NVM Express™ Infrastructure - Exploring Data Center PCIe® Topologies

January 29, 2015

Jonmichael Hands – Product Marketing Manager, Intel Non-Volatile Memory Solutions Group

Peter Onufryk – Sr. Director Product Development – PMC-Sierra

Moderator: Ravi Chari – Storage Technologist – Oracle

View recorded webcast at https://www.brighttalk.com/webcast/12367/141221

Legal Disclaimer

All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest Intel product specifications and roadmaps.

Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Learn more at intel.com, or from the OEM or retailer.

Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade.

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document. The products described may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Results have been estimated or simulated using internal Intel analysis or architecture simulation or modeling, and provided to you for informational purposes. Any differences in your system hardware, software or configuration may affect your actual performance.

Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.

No computer system can be absolutely secure. Intel does not assume any liability for lost or stolen data or systems or any damages resulting from such losses.

Cost reduction scenarios described are intended as examples of how a given Intel- based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.

For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.

*Other names and brands may be claimed as the property of others.

Agenda - NVM Express™(NVMe™) Infrastructure

- What is NVMe?
- NVMe advantages over SATA™
- NVMe driver ecosystem
- PCIe® form factors, cables, and connectors
- Link extension and port expansion for PCIe
- PCIe Solid-State Drive Topologies
- NVMe Management

NVM Express™ is a standardized high performance software interface for PCI Express® Solid-State Drives

NVMe is industry driven to be extensible for the needs of both the client and the data center

Architected from the ground up for SSDs to be more efficient, scalable, and manageable

What is nvm EXPRESS™?

"If I had asked people what they wanted, they would have said faster horses"
- Henry Ford

NVM Express™ Community

NVM Express, Inc.

Consists of more than 75 companies from across the industry

Promoter Group

Led by 13 elected companies

Technical Workgroup

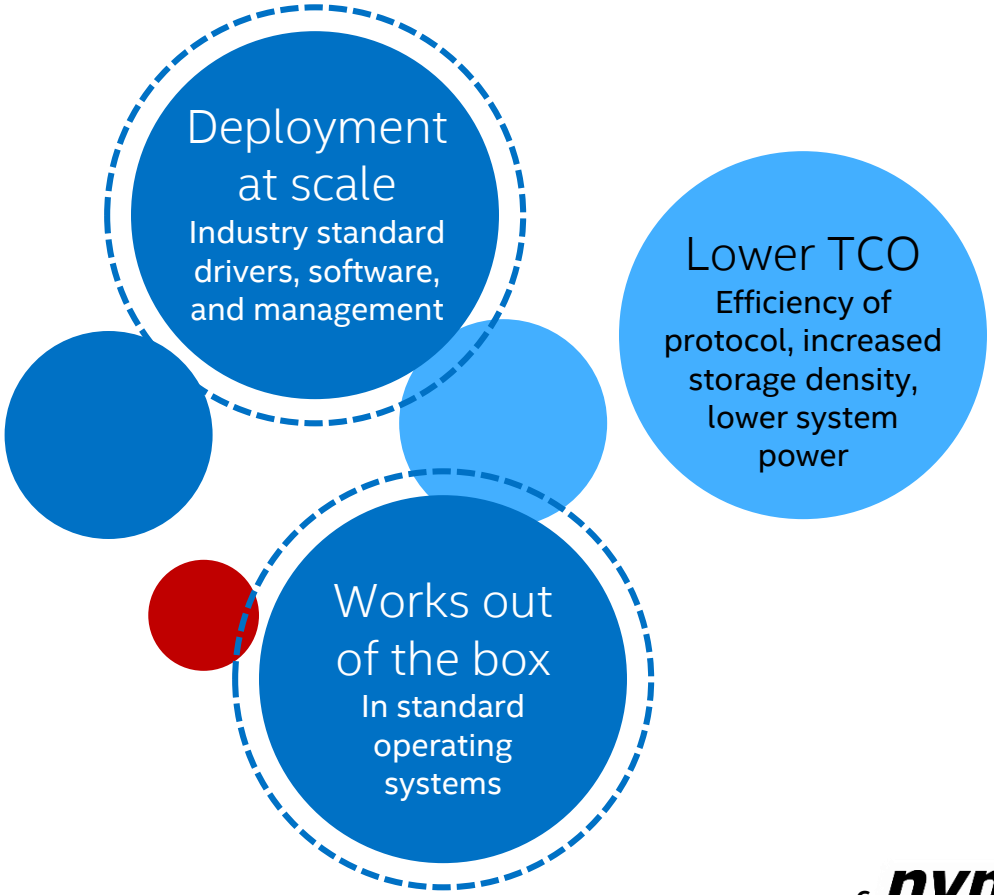
Queuing interface, NVMe I/O and Admin command set

Management Interface Workgroup

Out-of-band management over PCIe®
VDM and SMBus

CISCO DELL EMC² HGST a Western Digital company intel Micron
Microsoft NetApp ORACLE PMC SAMSUNG SanDisk Seagate

What NVM Express™ brings to the DATA CENTER



NVM Express™(NVMe™) Advantages over SATA™

PCIe® for **scalable** performance, **flexible** form factors, and industry **stability**

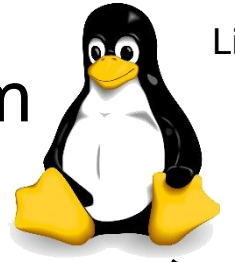
NVMe provides lower **latency** and increased **efficiency**:
lower CPU utilization, lower power, lower TCO

Increased **bandwidth**: 1 GB/s per lane – 1-16 lanes per drive
Directly attached to CPU, eliminate HBA cost and overhead

Low power features from both PCIe and NVMe
Security from Trusted Computing Group OPAL

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance.
Configurations: Intel® S2600CP server, Intel® Xeon® E5-2690v2 x2, 64GB DDR3, Intel® SSD DC P3700 Series 400GB, LSI 9207-8i, Intel® SSD DC S3700, HGST 6GBps SAS

NVMe™ Driver Ecosystem



Linux NVMe driver is open source

Windows Server 2012 R2 Certified

Windows 8 Compatible

Windows 8.1

redhat 6.5 | 7.0

suse SLES 11 SP3 SLES 12

ubuntu 13 | 14

solaris

FreeBSD

Native / in-box

Windows Server 2012 Certified

CERTIFIED FOR Windows Server 2008 R2

Windows 7

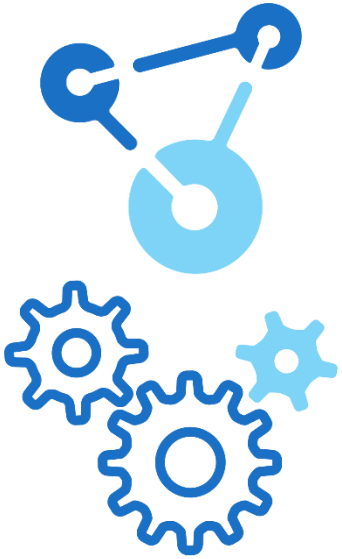
vmware ESXi 5.5

Install NVMe driver

*Other names and brands may be claimed as the property of others.

nvm EXPRESS™

What do I need to start using an NVM Express™ SSD?



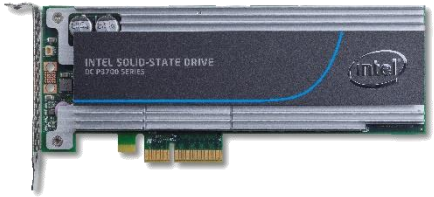
✓ Software: NVMe™ driver

✓ Hardware: PCIe® infrastructure

NVMe sits on top of PCIe

Form Factors for PCI Express®

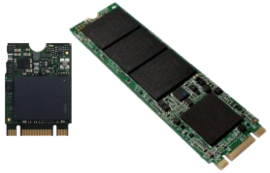
AIC

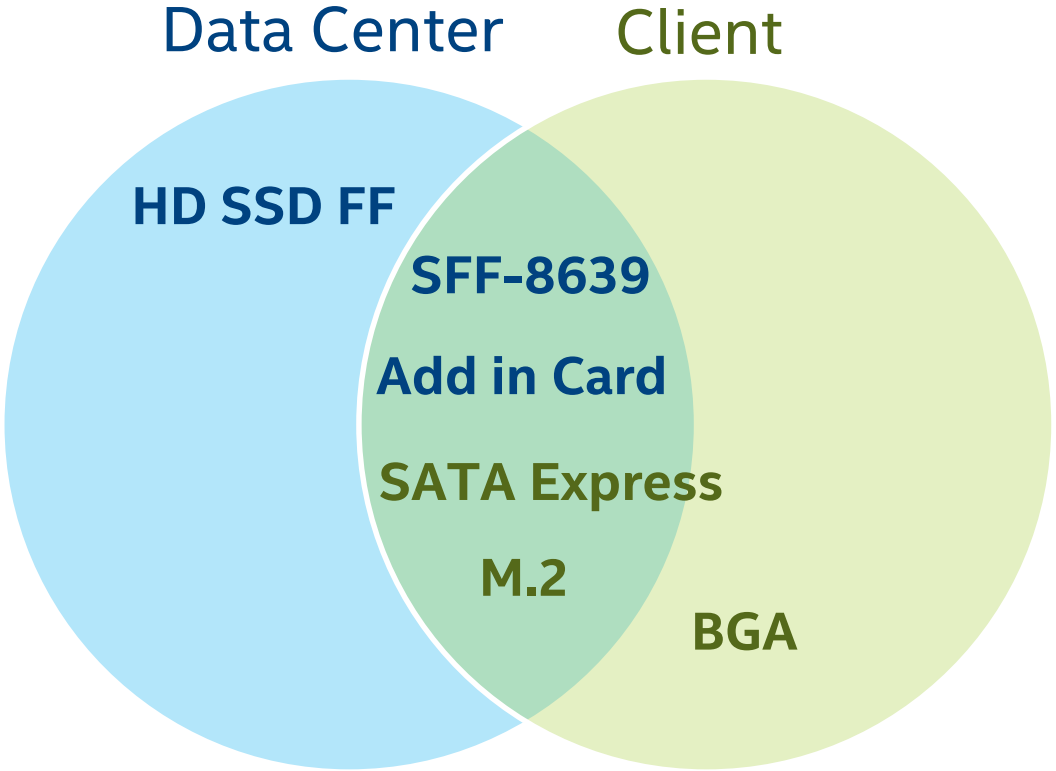


2.5in
SFF-8639

SATA Express™

M.2

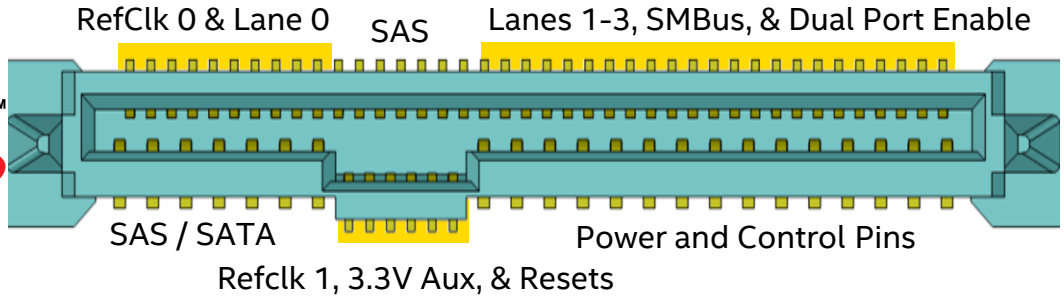




Drive Connectors

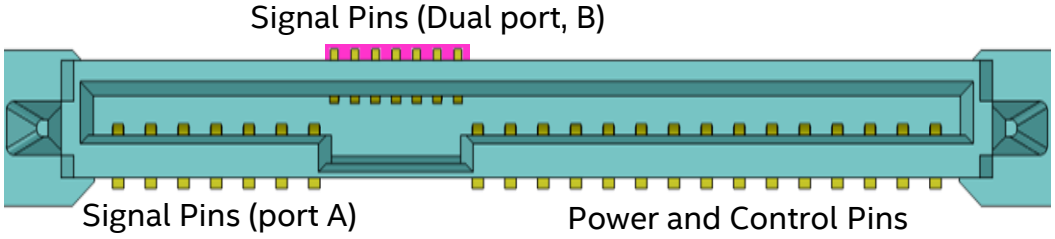
SFF-8639

- Supports SATA, SAS, and PCIe® x4 or two x2
- PCIe data, reference clock, and side band



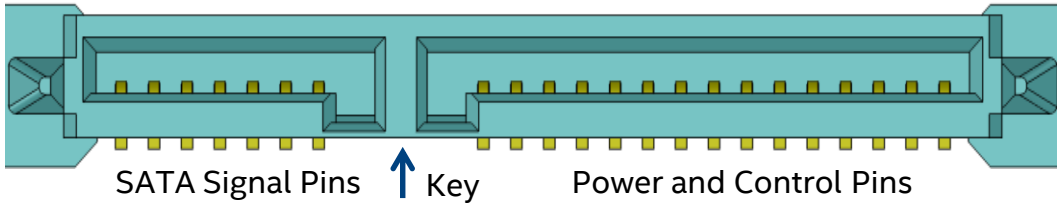
SAS®

- Backwards compatible with SATA
- Dual port

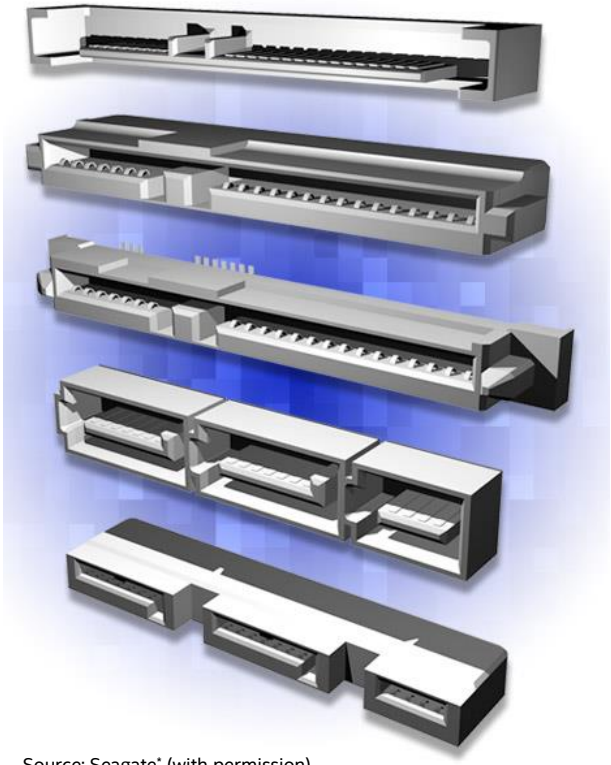


SATA™

- Keyed only for SATA drives
- Separate power and data



SATA Express™ and SFF-8639 Comparison



Source: Seagate* (with permission)

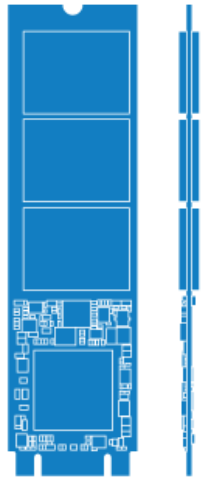
	SATAe	SFF-8639
SATA / SAS®	SATA	SATA / SAS
PCI Express®	x2	x4 or dual x2
Host Mux	Yes	No
Ref Clock	Optional	Required
EMI	SRIS	Shielding
Height	7mm	15mm
Max Performance	2 GB/s	4 GB/s
Bottom Line	Flexibility & Cost	Performance

SFF-8639 designed for data center, SATAe designed for Client

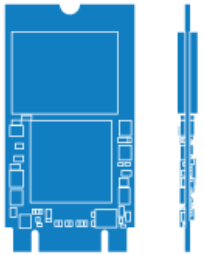
M.2 Form Factor Comparison

	M.2 Socket 2	M.2 Socket 3
SATA	Yes, Shared	Yes, Shared
PCIe® x2		
PCIe x4	No	Yes
Comms Support	Yes	No
Ref Clock	Required	Required
Max Performance	2 GB/s	4 GB/s
Bottom Line	Flexibility	Performance

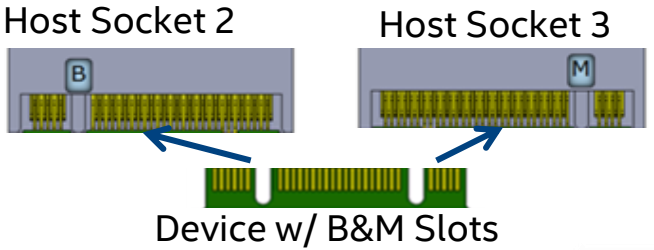
M.2 Socket 3 is the best option for Data Center PCIe SSDs



M.2 2280
22 x 80 x 3.58 (mm)

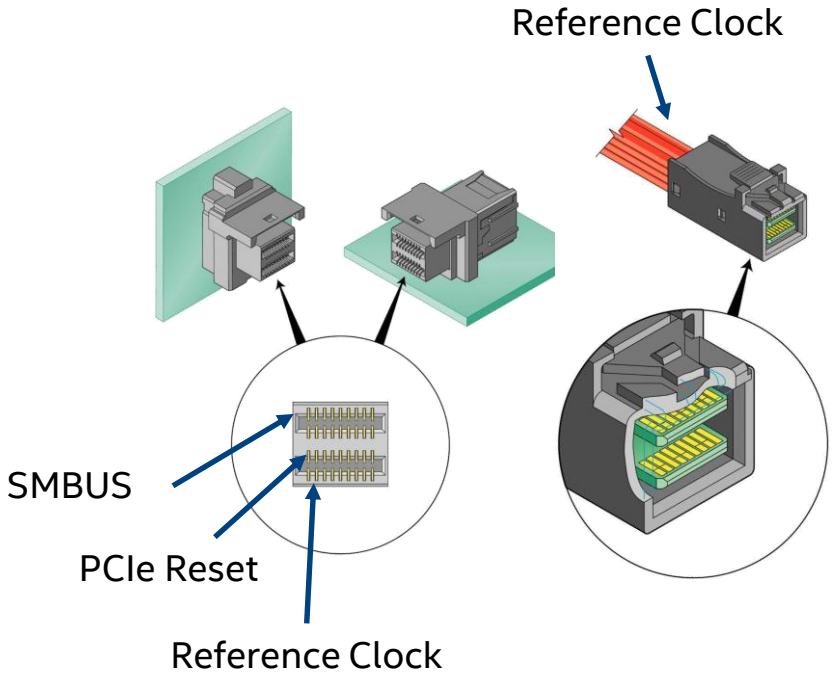


M.2 2242
22 x 42 x 3.58 (mm)



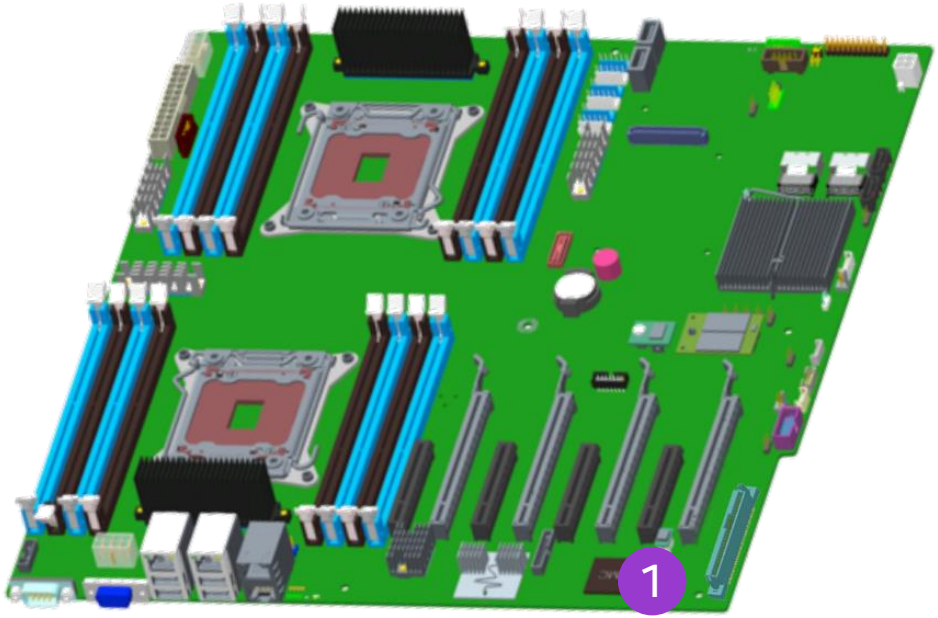
Cabling Options for Data Center PCIe® SSD Topologies

miniSAS HD cables lightly modified for PCIe are being used due to the robust connector and high volume manufacturing.

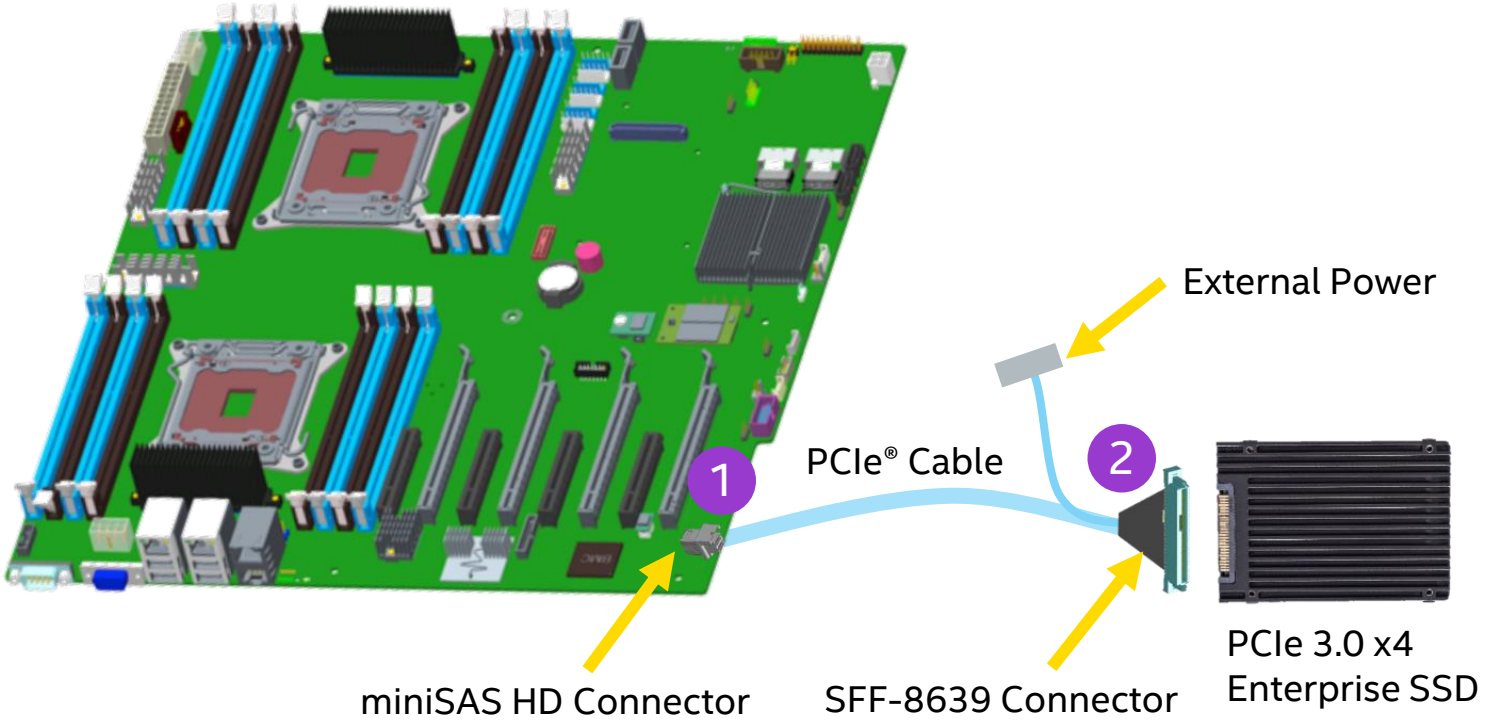


Basic PCI Express® SSD Topology – 1 Connector

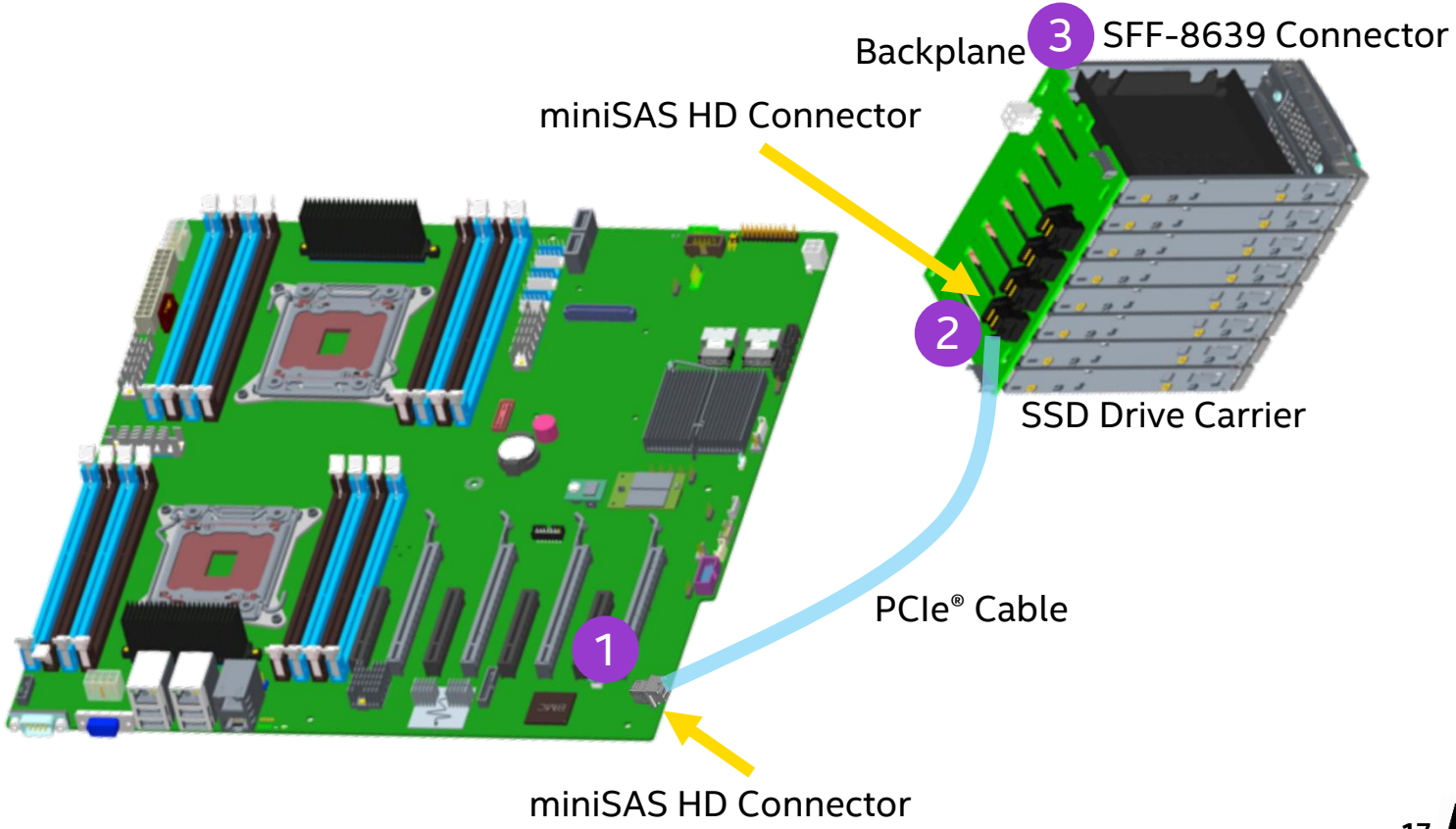
- SFF-8639 Connector directly attached to board
- Mostly used in small form factors such as compute node, blade, etc.



Basic PCI Express® SSD Topology – 2 Connector

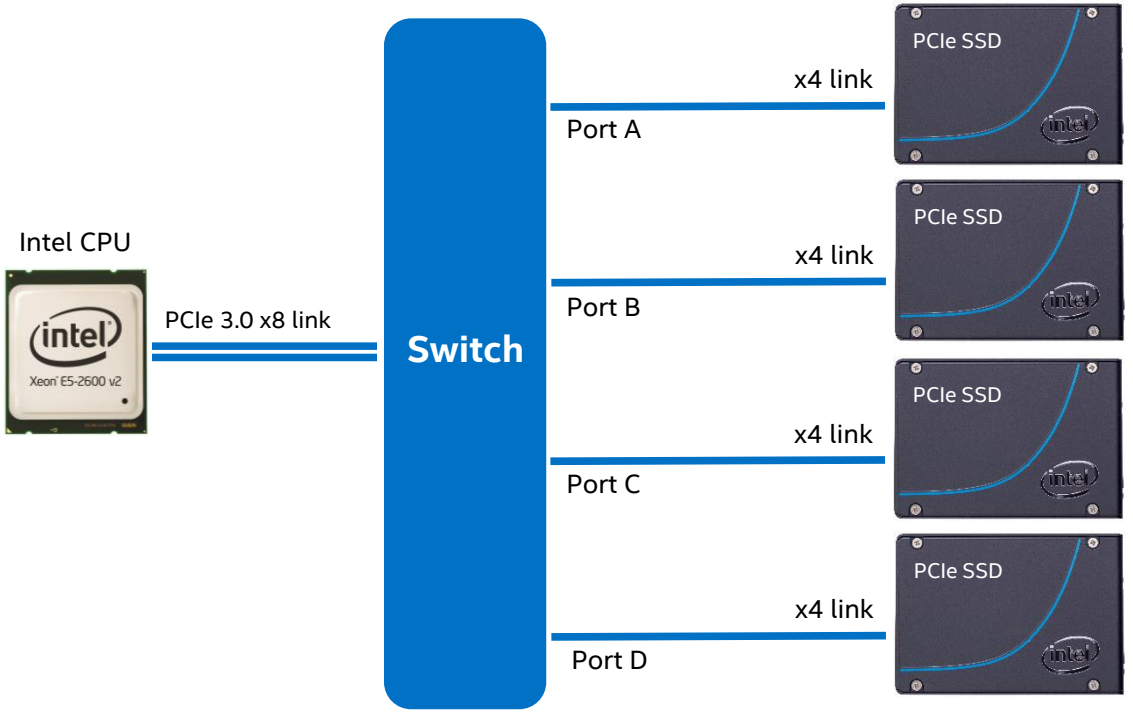


Basic PCI Express® SSD Topology – 3 Connector



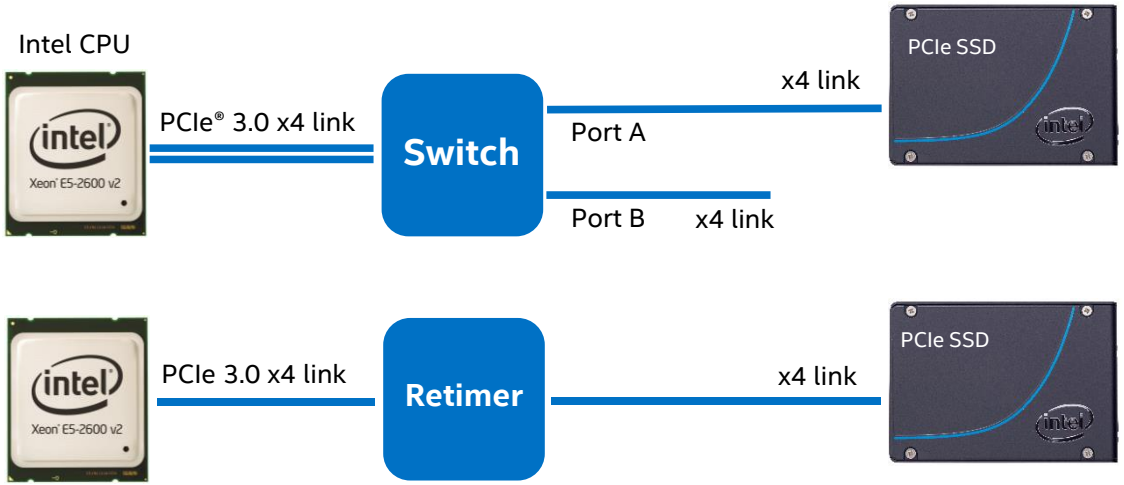
Port Expansion Devices - Switches

Use Switches to expand number of PCIe® SSDs



Link Extension Devices – Switches and Retimers

Use Link Extension Devices for longer topologies



PCI Express® (PCIe®) Switches and Retimers

PCIe Switches

- Use for link extension and/or port expansion
- Hot-plug and error isolation
- High performance peer-to-peer transfers
- Extra software features

Retimers

- Mostly transparent to software
- Retimers should be more common in PCIe 4.0

Recommend using only switches or retimers for link extension of PCIe

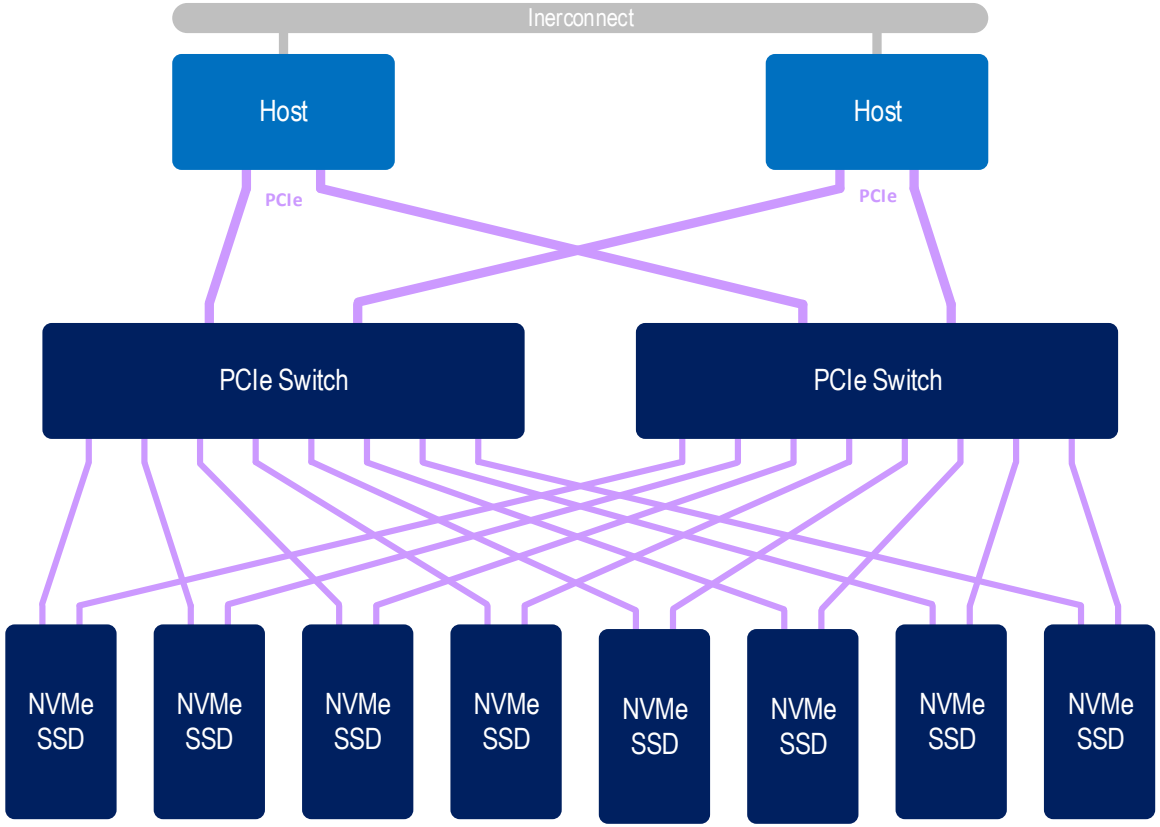
Link Extension Devices

- Use when channel has > **-20db loss**: at 8GT/s PCIe 3.0

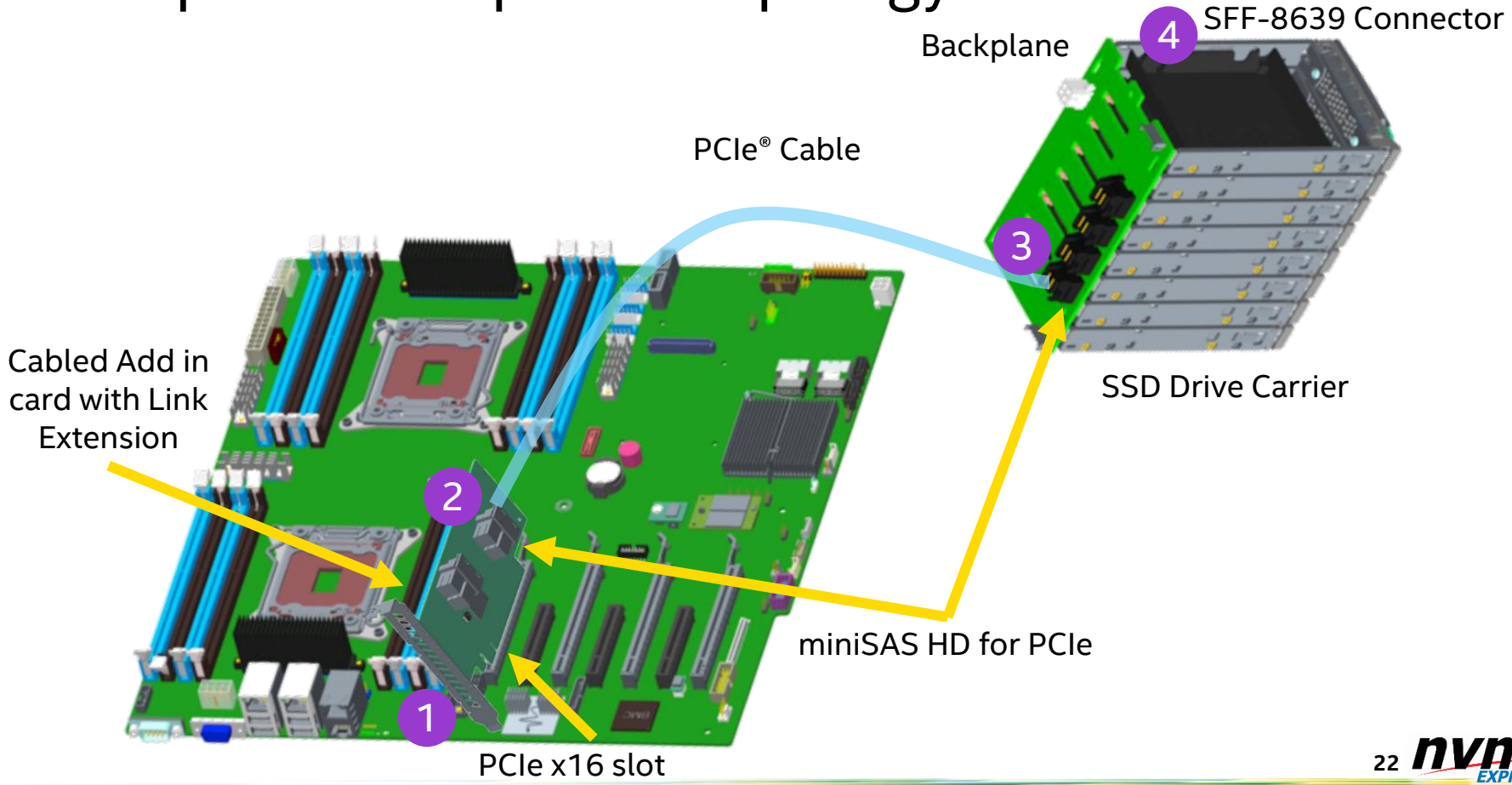
Retimer vs. Re-driver

- **Repeater**: A Retimer or a Re-driver
- **Re-driver**: Analog and not protocol aware
- ✓ **Retimer**: Physical Layer protocol aware, software transparent, Extension Device. Forms two separate electrical sub-links.
  - Executes equalization procedure on each sub-link

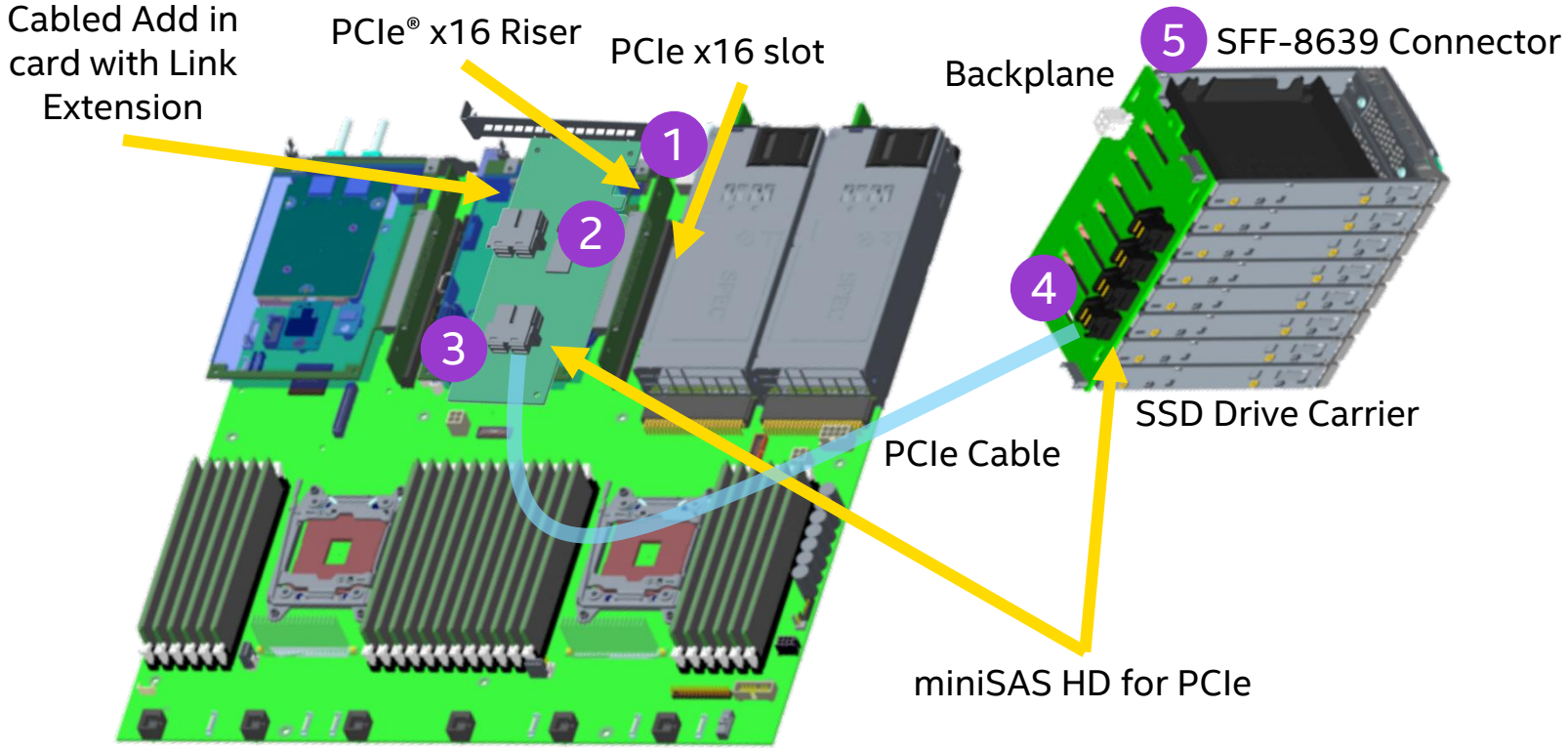
High Function Switches



Complex PCI Express® Topology – 4 Connector

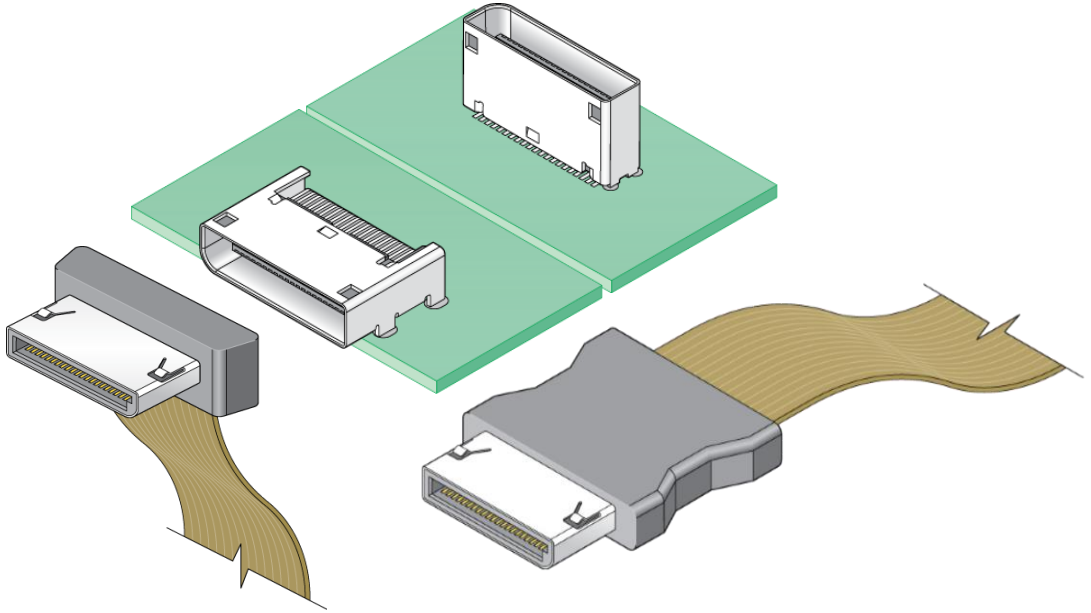


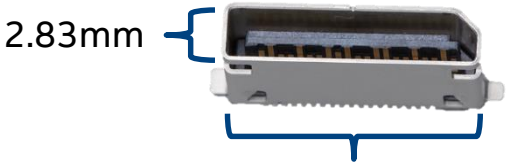
Complex PCI Express® Topology – 5 Connector



PCI Express® cabling for future topologies - OCuLink*

Category	OCuLink
Standard Based	PCI-SIG®
PCIe® Lanes	X4
Layout	Smaller footprint
Signal Integrity	Similar on loss dominated channels
PCIe 4.0 ready	16GT/s target
Clock, power	Supports SRIS and 3.3/5V power
Production Availability	Mid 2015

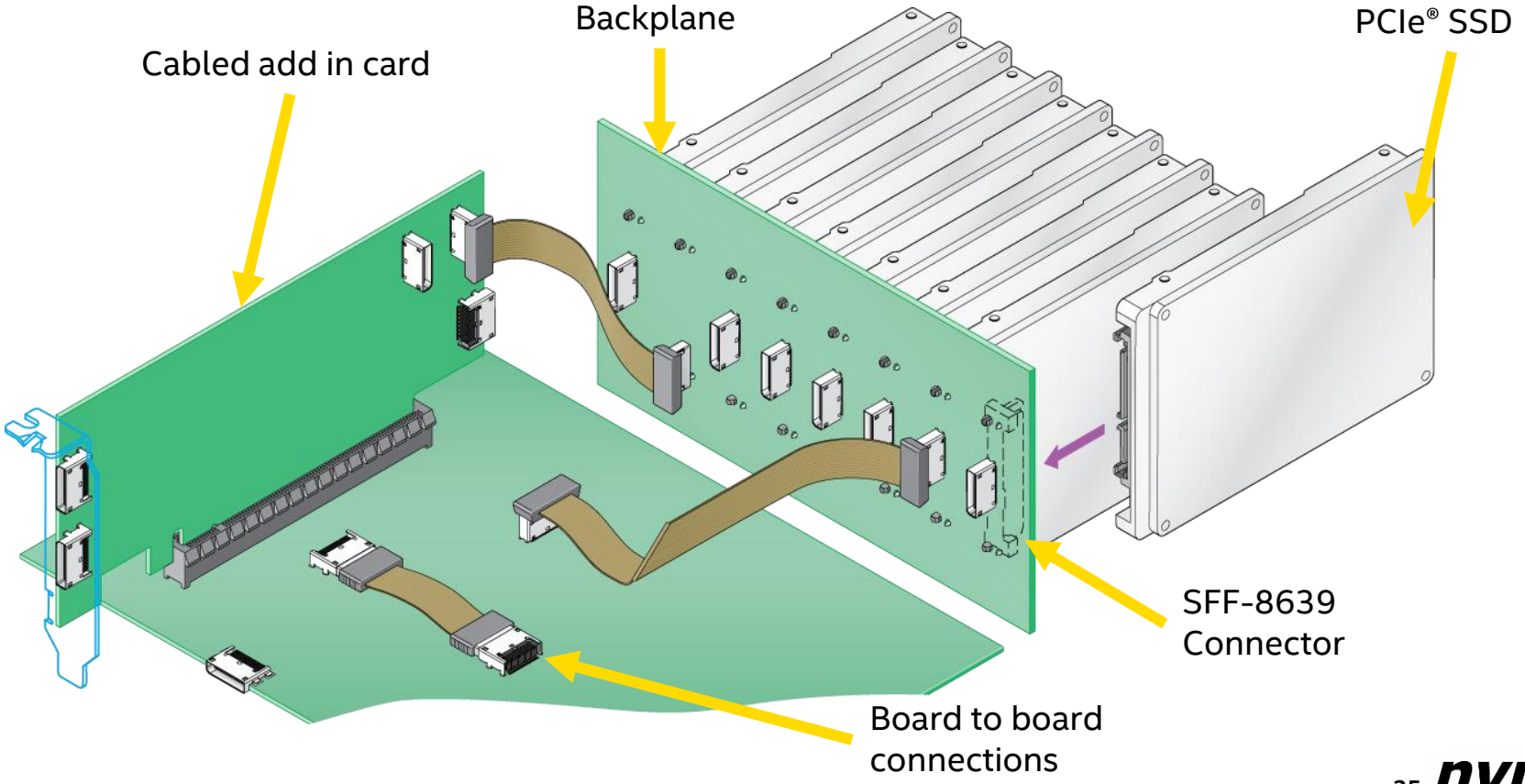




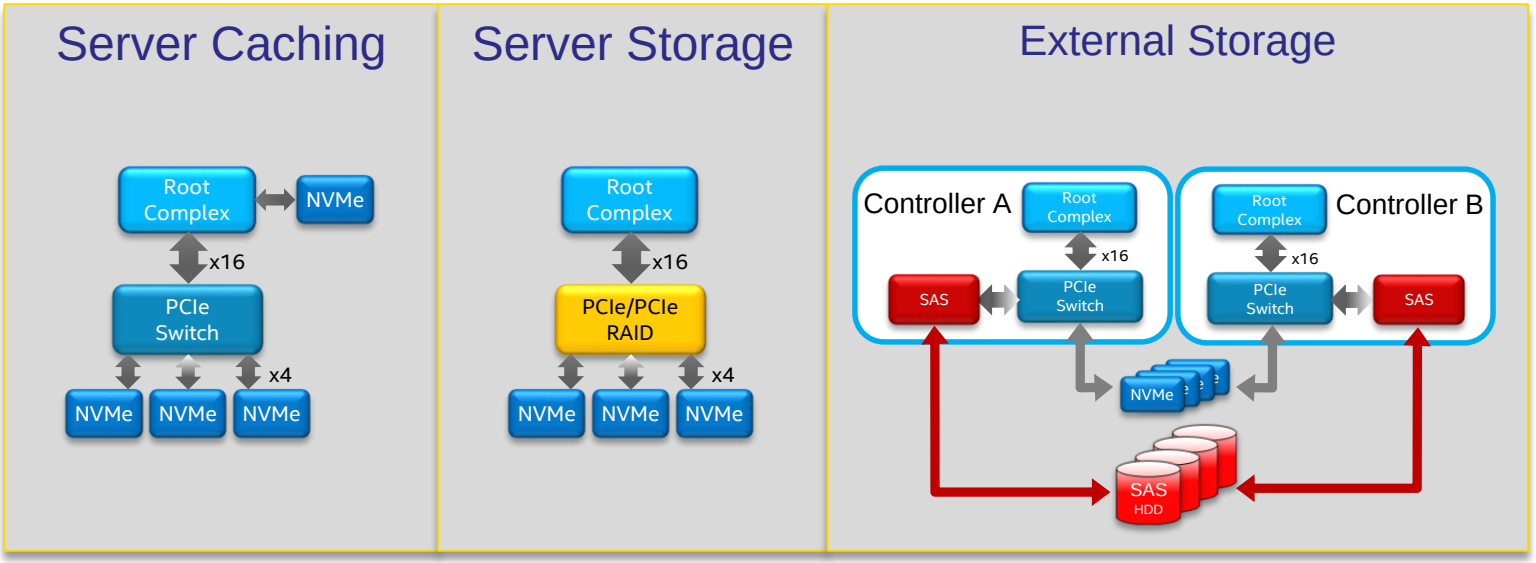
OCuLink internal cables and connectors

Source: molex one company › a world of innovation MSX Link

OCuLink* Provides Flexible Data Center Topologies



NVMe™ Storage Device Management



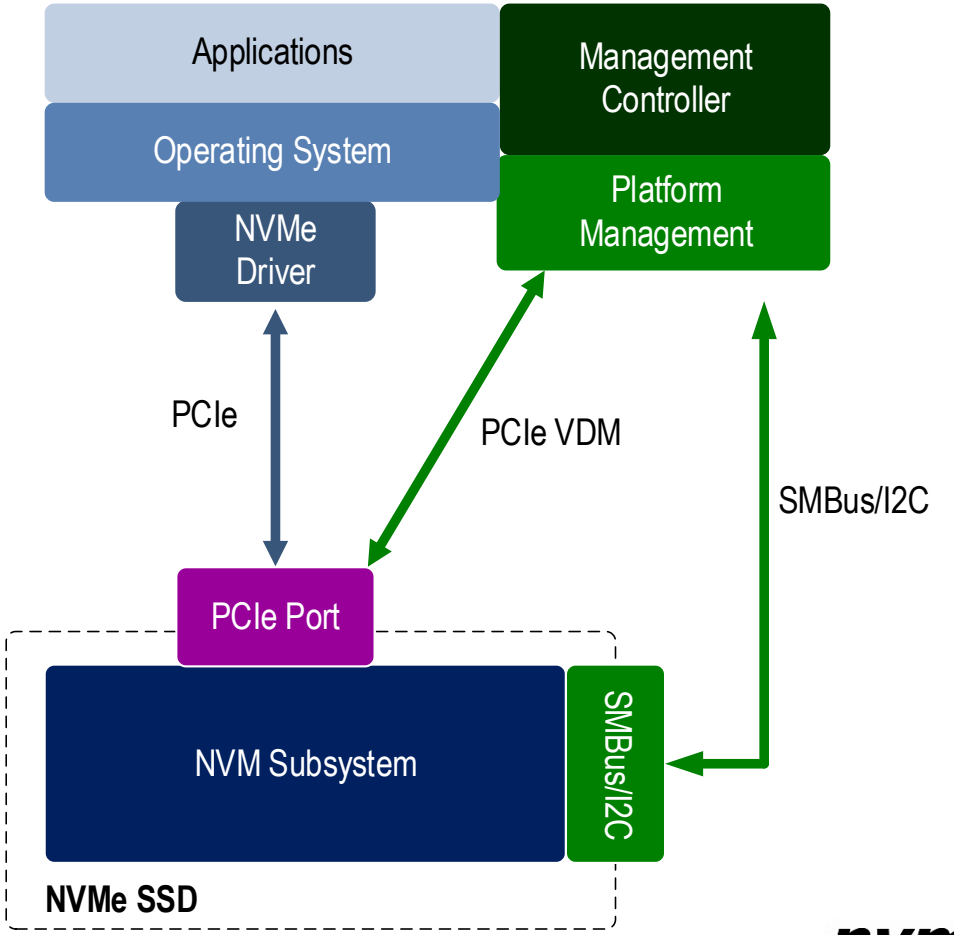
Example Pre-boot Management

- Inventory, Power Budgeting, Configuration, Firmware Update

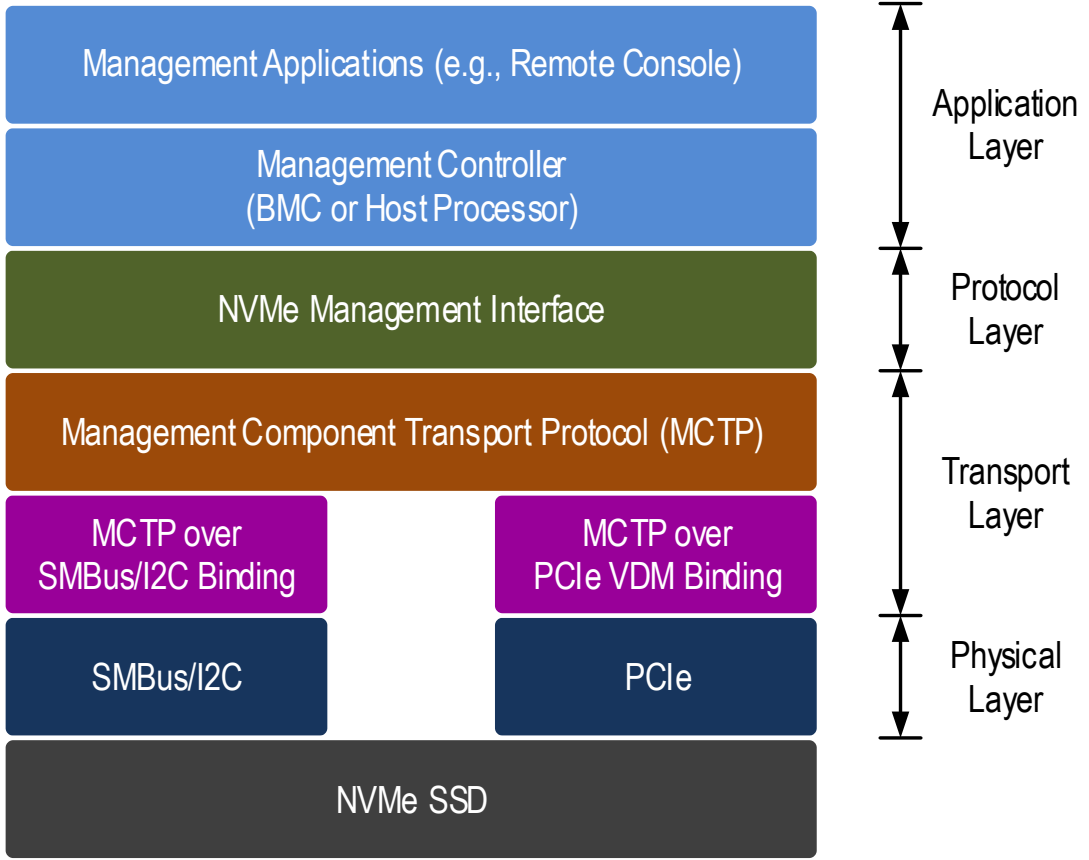
Example Out-of-Band Management During System Operation

- Health Monitoring, Power/Thermal Management, Firmware Update, Configuration

Driver vs. Out-of-Band Management



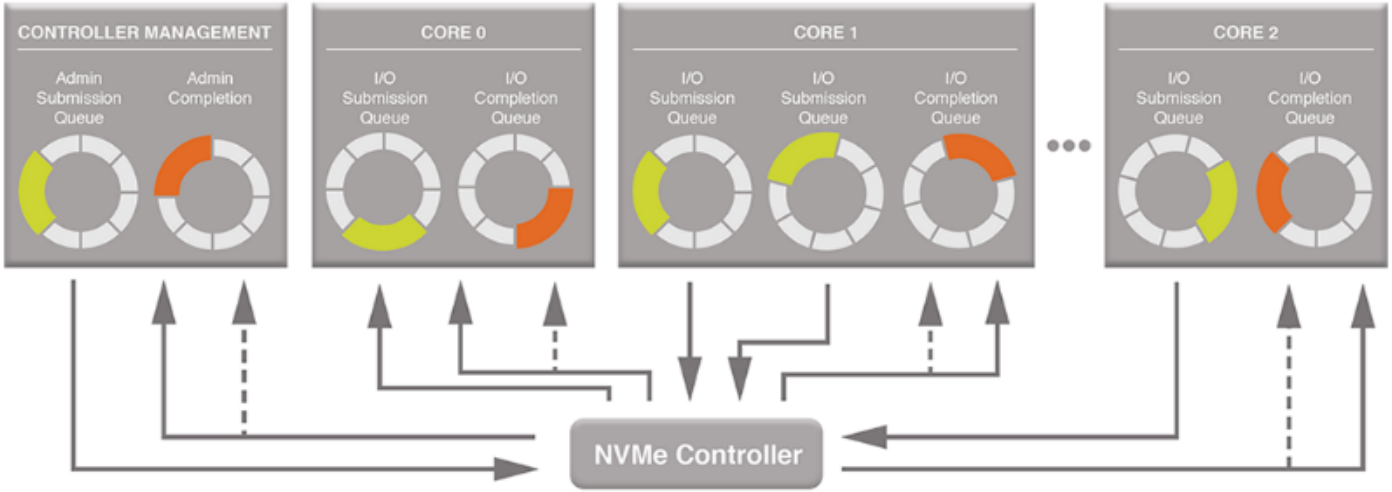
Management Interface Protocol Layering



Questions?

NVMe™ Technical Overview

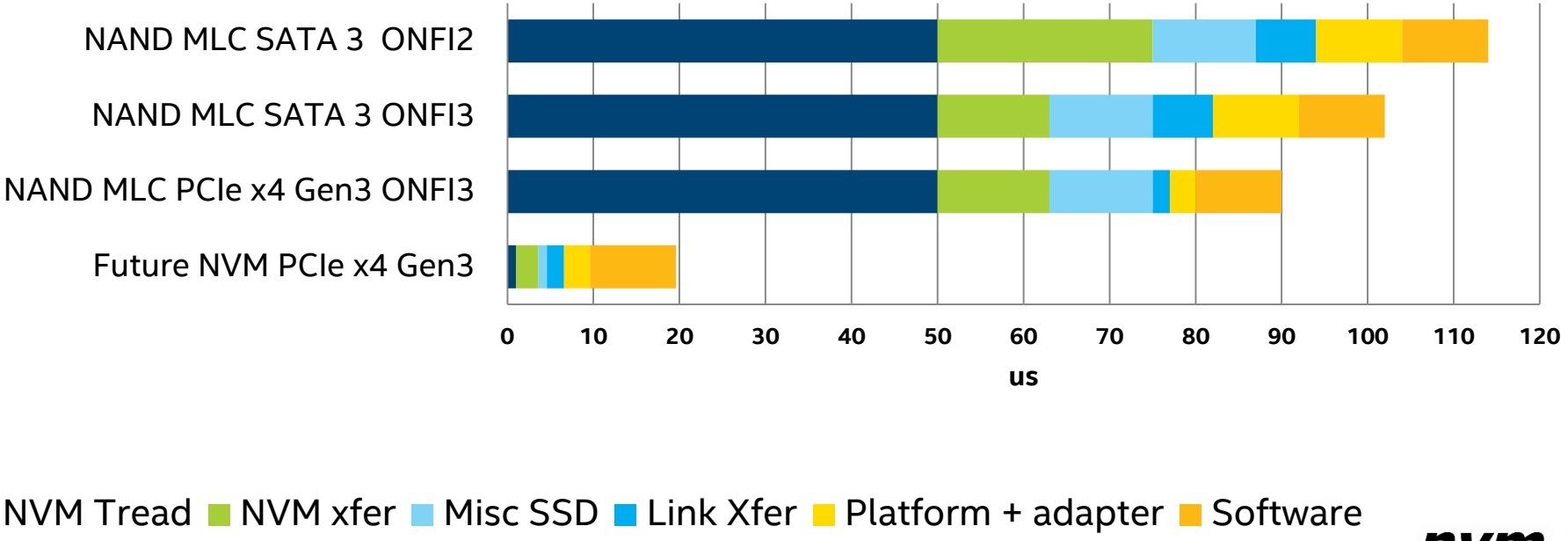
- Supports deep queues of 64K commands per queue, up to 64K queues
- Supports MSI-X and interrupt steering, enables even performance scaling
- Streamlined & simple command set (13 required commands), optional features to address target segments
- Built for the future, ready for next gen NVM



Fully Exploiting Next Gen NVM

With Next Gen NVM, the NVM is no longer the bottleneck

App to SSD read latency for 4KB transfer at Queue Depth of 1



NVMe™ Development History

NVMe 1.0 – Mar 2011

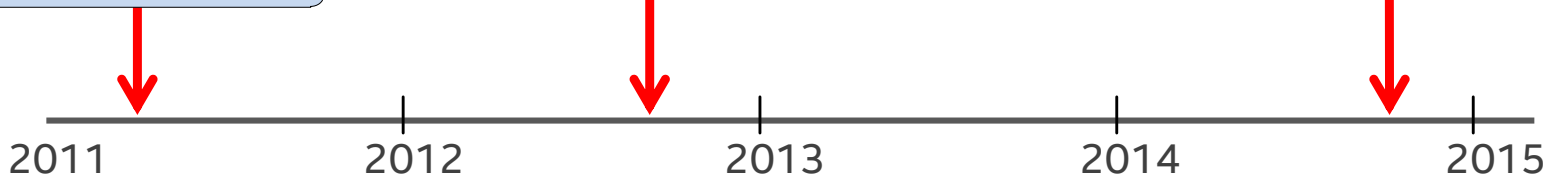
- Queuing Interface
- Command Set
- End-to-End Protection
- Security
- PRPs

NVMe 1.1 – Oct 2012

- Multi-Path IO
- Namespace Sharing
- Reservations
- Autonomous Power Transition
- Scatter Gather Lists

NVMe 1.2 – Q4 2014

- Host Memory Buffer
- Replay Protected Area
- Active/Idle Power and RTD3
- Temperature Thresholds
- Namespace Management
- Controller Memory Buffer
- Live Firmware Update
- Atomicity Enhancements



Architected for Performance