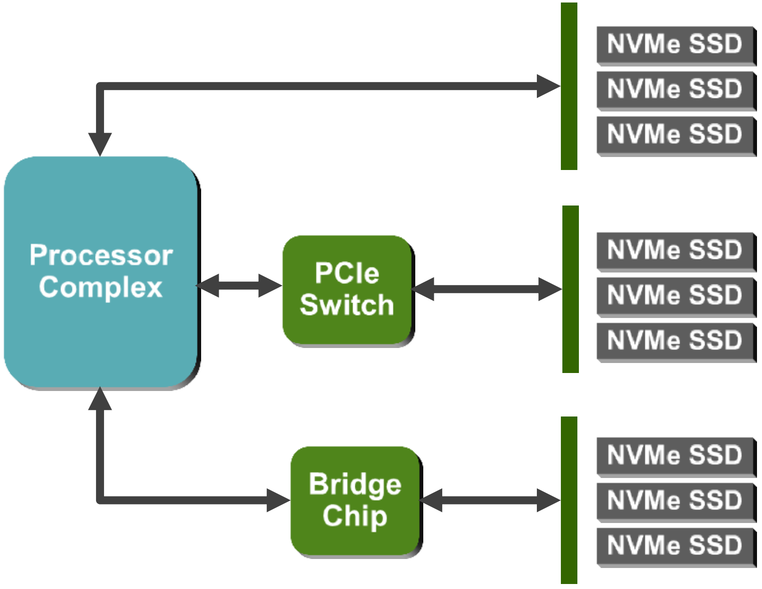
Solution

- 2.5” 15mm FF
- 25W power envelope
- SFF-8639 (U.2 profile)
  - x4 or x2/x2 ports PCIe
  - x2 SAS or x1 SATA
- NVMe protocol

Benefits

- High performance of PCIe
- Hot Serviceability
- Compatibility with 2.5” SAS/SATA
- 25W power envelope
- NVMe: higher performance
- Open driver with inbox support
- Reduce component counts

Platform Solution



Challenges

- Performance balance
- Power
- Thermal
- Mechanical
- System Management
- Serviceability
- Reliability
- Availability
- Security
- Co-existence with SAS/SATA
- Connectivity options
- Clocking
- Resets
- Dual-port

Chris Petersen

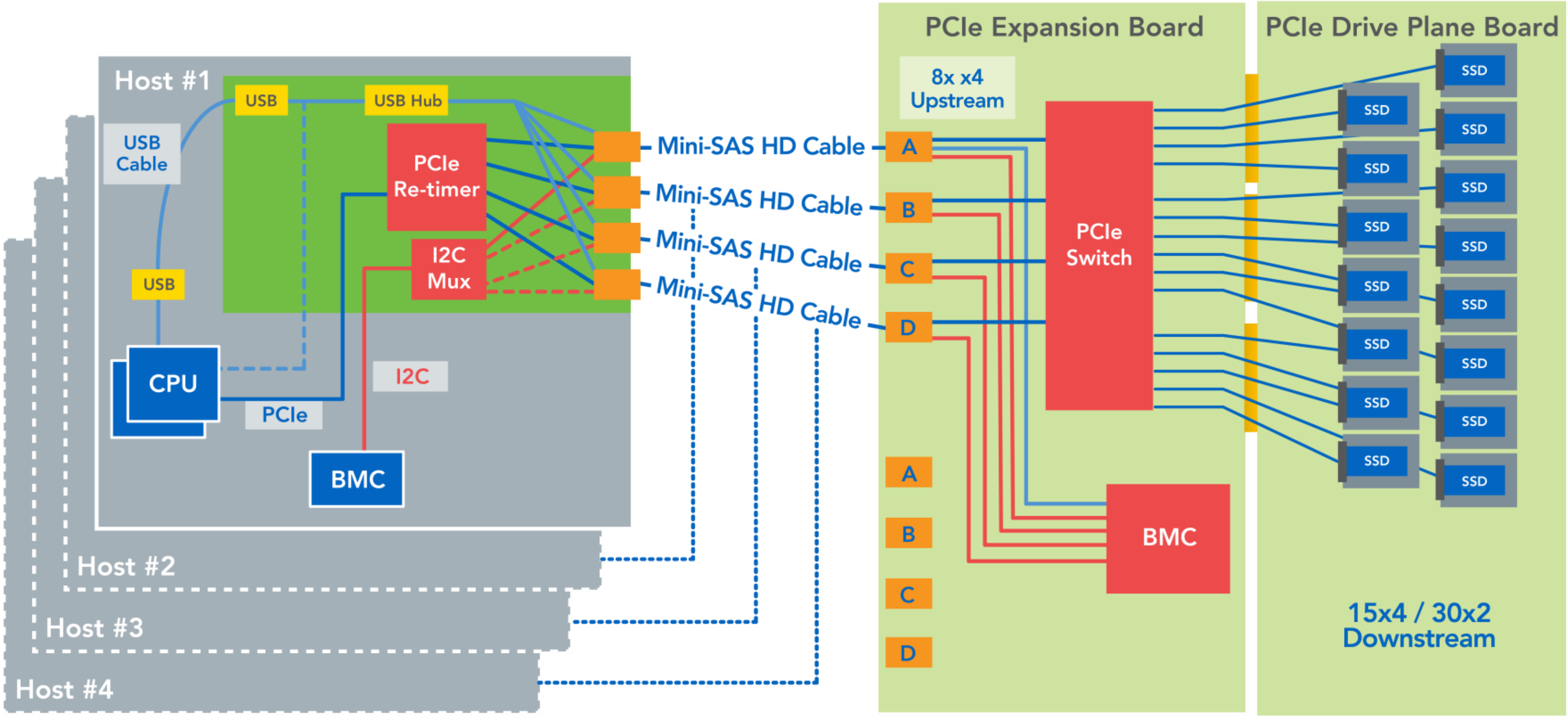
Hardware Systems Architect

facebook

NVMe™ System Objectives

- Modular and Flexible
- Scalable
- Efficient

JBOF Architecture



Challenges and solutions

- Open source + Surprise hot-plug
  - NVMe and PCIe Advanced Error Reporting (AER) drivers
  - Downstream Port Containment (DPC) driver
  - All 1’s completions
  - NVMeCLI

Challenges and solutions

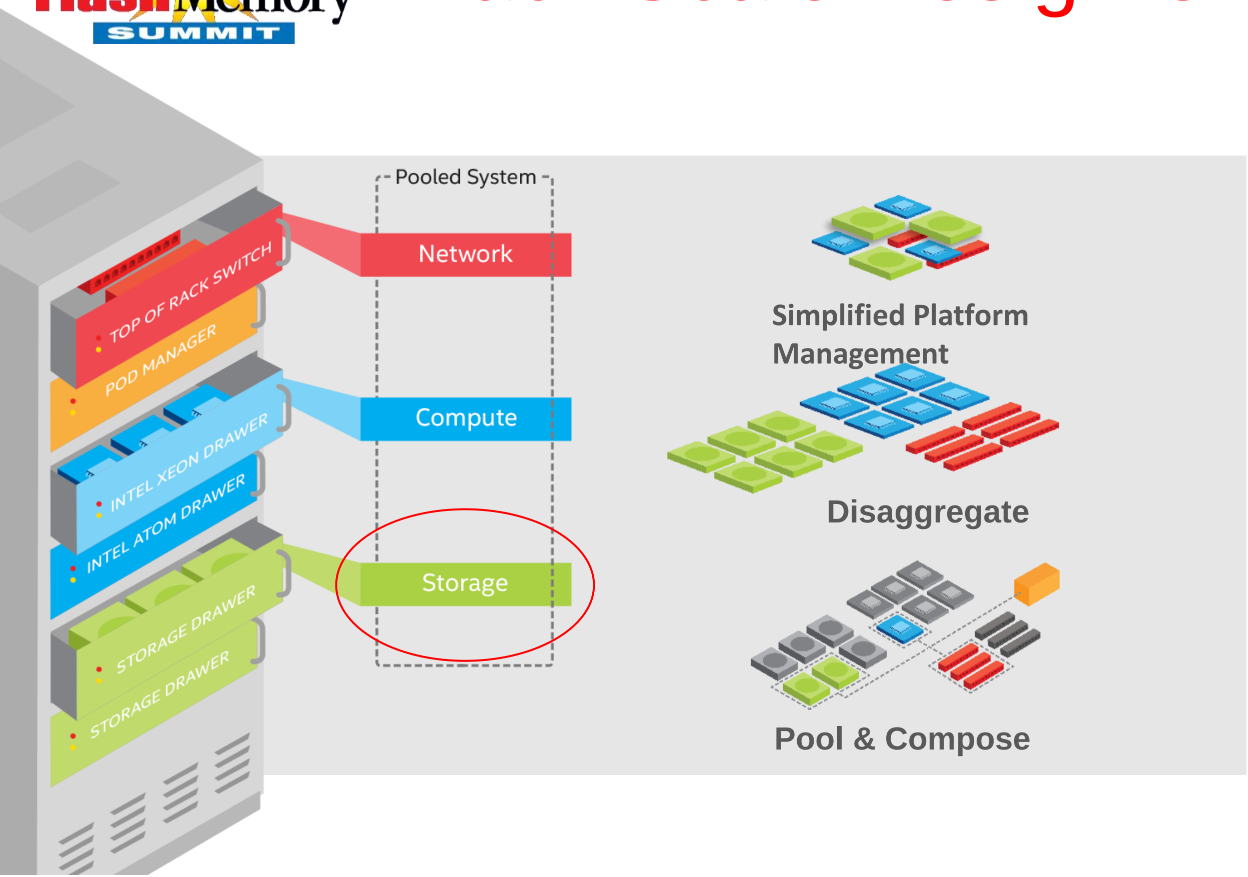
- Cabling
  - Mini-SAS HD cables with full sideband
- M.2
  - Add hot-plug and thermal management

Hot Tier Pooled Storage

Don Faw

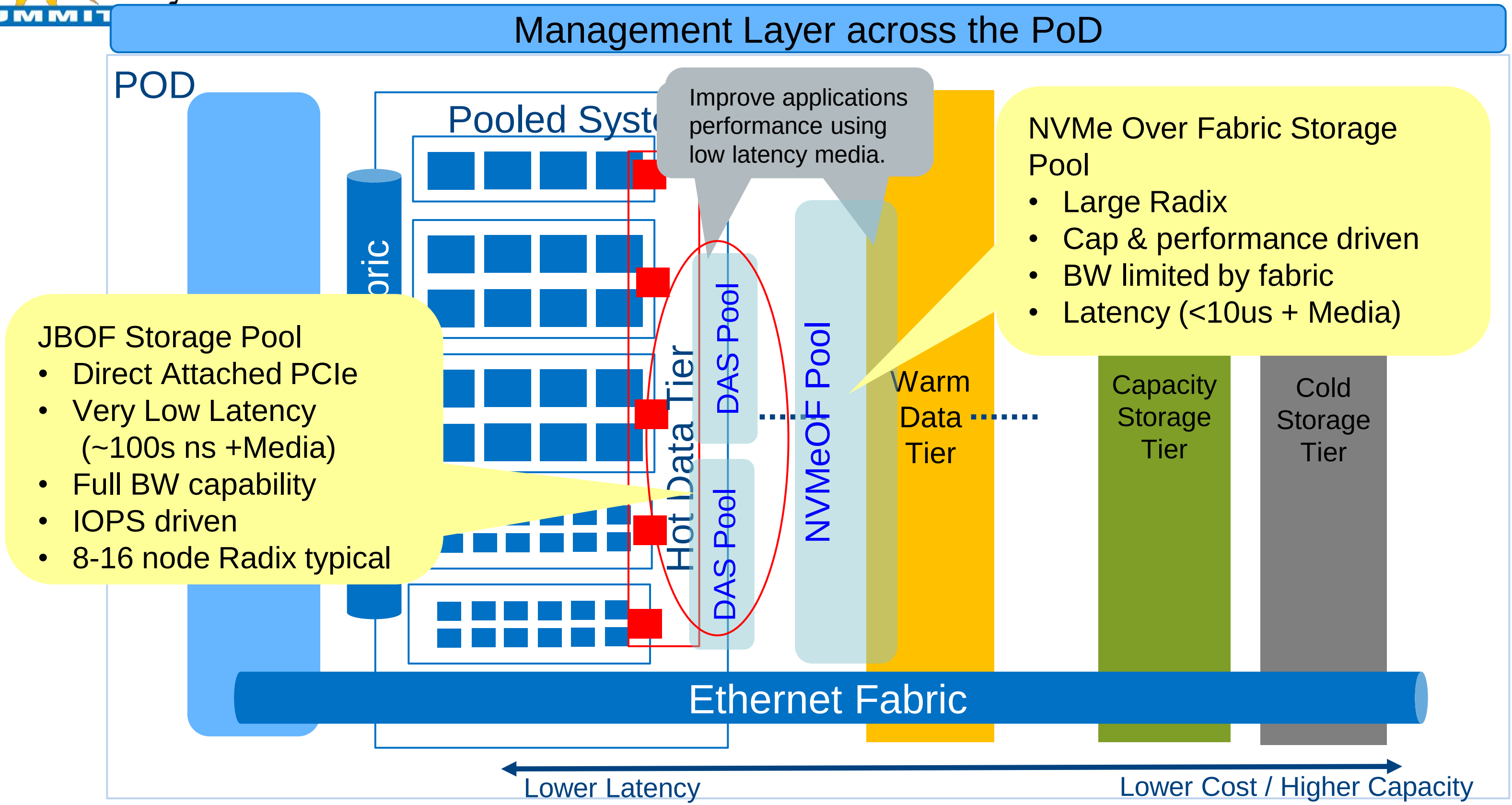
Principal Engineer, Intel Corp

Rack Scale Design Overview

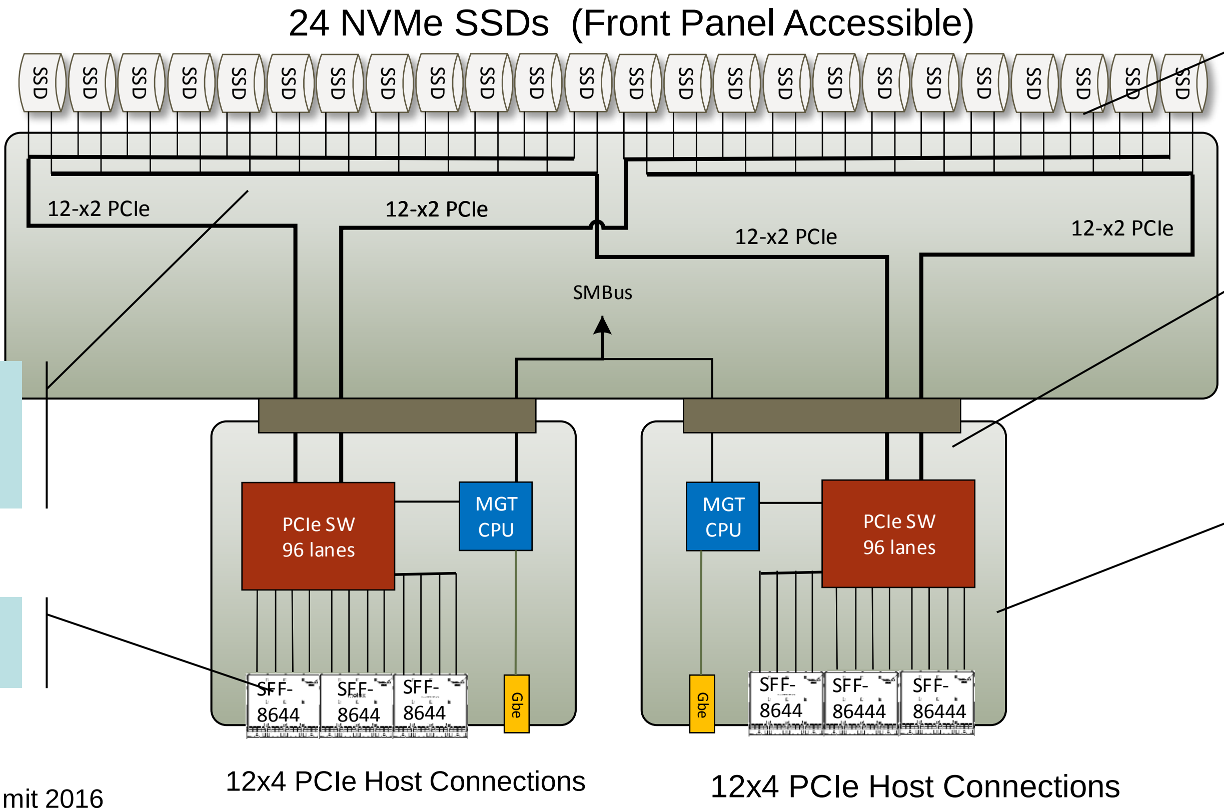


- ✓ **User-Defined Performance**
- ✓ **Maximum Utilization**
- ✓ **Interoperable Solutions**

Intel® RSD Vision



Hot Tier JBOF



Dual Port SSD Active-Active

Independent paths to SSD via multiple switches.

Serviceable Modular Switch module.

x2 Provides more efficient connectivity model, Flat Topology

Each host port has access to all SSD

Challenges

- Multi host connection
  - Lots of Bulky cables vs drive serviceability
    - Keep cables stationary if possible, reduce cable mgt
    - SSD density vs serviceability, Font panel accessibility
  - Host radix within the rack
    - Multiple JBOF pools/rack vs ganging PCIe switches
    - NVMe Over Fabric for larger radix, multi-rack
  - Host Clocking requires SRIS
    - Repeaters to drive PCIe cables need SRIS capability
- Drive telemetry
  - Allocation of storage resources based on drive perf parameters
  - R/W BW perf/namespace, QoS, OOB accessible

Ziv Serlin

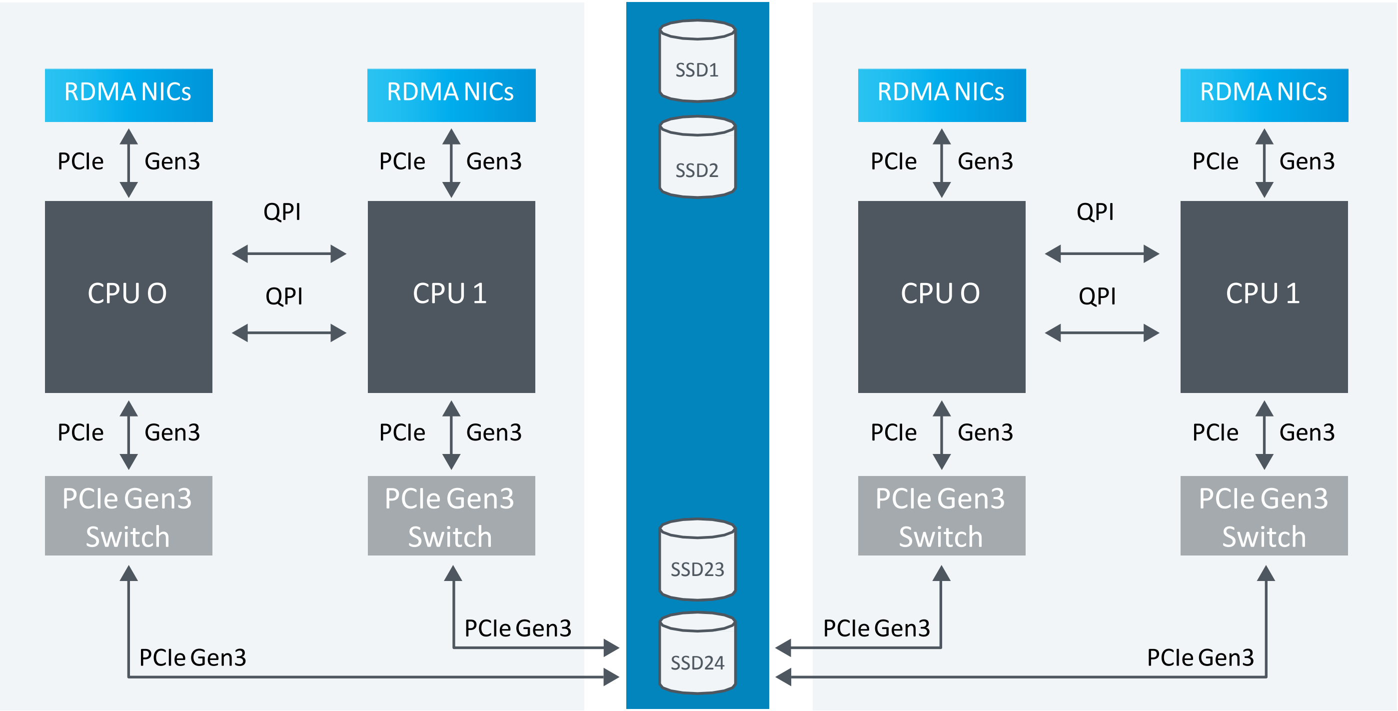
Director System Architecture

E8 storage

Overall E8 System Objective

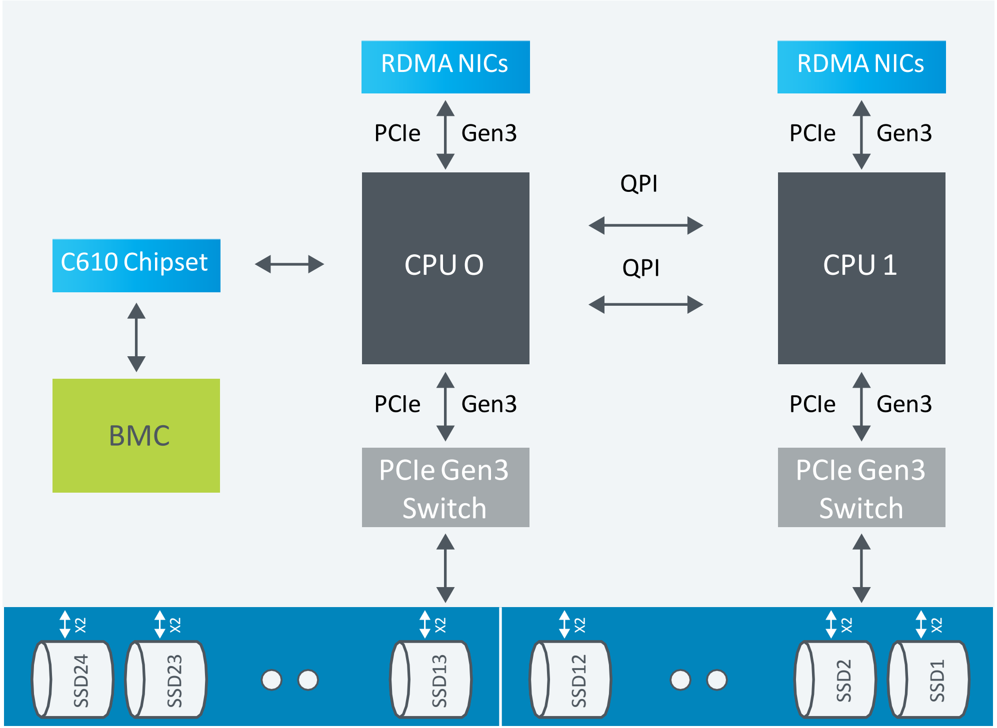
- Very high density
- Latency on par with local NVMe
- Extract the full performance of the SSDs
- Easy to use & highly scalable
- Tier-1 reliability
- Rich feature set
- Cost-effective and low TCO

Dual controller NVMe™ architecture



Highly Available Spec
No SPOF
Passive Midplane
Power Control / Protection
MI I/F
Node 2 Node Communication

Node concept NVMe™ Dual controller

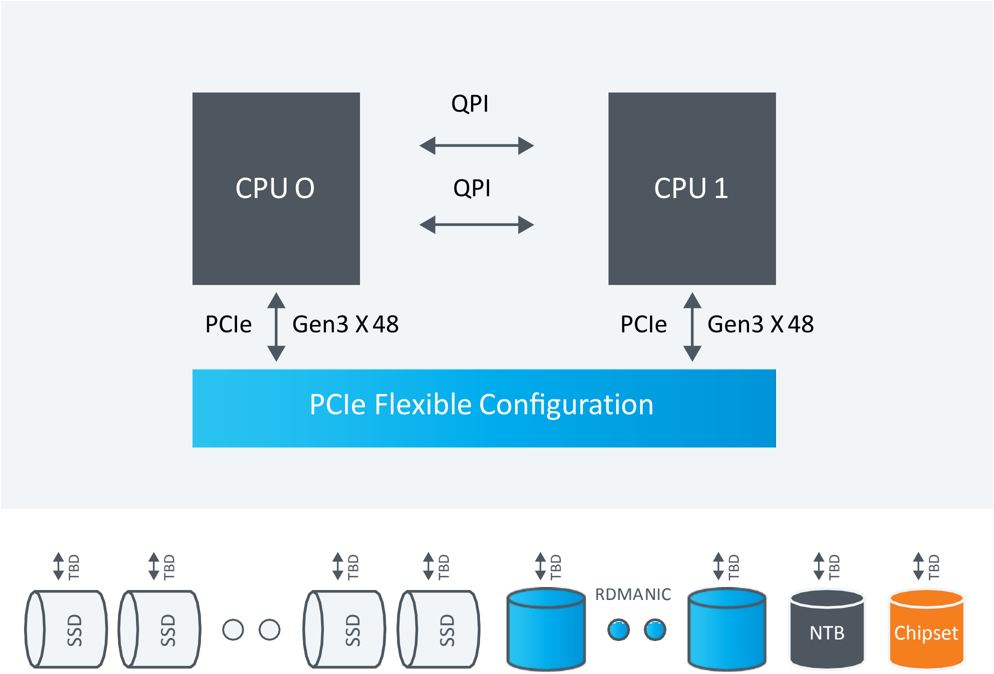


Node Spec

- QPI Overhead
- Networking Vs. SSD BW
- BMC Mngmt.
- SR/SI of NVMe SSD
- Error recovery

*SR/SI – Surprise removal / Surprise insertion

Future PCIe Fabric with NVMe™



Node Spec
Peer 2 Peer

E8 System Challenges

- PCIe physical layer with HA enclosure (backplane)
- Power/Reset sequence in dual port environments
- SI & SR of controller and NVMe devices
- BIOS support for unexpected errors during bootup
- Cooling, Power & power protection

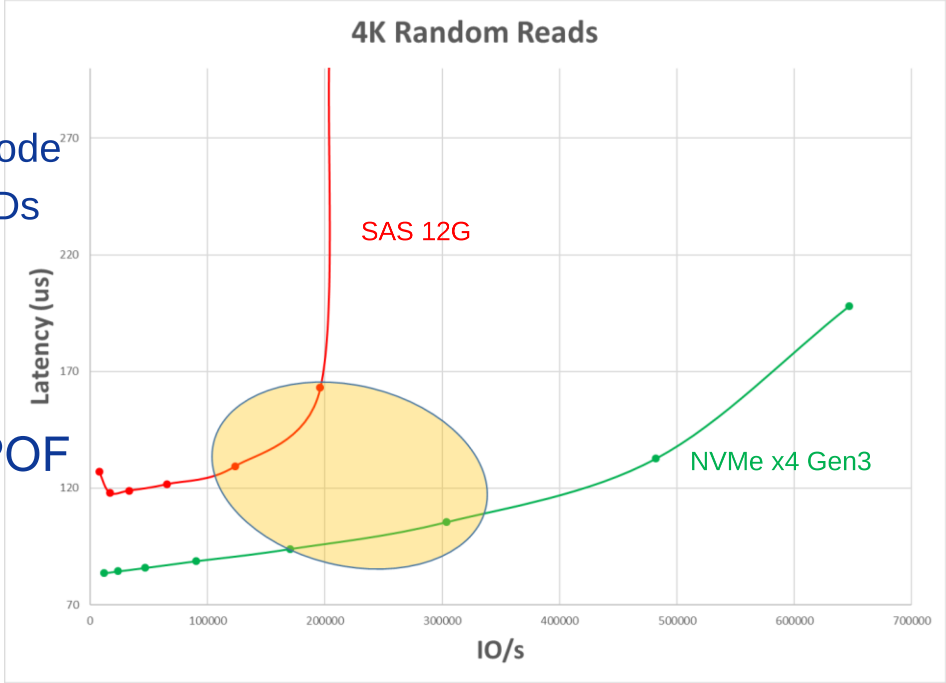
*SR/SI – Surprise removal / Surprise insertion

Tim Emami

Technical Director

All-Flash Arrays need "Value" Dual-Ported NVMe™ SSDs

- Parity Raid Economics
  - lowest $/GB ("Data-Center grade" NAND)
  - Aggregated capacity and performance per node
  - "Repair-in-place" mechanisms for hi-cap SSDs
- "Modest" peak performance
  - Reduce latency at lower concurrencies
  - Throughput improvements over 12G SAS
- "Hi-Availability" configurations with no SPOF
  - Reliable dual-ported NVMe..
  - PI and SGL support
  - NS Reservations
  - Robust management infrastructure
    - NVM-MI development is now under-way



NVMe-oF™ (NVMe™ over Fabrics) – Overlaps with AFAs

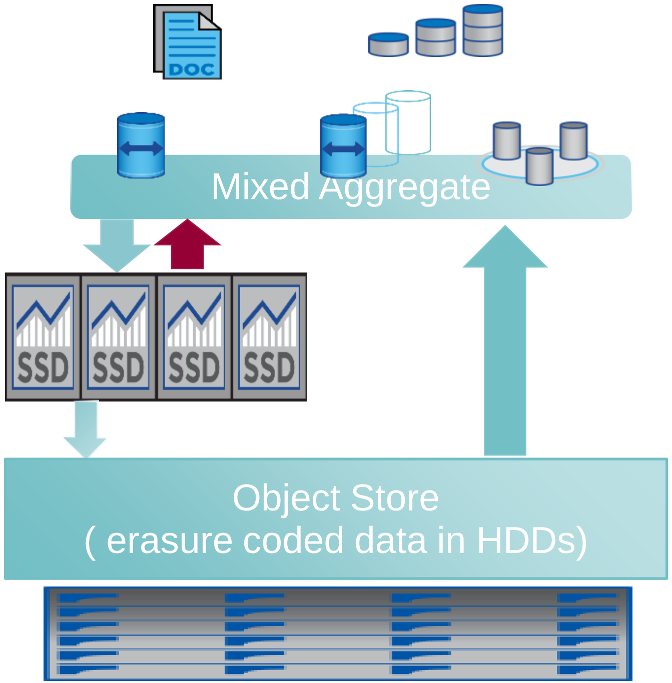
- DASD created under-utilized ”Islands of Flash”
  - NVMe-over-Fabric expansion allows dynamic provisioning of Compute to Flash ratio
- RDMA enabled Fabrics (IB, RoCE..) dramatically reduce the cost of remote vs local access
  - Addresses OS bottlenecks with traditional Block interfaces for AFAs
  - NVMe-over-Fabric enables “scale-up” expansion
- Needs “Value” Dual-Ported NVMe SSDs plus;
  - Controller Memory Buffer
  - SR-IOV/Multi NS
  - SRIS/NRIS

NVMe™ Device/Enclosure Management

- Traditional Enterprise “Scale-up” Storage architectures require an “Enclosure Management Plane”
  - Physical/Protocol
  - SW/Logical
- SAS was meant for expansion and included a SMP/SES management overlay
- NVMe started life as a “fast path” interconnect for locally attached NVM
- NVMe-MI is meant to close the management gap
  - Mix of requirements from “Server” and “Storage” folks..
  - SMBus, PCIe “In-Band”, Ethernet
  - NVMe-MI command set
    - MCTP binding..?
  - Needs some processing power/SW in the enclosure
  - MI Standard is still evolving (e.g. PD pin..)

Tiering / Hybrid Use-Cases

Hybrids remain relevant, but..



- Performance optimized SFF HDDs are being displaced by SSDs
- Capacity optimized LFF HDDs offer the lowest bit cost for colder data
- Tiering remains relevant but; "Hybrid" implies a very different mix of devices/media
  - Not about "IOPs Efficiency" anymore..
  - The right data on the right media; to reduce the Total Cost Of Ownership

Panel Members

Name	Title	Company	Email
Tom Heil	Senior Systems Architect Distinguished Engineer	Broadcom	tom.heil@broadcom.com
Gary Kotzur	Executive Director/Senior Distinguished Engineer	Dell	Gary_Kotzur@dell.com
Chris Petersen	Hardware Systems Architect	Facebook	cpetersen@fb.com
Don Faw	Principal Engineer, Platform Architect	Intel	donald.l.faw@intel.com
Ziv Serlin	Director System Architecture	E8 Storage	ziv@e8storage.com
Tim Emami	Technical Director	Network Appliance	Tim.Emami@netapp.com

Q&A

Thank You!