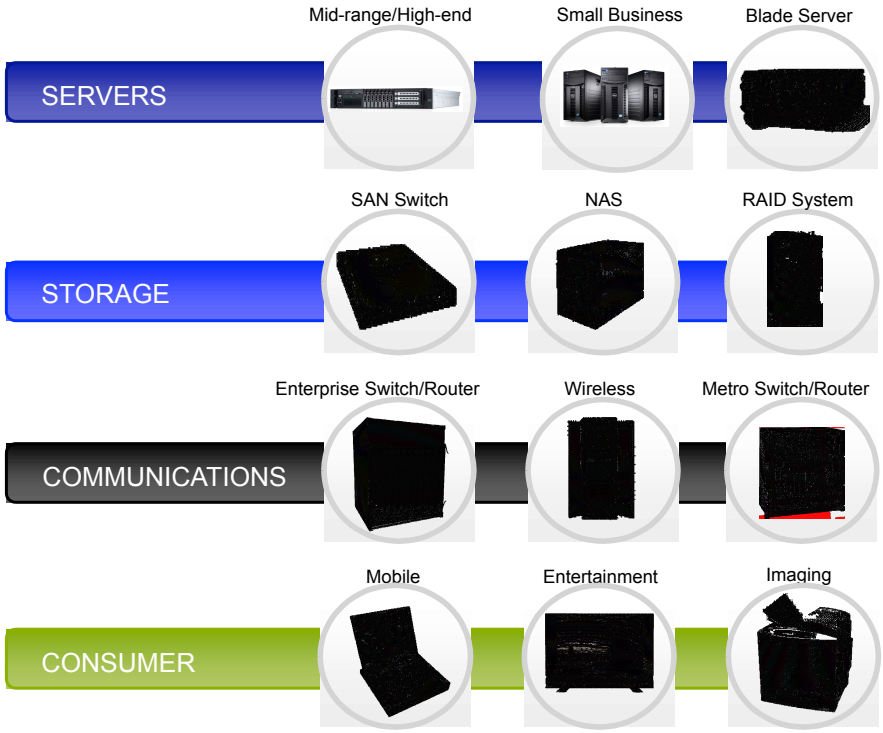
NVM
PCIe® Networked Flash Storage

Peter Onufryk
Microsemi Corporation

PCI Express (PCIe)

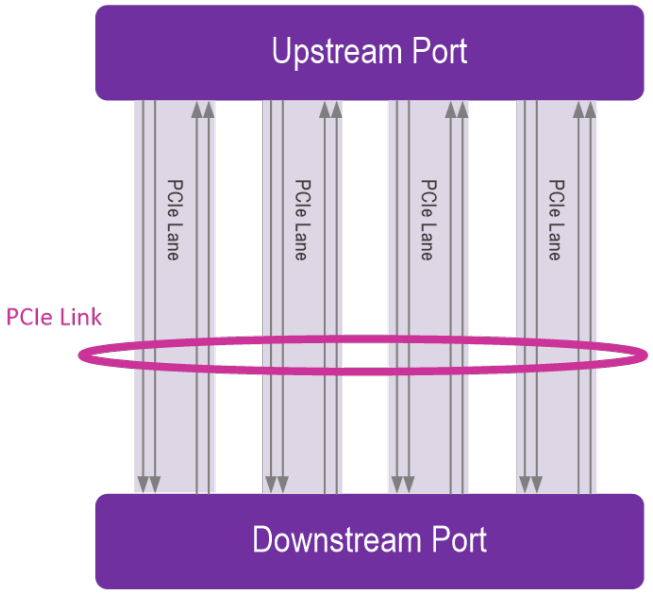
- Specification defined by PCI-SIG
  - www.pcisig.com
- Packet-based protocol over serial links
  - Software compatible with PCI and PCI-X
  - Reliable, in-order packet transfer
- High performance and scalable from consumer to Enterprise
  - Scalable link speed (2.5 GT/s, 5.0 GT/s, 8.0 GT/s, 16 GT/s, and 32 GT/s)
    - Gen5 (32 GT/s) is still being standardized
  - Scalable link width (x1, x2, x4, …, x32)
- Primary application is as an I/O interconnect



PCIe Characteristics

- Scalable speed
  - Encoding
    - 8b10b: 2.5 GT/s (Gen 1) and 5 GT/s (Gen 2)
    - 128b/130b: 8 GT/s (Gen 3), 16 GT/s (Gen4) and 32 GT/s (Gen5)
- Scalable width: x1, x2, x4, x8, x12, x16, x32

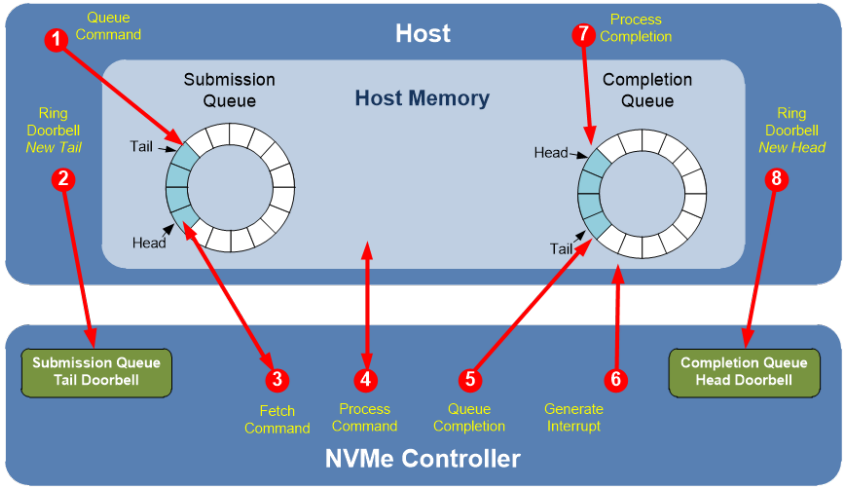
Generation	Raw Bit Rate	Bandwidth Per Lane Each Direction	Total x16 Link Bandwidth
Gen 1*	2.5 GT/s	~ 250 MB/s	~ 8 GB/s
Gen 2*	5.0 GT/s	~500 MB/s	~16 GB/s
Gen 3*	8 GT/s	~ 1 GB/s	~ 32 GB/s
Gen 4	16 GT/s	~ 2 GB/s	~ 64 GB/s
Gen 5	32 GT/s	~4 GB/s	~128 GB/s



*Source – PCI-SIG PCI Express 3.0 FAQ

NVM Express™ (NVMe™)

- Two specifications
  1. NVM Express (PCIe)
  2. NVM Express over Fabrics (RDMA and Fibre Channel)
- Architected from the ground up for NVM
  - Simple optimized command set
  - Fixed size 64 B commands and 16 B completions
  - Supports many-core processors without locking
  - No practical limit on the number of outstanding requests
  - Supports out-of-order data deliver



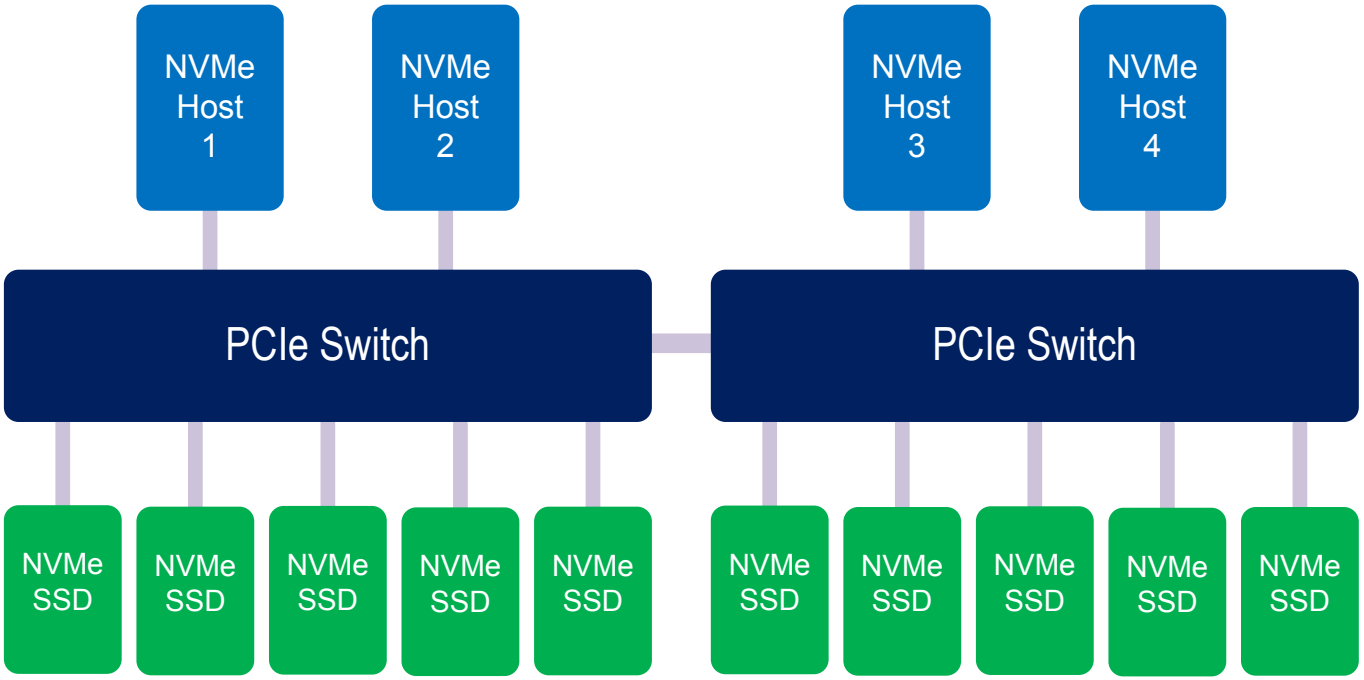
PCIe SSD = NVMe SSD

Ideal NVM Fabric

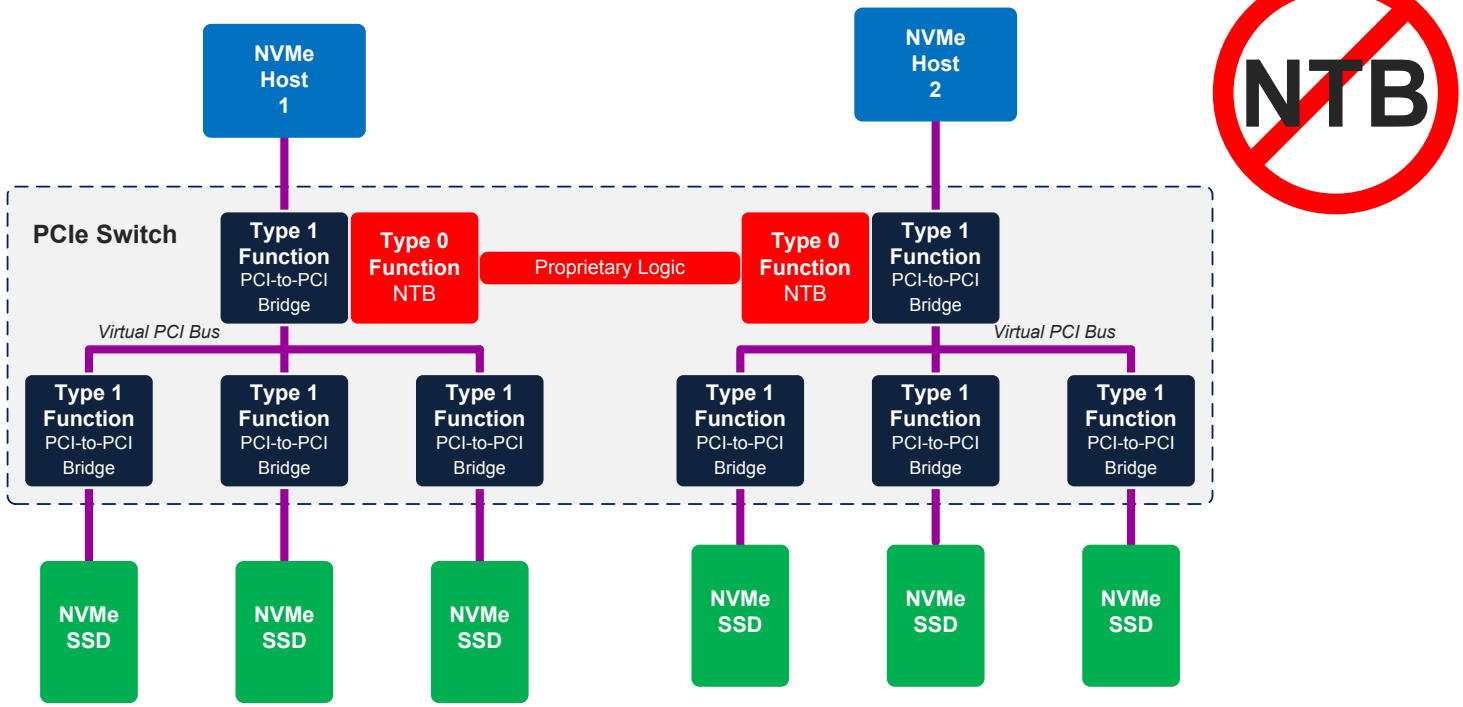
Property	Ideal Characteristic
Cost	Free
Complexity	None
Performance	High
Power consumption	None
Standards-based	Yes
Scalability	Infinite


Flash Memory Summit

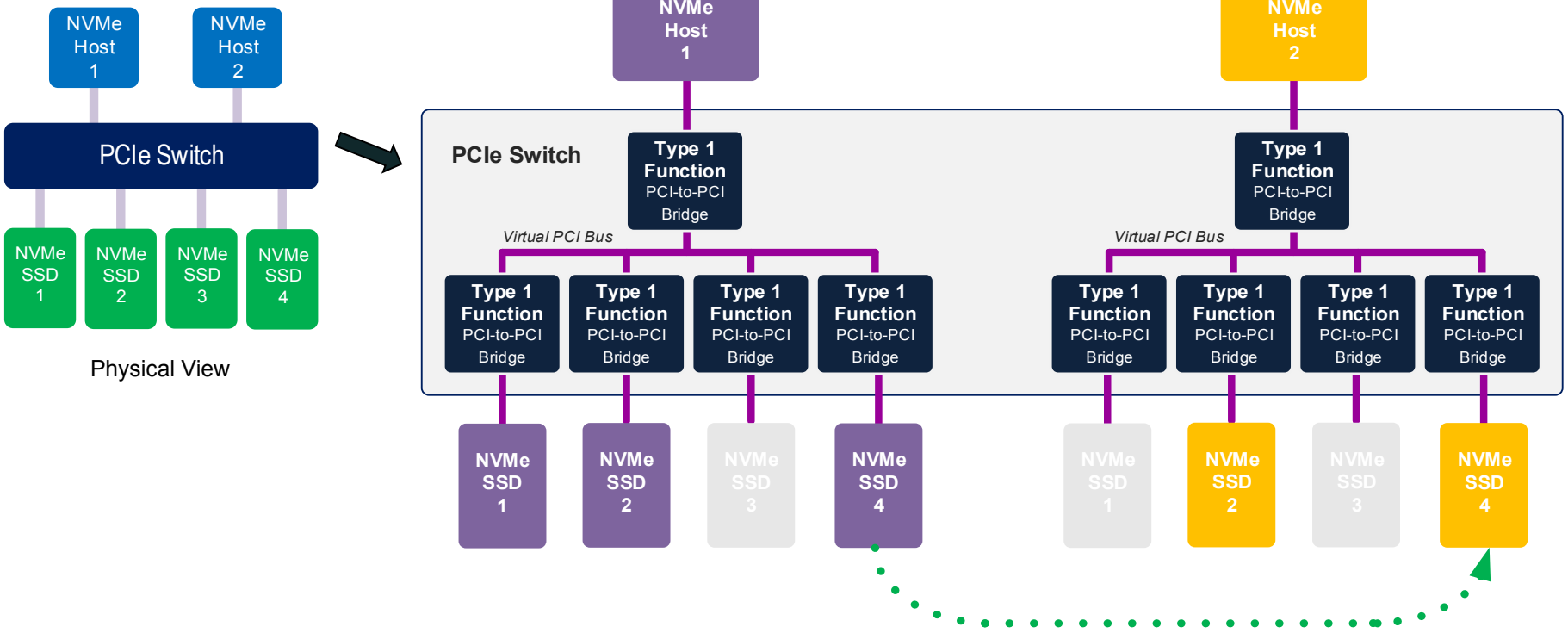
PCIe Fabric


NVMe Host 1
NVMe Host 2
NVMe Host 3
NVMe Host 4
PCIe Switch
PCIe Switch
NVMe SSD
NVMe SSD
NVMe SSD
NVMe SSD
NVMe SSD
NVMe SSD
NVMe SSD
NVMe SSD
NVMe SSD
NVMe SSD

Non-Transparent Bridging (NTB)

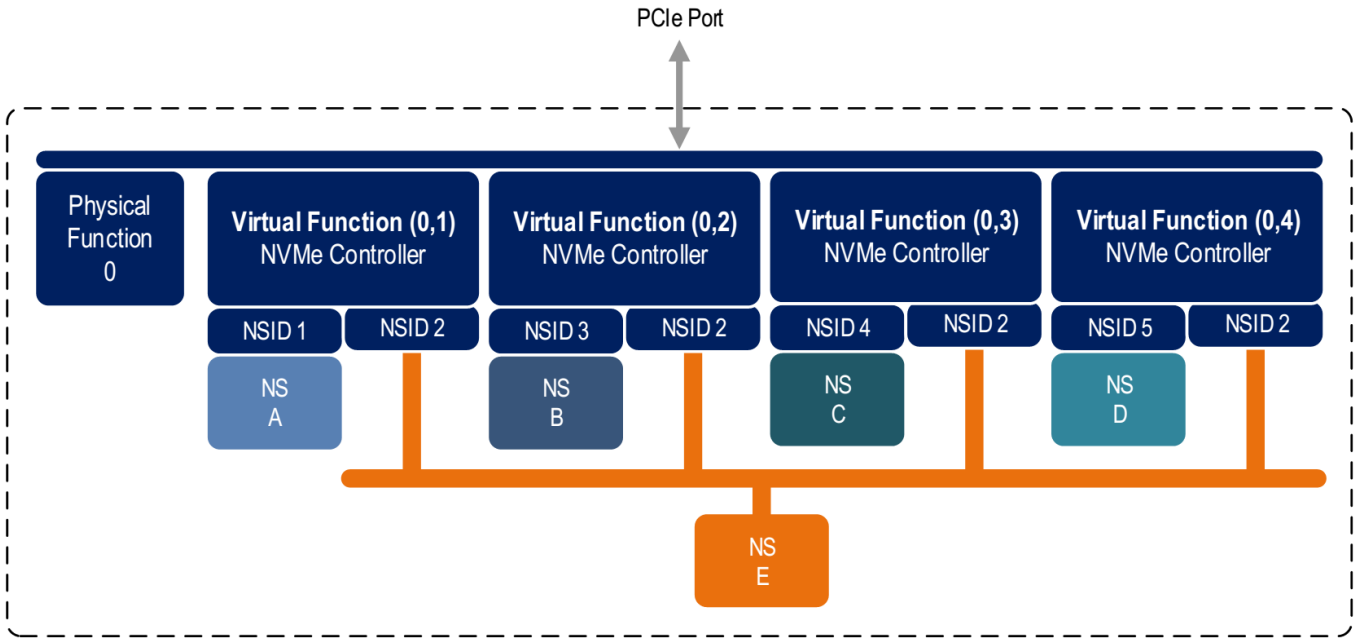


Dynamic Partitioning

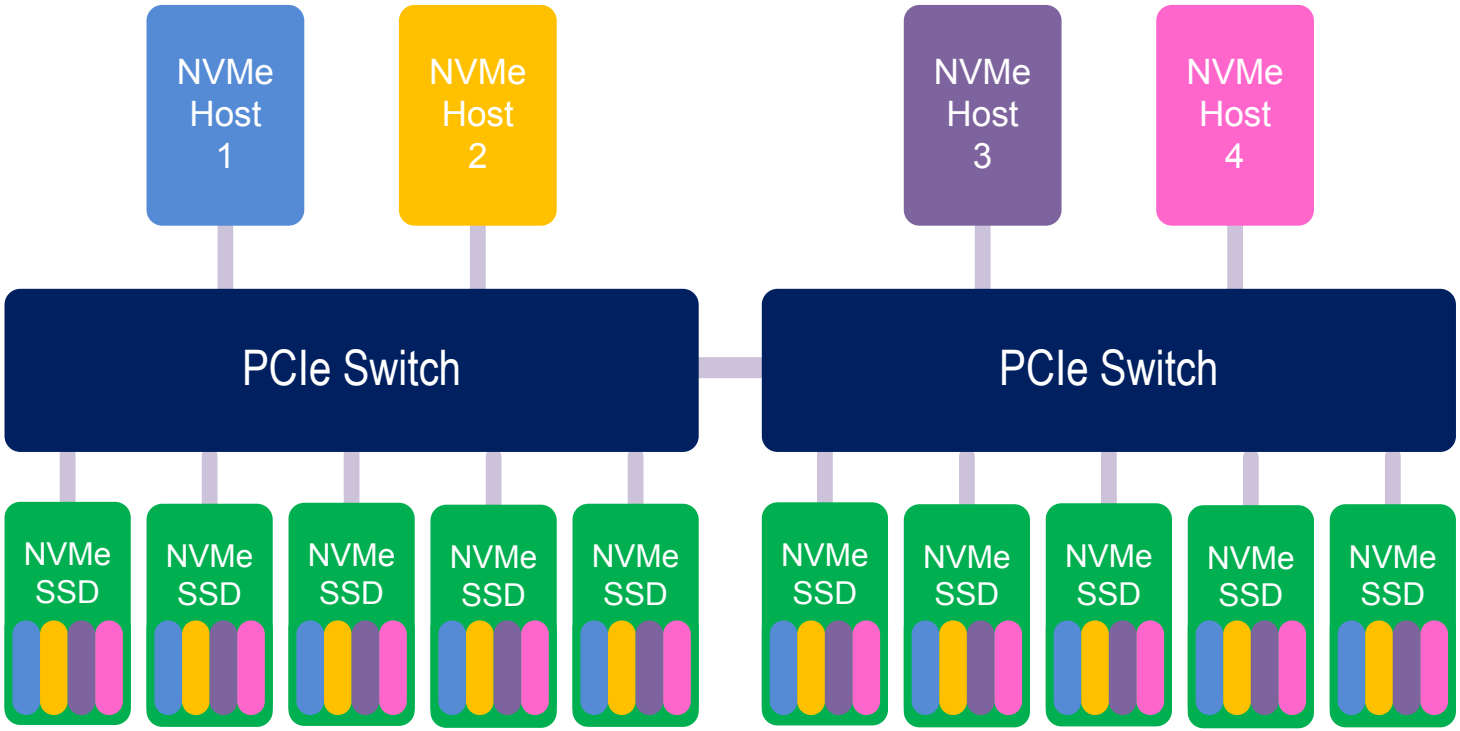



Flash Memory Summit

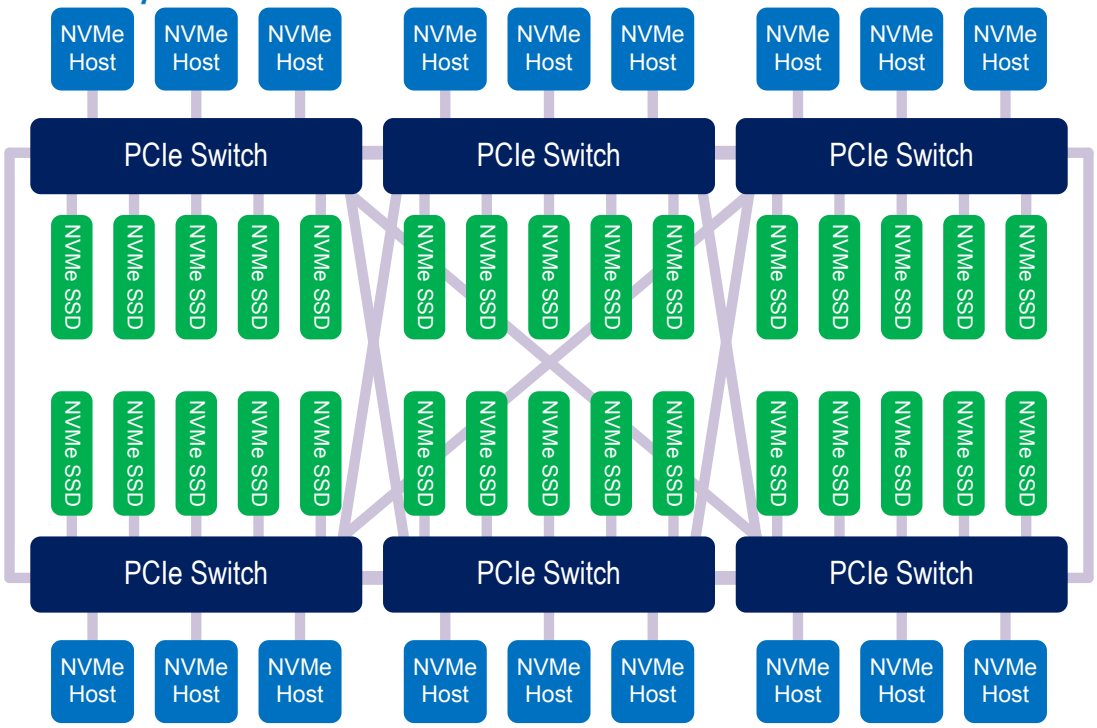
NVMe SR-IOV


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



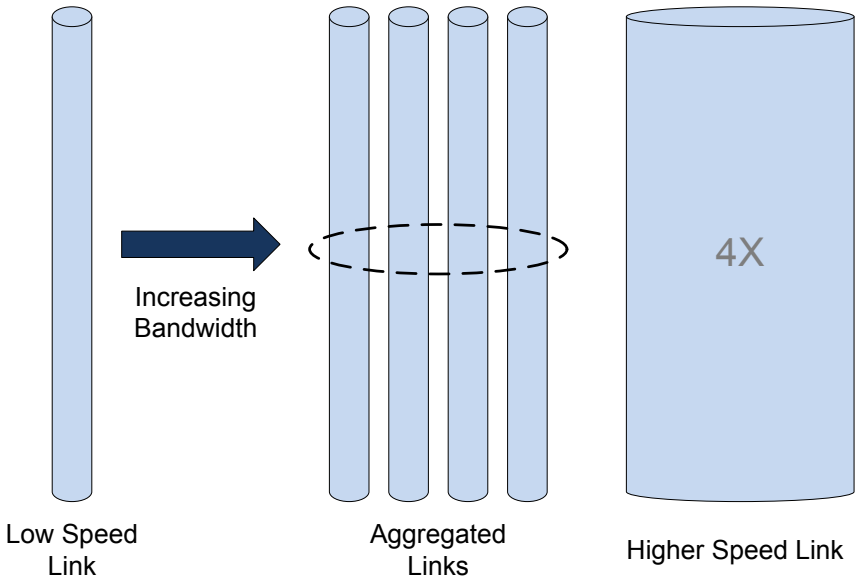
PCIe Fabric



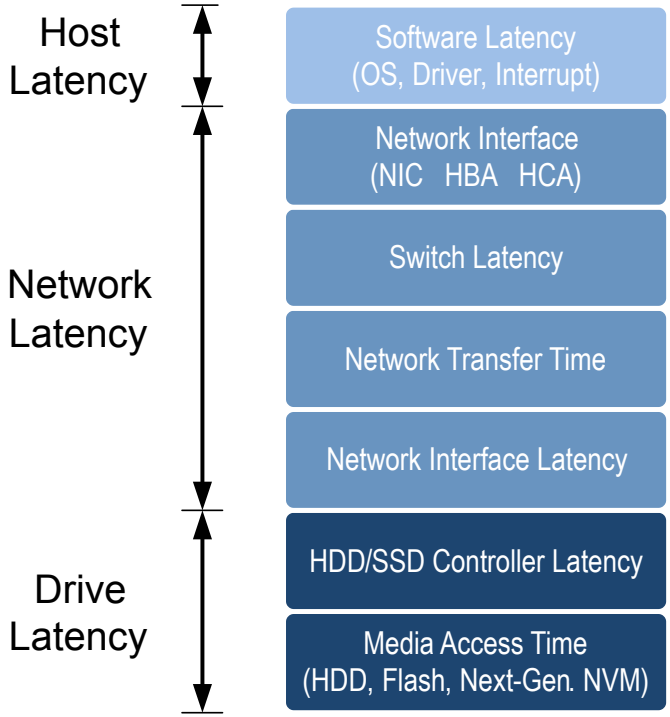
- Storage Functions
  - Dynamic partitioning (drive-to-host mapping)
  - NVMe shared I/O (shared storage)
  - Ability to share other storage (SAS/SATA)
- Host-to-Host Communications
  - RDMA
  - Ethernet emulation
- Manageability
  - NVMe controller-to-host mapping
  - PCIe path selection
  - NVMe management
- Fabric Resilience
  - Supports link failover
  - Supports fabric manager failover

Fabric Performance

- A high-performance fabric means
  - High bandwidth
  - Low latency
- Increasing bandwidth is easy
  - Aggregate parallel links
  - Increase link speed (fatter pipe)
- Reducing latency is hard
  - Transfer latency is typically a small component of overall latency
  - Other sources of latency:
    - Software (drivers)
    - Complex protocols
    - Protocol translation
    - Fabric switches/hops

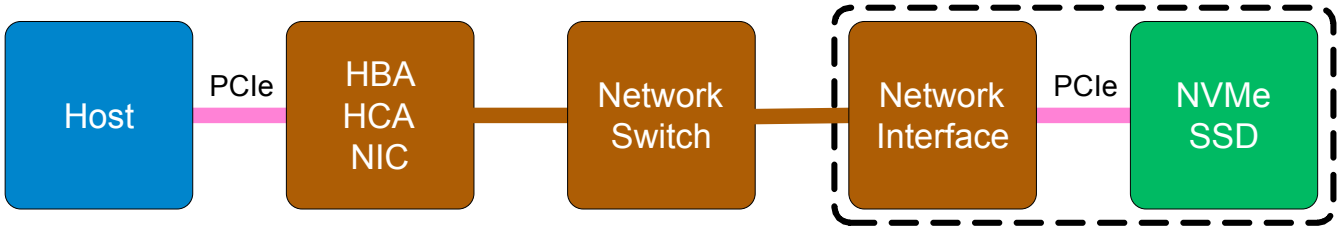


Latency



- Media Access Time
  - Hard drive: Milliseconds
  - NAND flash: Microseconds
  - Next-generation NVM: Nanoseconds

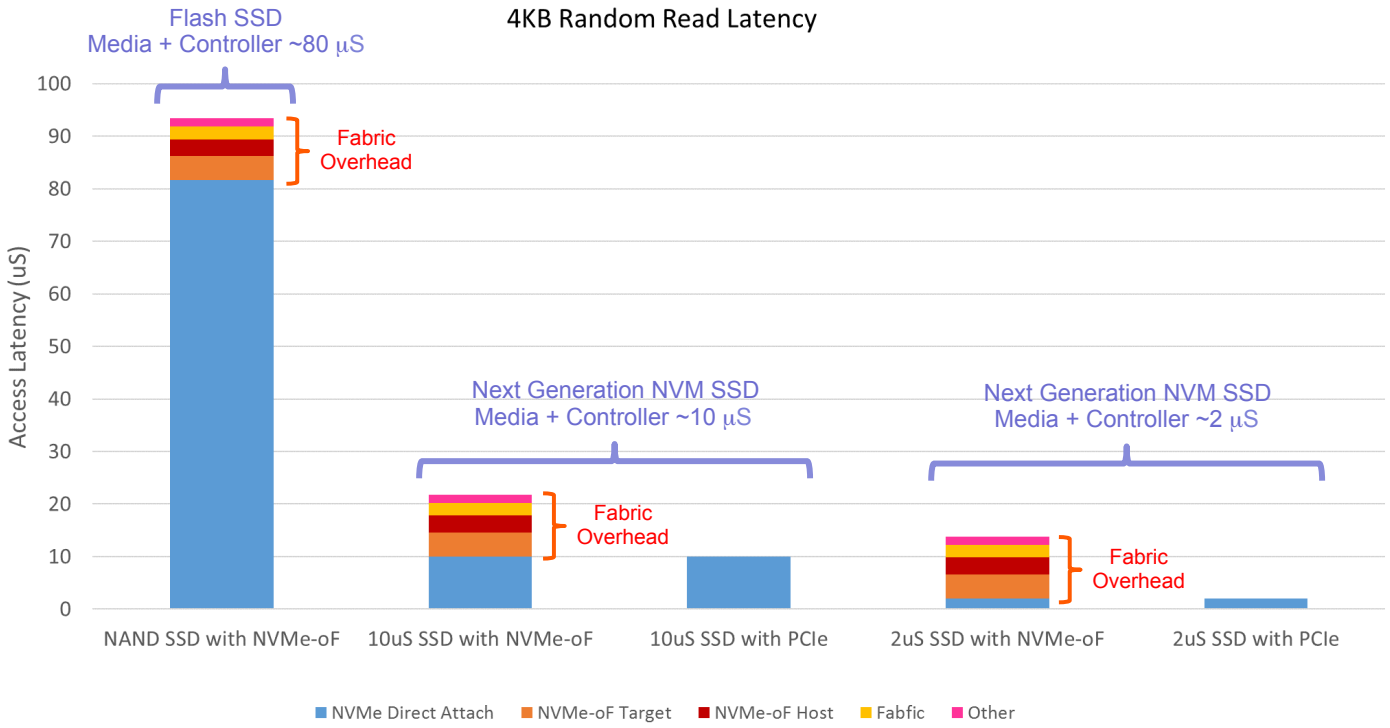
The PCIe Advantage



Other Flash Storage Networks

PCIe Fabric

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

PCIe Fabric Characteristics

Property	Ideal Characteristic	PCIe Fabric	Notes
Cost	Free	Low	• PCIe built into virtually all hosts and NVMe drives
Complexity	None	Medium Complexity	• Builds on existing NVMe ecosystem with no changes • PCIe fabrics are an emerging technology • Requires PCIe SR-IOV drives for low-latency shared storage
Performance	High	High	• High bandwidth • The absolute lowest latency
Power consumption	None	Low	• No protocol translation
Standards based	Yes	Yes	• Works with standard hosts and standard NVMe SSDs
Scalability	Infinite	Limited	• PCIe hierarchy domain limited to 256 bus numbers • PCIe has limited reach (cables) • PCIe fabrics have limited scalability (less than 256 SSDs and 128 hosts)

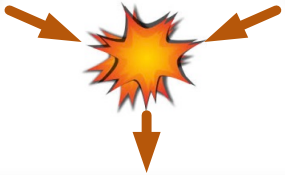
Persistent Memory and Next Gen. NVM

Traditional Memory

- Volatile
- Byte addressable
- Memory load/store operations
- Memory bus

Traditional Storage

- Non-volatile (persistent)
- Block, file, or object addressable
- I/O operations
- Storage interconnect



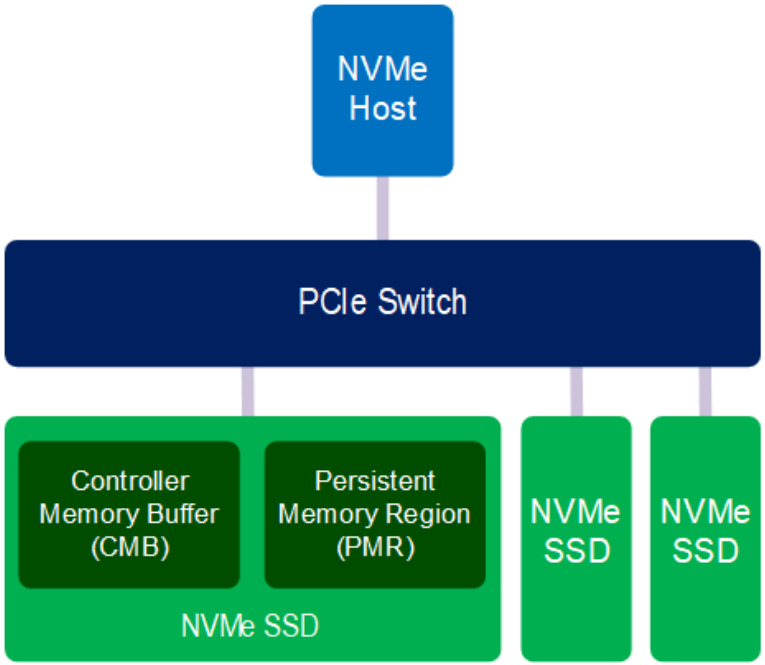
Next Generation NVM

- Non-volatile (persistent)
- Byte, block, file, or object addressable
- Memory load/store operations and I/O operations

Examples: phase-change memory (PCM), resistive RAM (RRAM), spin-transfer-torque magnetic RAM (STT_MRAM), ferroelectric RAM (fRAM)

NVMe and Memory Operations

- Controller Memory Buffer (CMB)
  - PCI memory space exposed to host (byte addressable)
  - May be used to store commands and data
  - Contents do not persist across power cycles and resets
- Persistent Memory Region (PMR)
  - PCI memory space exposed to host (byte addressable)
  - May be used to store data
  - Content persist across power cycles and resets



Storage is Not Just About CPU I/O Anymore

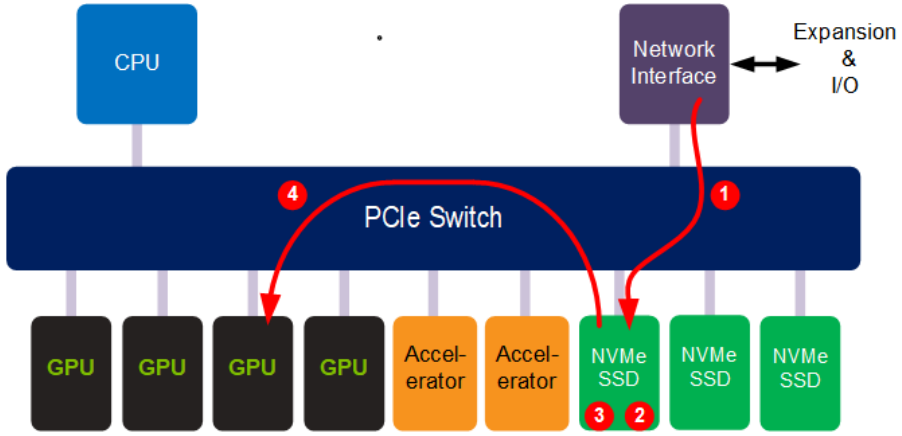
- NVMe together with a PCIe fabric allows direct network-to-storage and accelerator-to-storage communications

Example:

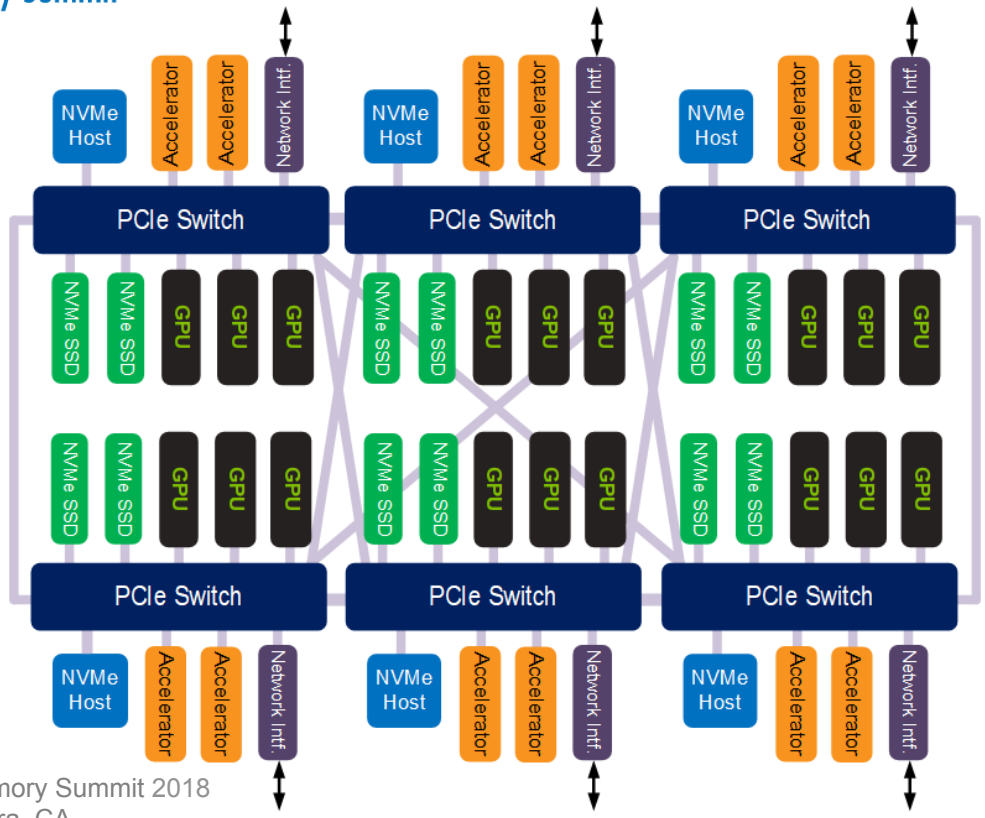
1. Data transferred from network to NVMe CMB
2. NVMe block write operation imitated from CMB to NVM

… sometime later …

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



Putting It All Together



- NVMe Storage Functions
  - Dynamic partitioning (drive-to-host mapping)
  - NVMe shared I/O (shared storage)
- Direct accelerator-to-NVMe and network-to-NVMe transfers
- Byte-addressable persistent memory

Summary

- PCIe fabrics build on the existing PCIe and NVMe ecosystem
  - Work with standard NVMe SSDs, OS drivers, and PCIe infrastructure
- PCIe fabrics support both byte-addressable memory and traditional storage operations
- PCIe fabrics are well-suited for applications that require low cost, the absolute lowest latency, and limited scalability
  - NVMe SSD sharing inside a rack and small clusters
- PCIe fabrics are not well-suited for long-reach applications or where a high degree of scalability is required
  - NVM Express over Fabrics (NVMe-oF™) is well-suited for these applications