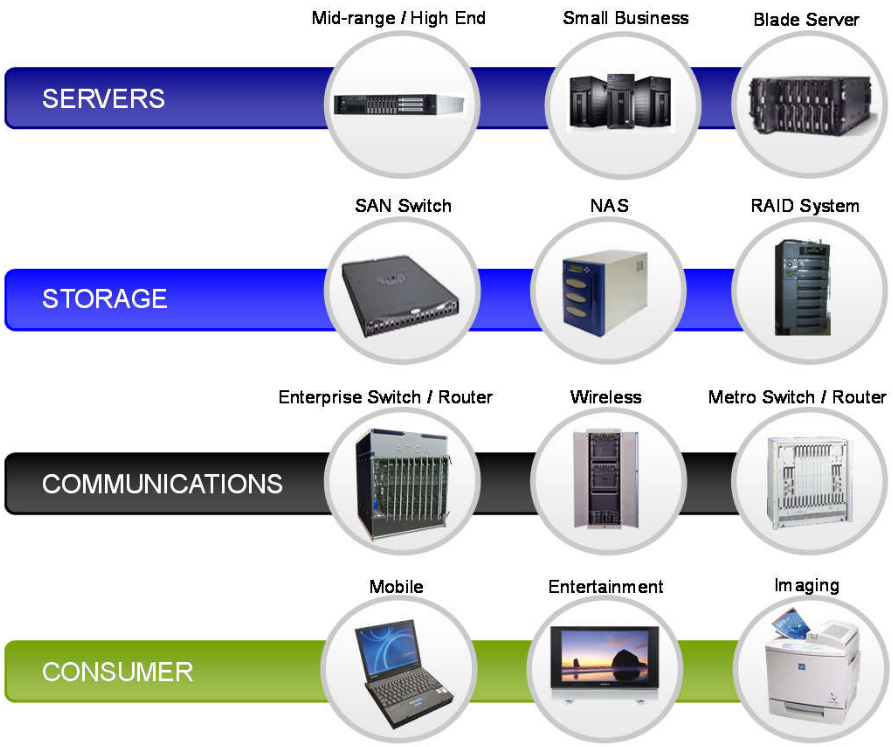
PCIe Networked ~~Flash~~ NVM Storage

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)
  - Scalable link width (x1, x2, x4, …. x32)
- Primary application is as an I/O interconnect

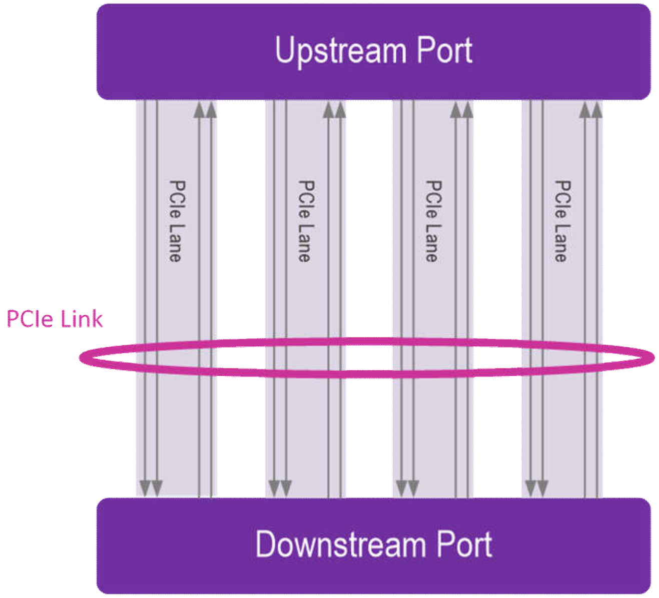


PCIe Characteristics

- Scalable speed
- Scalable width: x1, x2, x4, x8, x12, x16, x32
- Encoding
  - 8b10b: 2.5 GT/s and 5 GT/s
  - 128b/130b: 8 GT/s and 16 GT/s

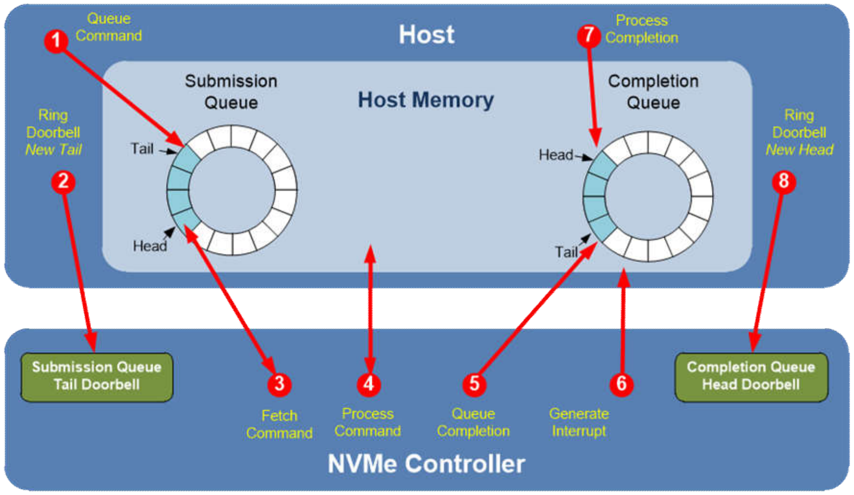
Generation	Raw Bit Rate	Bandwidth Per Lane Each Direction	Total x16 Link Bandwidth
Gen 1*	2.5 GT/s	~ 250 MB/s	~ 8 GB/s
Gen 2*	5.0 GT/s	~500 MB/s	~16 GB/s
Gen 3*	8 GT/s	~ 1 GB/s	~ 32 GB/s
Gen 4	16 GT/s	~ 2 GB/s	~ 64 GB/s

Note
** Source – PCI-SIG PCI Express 3.0 FAQ*



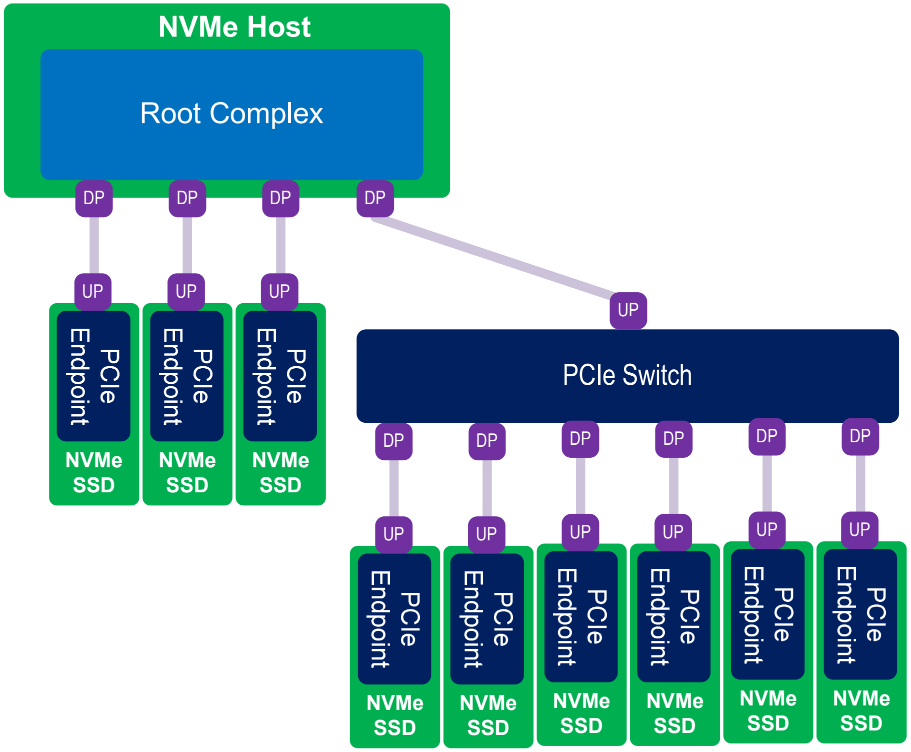
NVM Express (NVMe)

- 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

PCIe and NVMe



Card

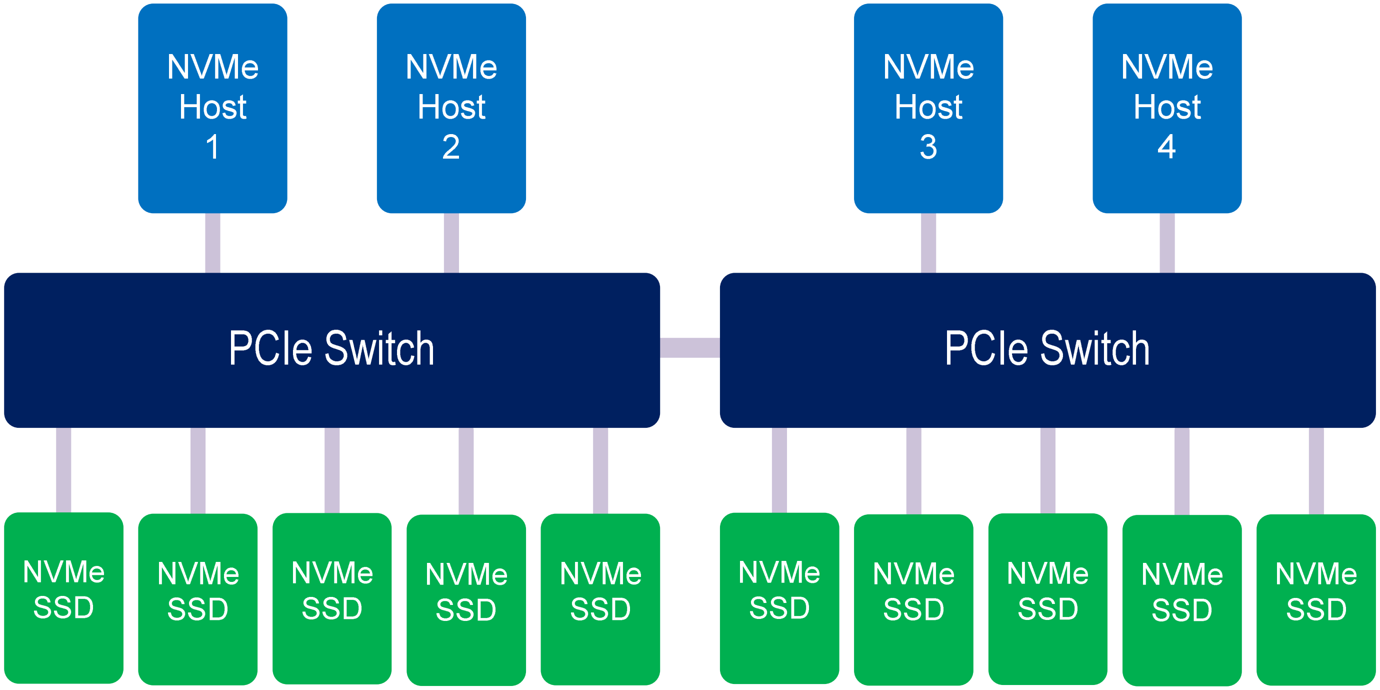
U.2

M.2

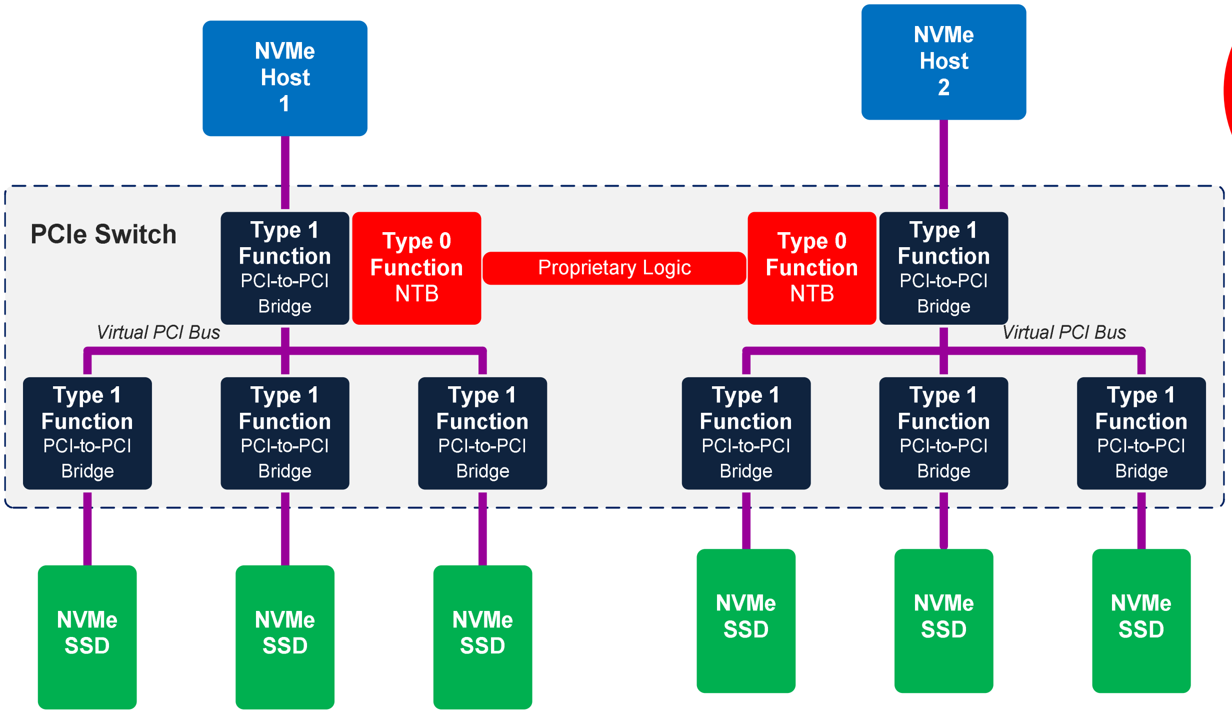
Ideal NVM Fabric

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

PCIe Fabric

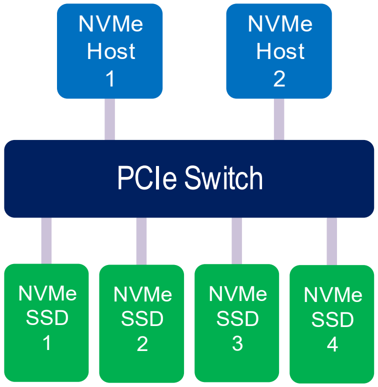

NVMe Host 1
NVMe Host 2
NVMe Host 3
NVMe Host 4
PCIe Switch
PCIe Switch
NVMe SSD
NVMe SSD
NVMe SSD
NVMe SSD
NVMe SSD
NVMe SSD
NVMe SSD
NVMe SSD
NVMe SSD
NVMe SSD

Non-Transparent Bridging (NTB)

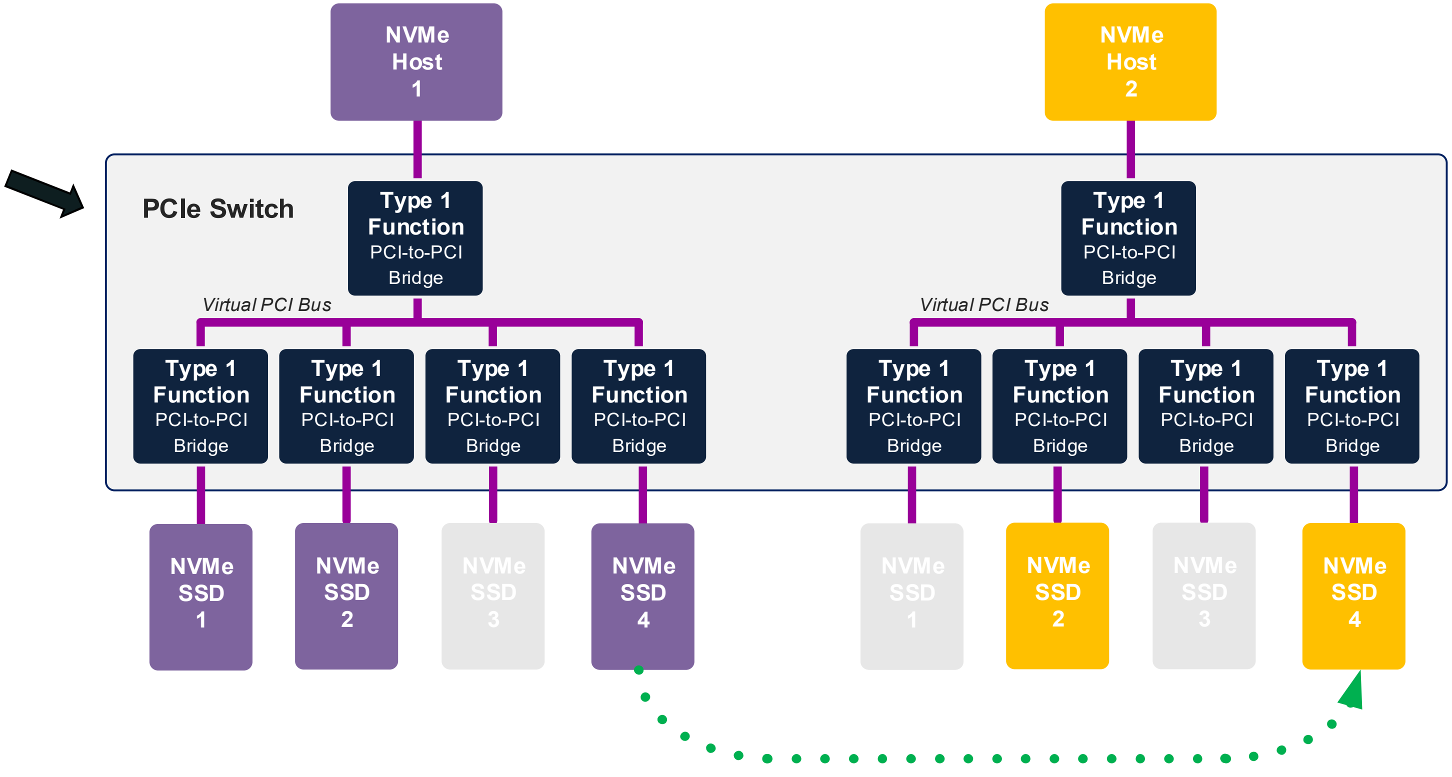

NVMe Host 1
NVMe Host 2
PCIe Switch
Type 1 Function PCI-to-PCI Bridge
Type 0 Function NTB
Proprietary Logic
Type 0 Function NTB
Type 1 Function PCI-to-PCI Bridge
Virtual PCI Bus
Virtual PCI Bus
Type 1 Function PCI-to-PCI Bridge
Type 1 Function PCI-to-PCI Bridge
Type 1 Function PCI-to-PCI Bridge
Type 1 Function PCI-to-PCI Bridge
Type 1 Function PCI-to-PCI Bridge
Type 1 Function PCI-to-PCI Bridge
NVMe SSD
NVMe SSD
NVMe SSD
NVMe SSD
NVMe SSD
NVMe SSD

NTB

Dynamic Partitioning

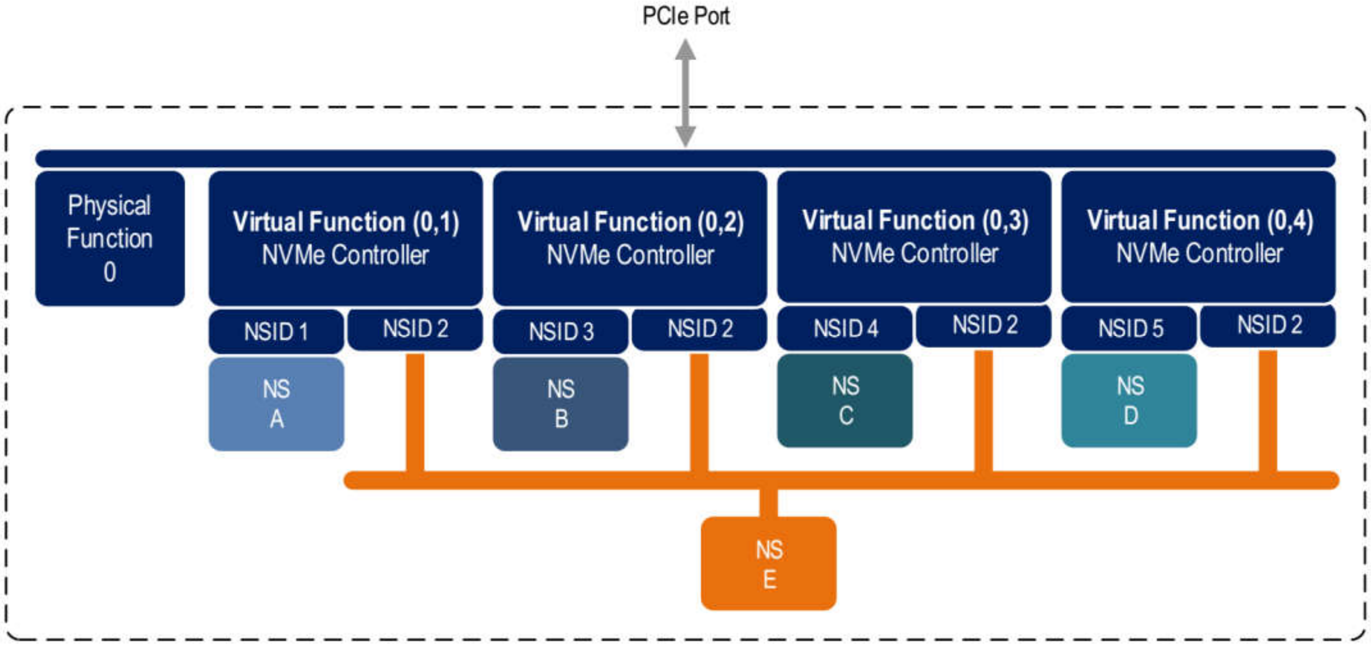


Physical View

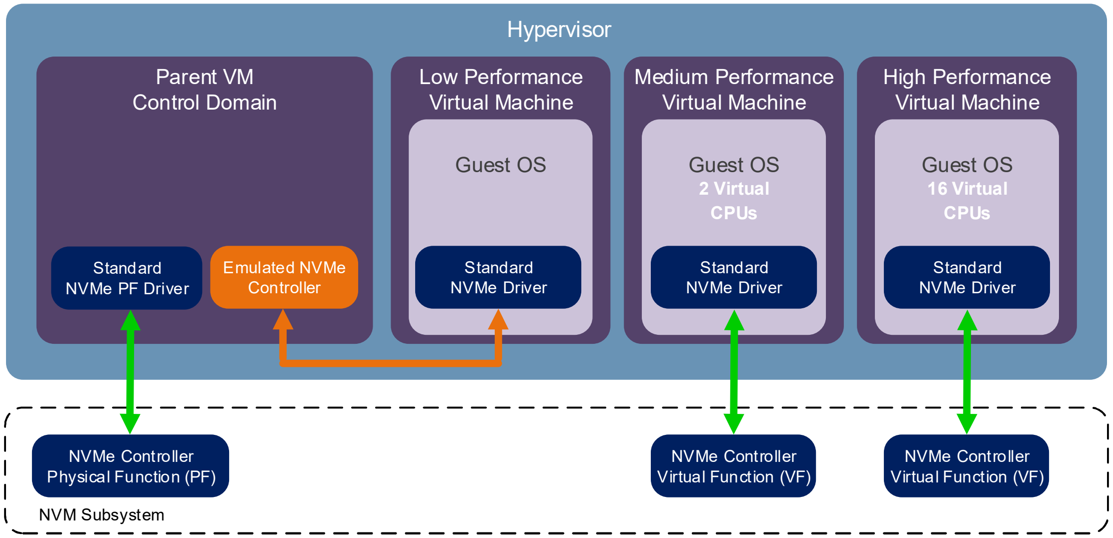

NVMe Host 1
NVMe Host 2
PCIe Switch
Type 1 Function PCI-to-PCI Bridge
Virtual PCI Bus
Type 1 Function PCI-to-PCI Bridge
Type 1 Function PCI-to-PCI Bridge
Type 1 Function PCI-to-PCI Bridge
Type 1 Function PCI-to-PCI Bridge
Type 1 Function PCI-to-PCI Bridge
Virtual PCI Bus
Type 1 Function PCI-to-PCI Bridge
Type 1 Function PCI-to-PCI Bridge
Type 1 Function PCI-to-PCI Bridge
Type 1 Function PCI-to-PCI Bridge
NVMe SSD 1
NVMe SSD 2
NVMe SSD 3
NVMe SSD 4
NVMe SSD 1
NVMe SSD 2
NVMe SSD 3
NVMe SSD 4

Functional View

NVMe SR-IOV

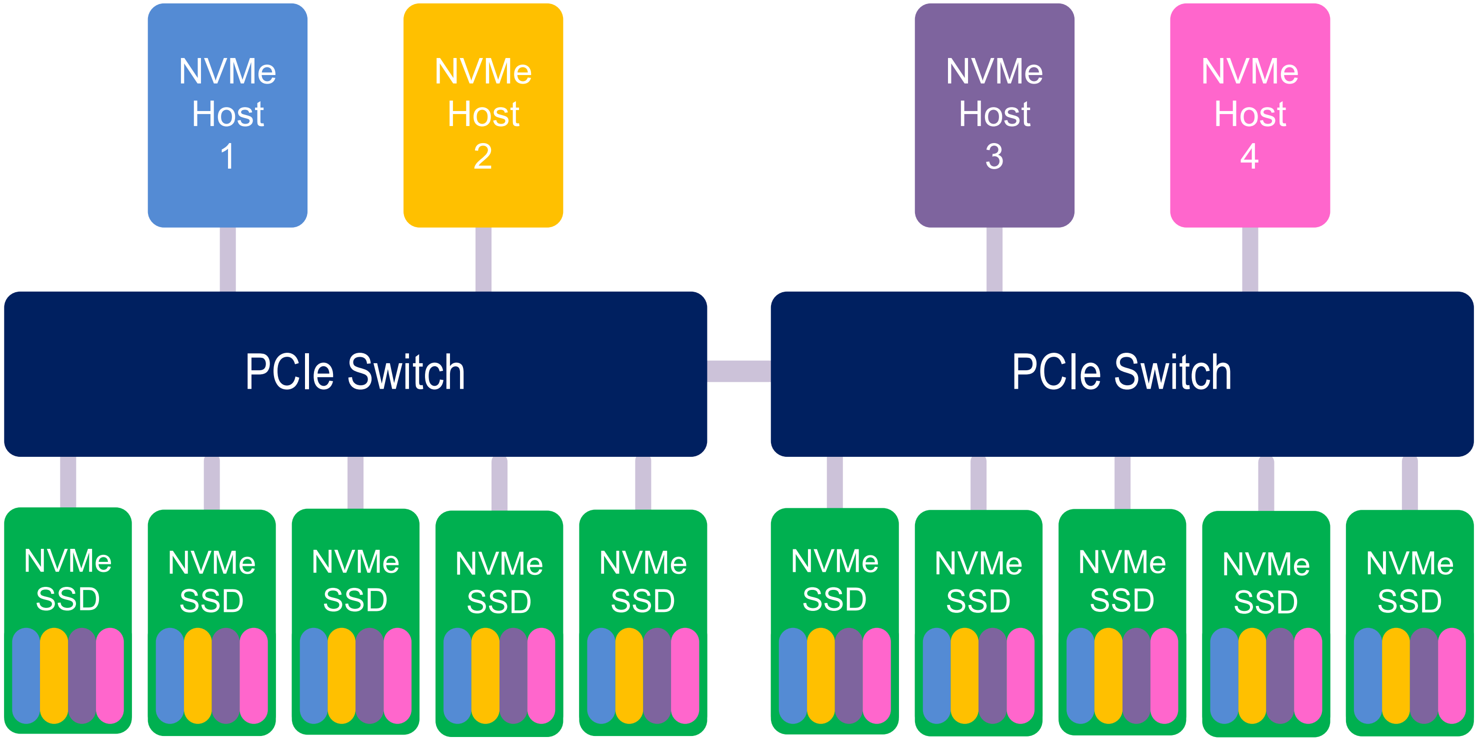

PCIe Port
Physical Function 0
Virtual Function (0,1) NVMe Controller
Virtual Function (0,2) NVMe Controller
Virtual Function (0,3) NVMe Controller
Virtual Function (0,4) NVMe Controller
NSID 1
NSID 2
NSID 3
NSID 2
NSID 4
NSID 2
NSID 5
NSID 2
NS A
NS B
NS C
NS D
NS E

NVMe Single Root I/O Virtualization (SR-IOV)

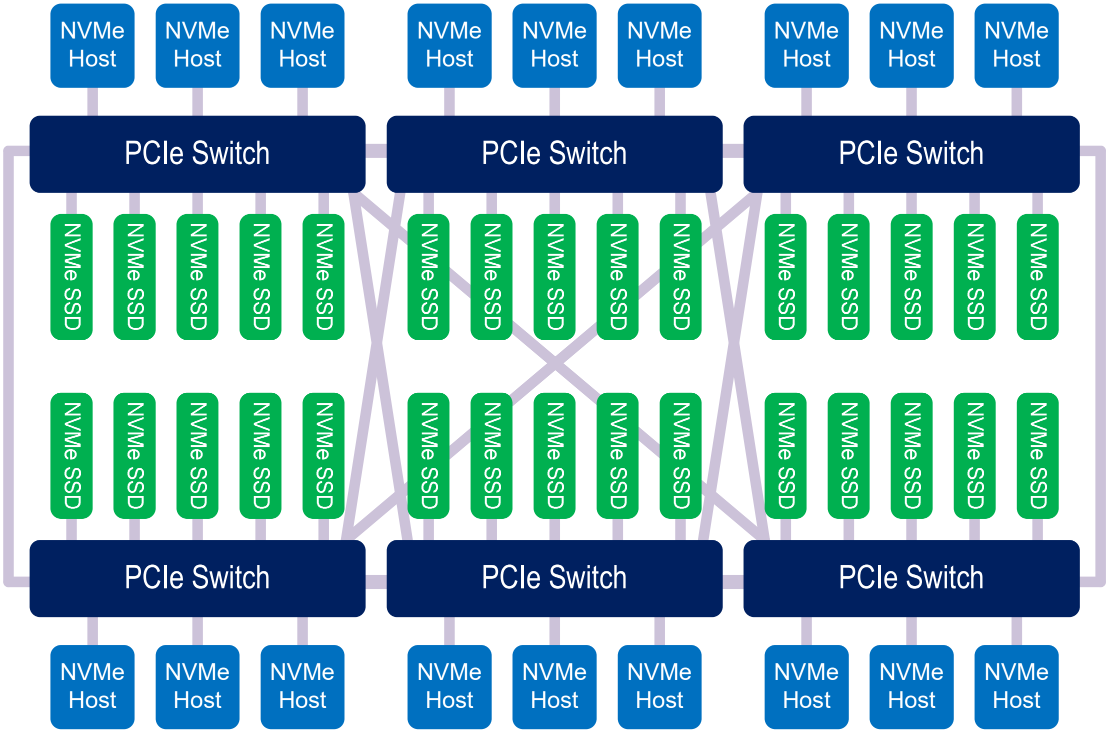


- Datacenter service providers that host VMs win with high density, oversubscription, and differentiation
  - Multi-tenancy is the norm
  - Premium differentiator offering high speed storage
- Virtual machines are inherently mobile and hosts are inherently dynamic

Multi-Host I/O Sharing



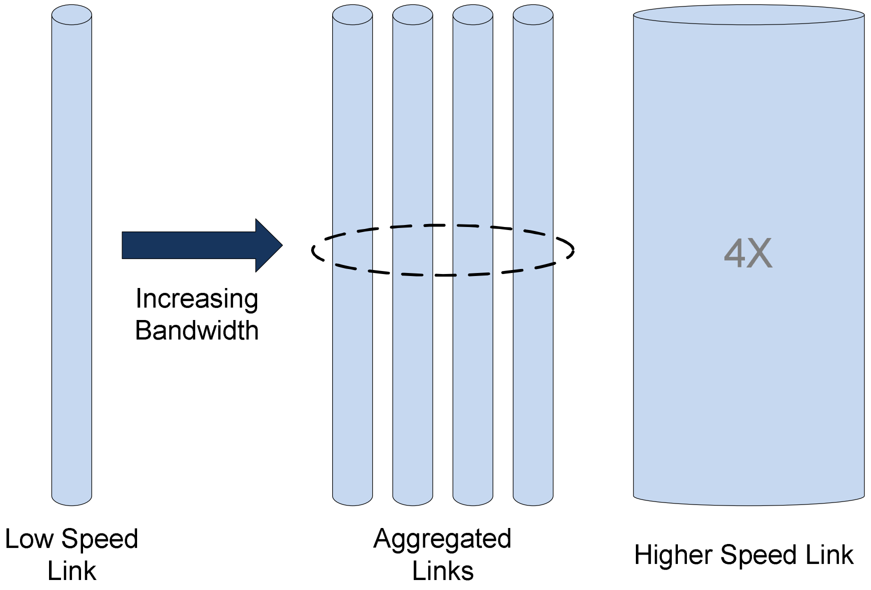
PCIe Fabric



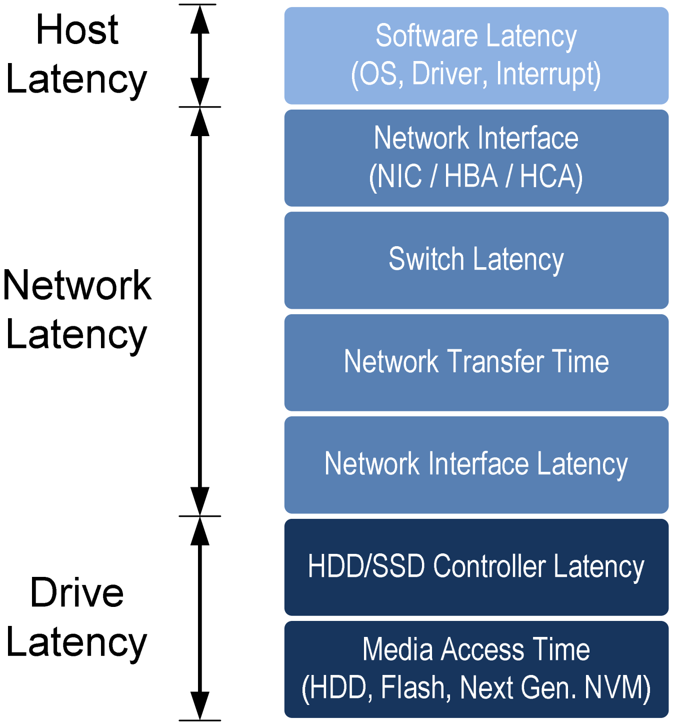
- Storage Functions
  - Dynamic partitioning (drive-to-host mapping)
  - NVMe shared I/O (shared storage)
  - Ability to share other storage (SAS/SATA)
- Host-to-Host Communications
  - RDMA
  - Ethernet emulation
- Manageability
  - NVMe controller-to-host mapping
  - PCIe path selection
  - NVMe management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

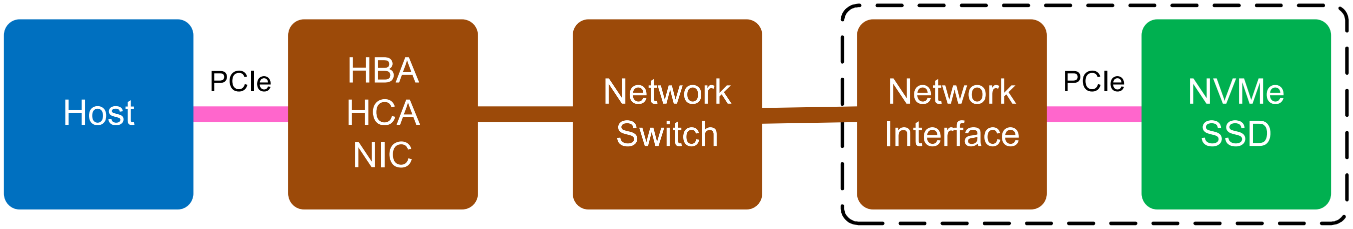


Latency and Next-Generation NVM

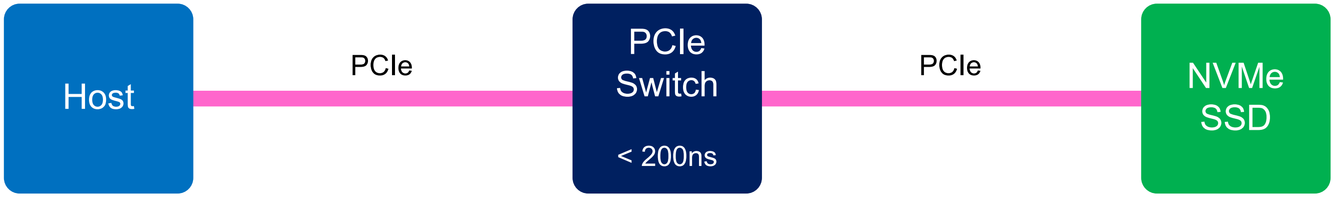

Host Latency
Network Latency
Drive Latency
Software Latency (OS, Driver, Interrupt)
Network Interface (NIC / HBA / HCA)
Switch Latency
Network Transfer Time
Network Interface Latency
HDD/SSD Controller Latency
Media Access Time (HDD, Flash, Next Gen. NVM)

- Media Access Time
  - Hard drive – Milliseconds
  - NAND flash –Microseconds
  - Next-gen. NVM – Nanoseconds

The PCIe Advantage



Other Flash Storage Networks



PCIe Fabric

PCIe Fabric Characteristics

Property	Ideal Characteristic	PCIe Fabric	Notes
Cost	Free	Low	• PCIe built into virtually all hosts and NVMe drives
Complexity	None	Medium	• 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)

Summary

- PCIe fabrics build on the existing PCIe and NVMe ecosystem
  - Work with standard NVMe SSDs and OS drivers
  - Leverage standard PCIe infrastructure

- PCIe fabrics are well suited for applications that require the absolute lowest latency and limited scalability
  - NVMe SSD sharing inside a rack
  - Small clusters

- PCIe fabrics are not well suited for long reach applications or where a high degree of scalability is required
  - NVMe over fabrics is well suited for these applications