Motti Beck

Motti Beck is Sr. Director of Marketing, Enterprise Data Center market segment at Mellanox Technologies, Inc. Before joining Mellanox, Motti was a founder of several start-up companies including BindKey Technologies that was acquired by DuPont Photomask (today Toppan Printing Company LTD) and Butterfly Communications that was acquired by Texas Instrument. Prior to that he was a Business Unit Director at National Semiconductors. Motti hold B.Sc in computer engineering from the Technion - Israel Institute of Technology.

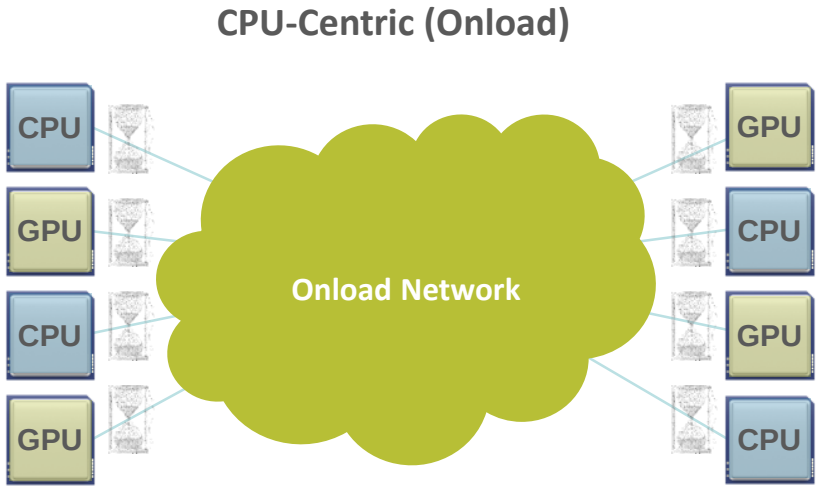
InfiniBand Networked Flash Storage

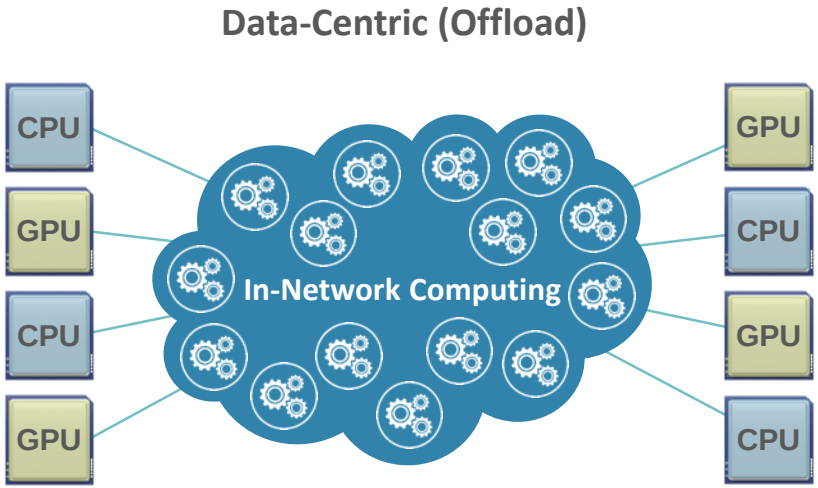
Superior Performance, Efficiency and Scalability

Motti Beck – Sr. Director Enterprise Market Development, Mellanox Technologies

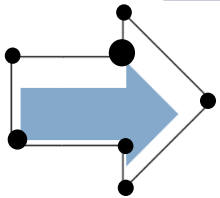
The Need for Intelligent and Faster Interconnect

Faster Data Speeds and In-Network Computing Enable Higher Performance and Scale





Must Wait for the Data
Creates Performance Bottlenecks



Analyze Data as it Moves!
Higher Performance and Scale

Flash Memory Summit

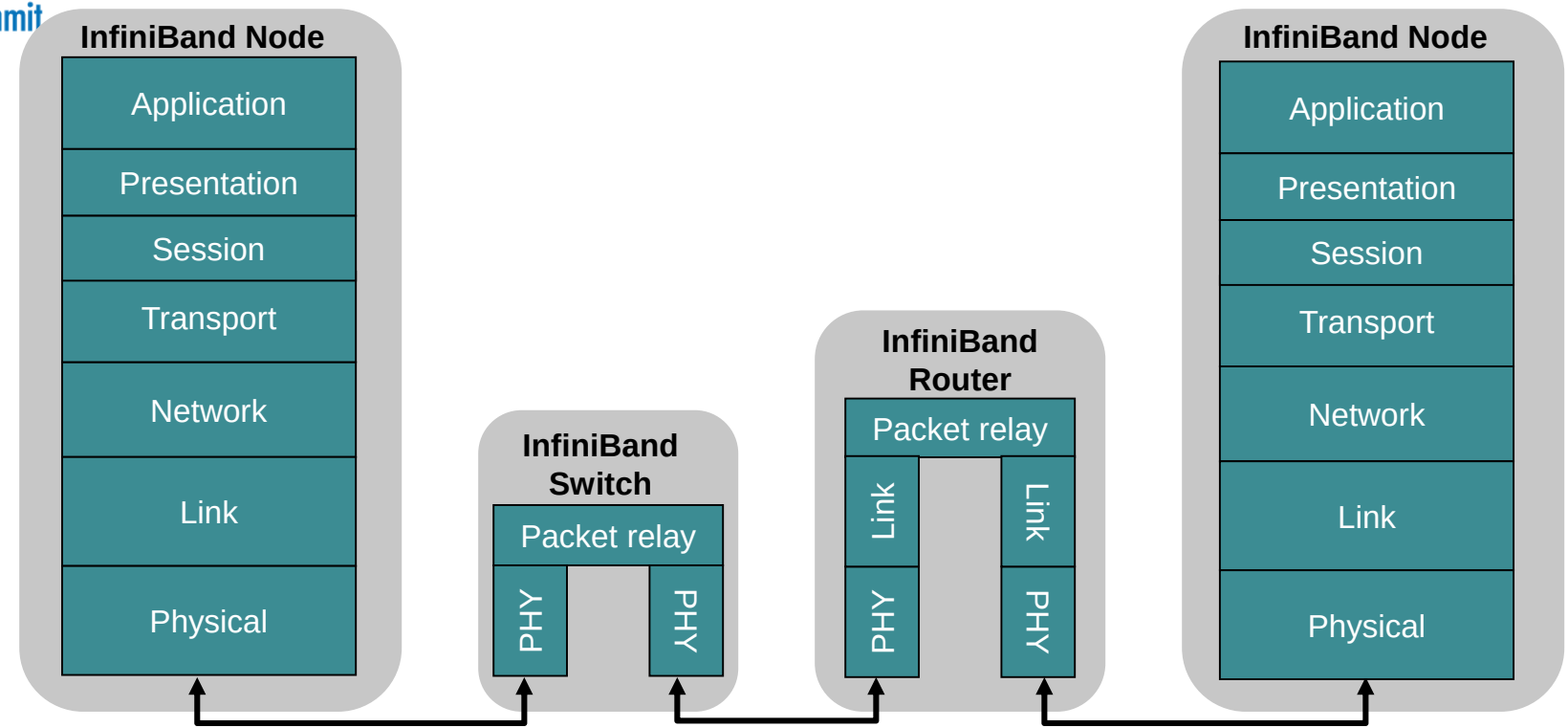
In-Network Processing Enables Higher Efficiency

- Higher Scalability
- Lower latency
- Higher ROI

InfiniBand Technical Overview

- What is InfiniBand?
  - InfiniBand is an open standard, interconnect protocol developed by the InfiniBand® Trade Association: http://www.infinibandta.org/home
  - First InfiniBand specification was released in 2000
- What does the specification includes?
  - The specification is very comprehensive
  - From physical to applications
- InfiniBand SW is open and has been developed under OpenFabrics Alliance
  - http://www.openfabrics.org/index.html

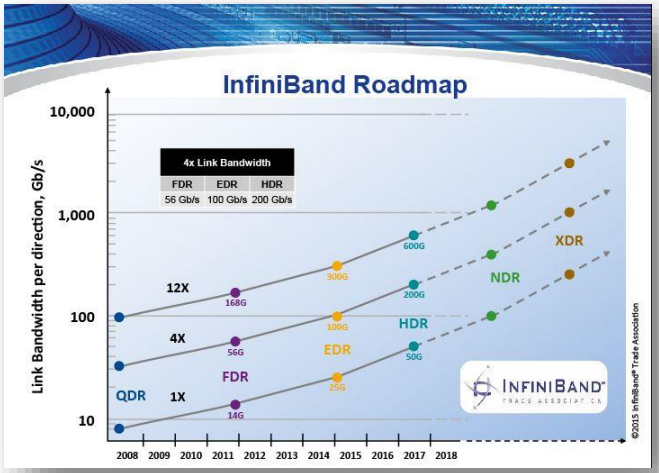
InfiniBand Protocol Layers

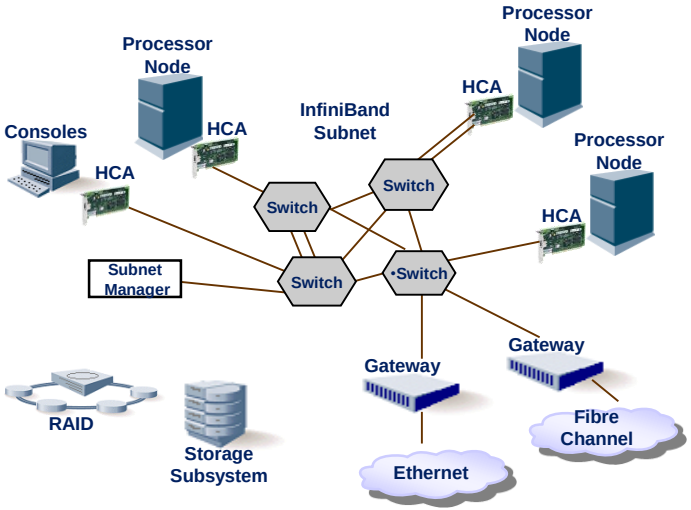


Packet Format

InfiniBand Architecture Highlights

- Reliable, lossless, self-managed fabric
- Hardware based transport protocol- Remote Direct Memory Access (RDMA)
- Centralized fabric management – Subnet Manger (SM)

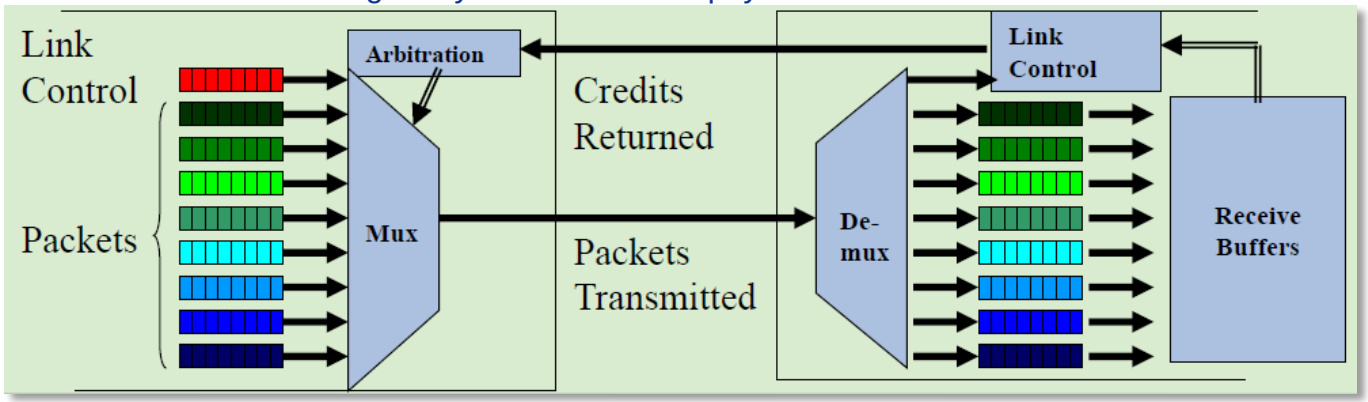




Reliable, Lossless, Self-Managed Fabric

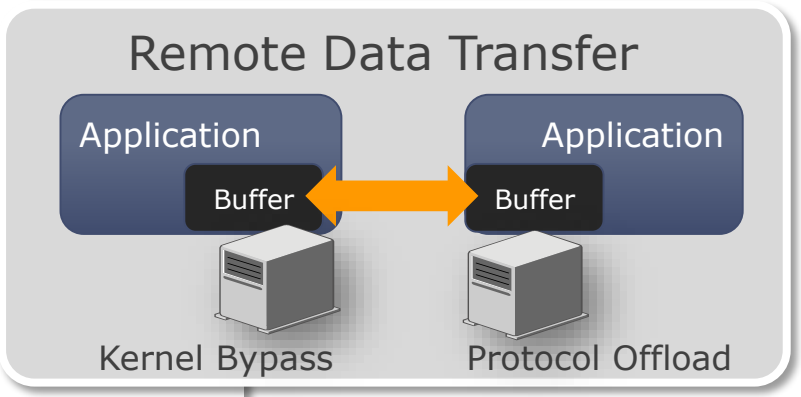
- Credit-based link-level flow control
  - Link Flow control assures **NO packet loss** within fabric even in the presence of congestion
  - Link Receivers grant packet receive buffer space credits per Virtual Lane
  - Flow control credits are issued in 64 byte units

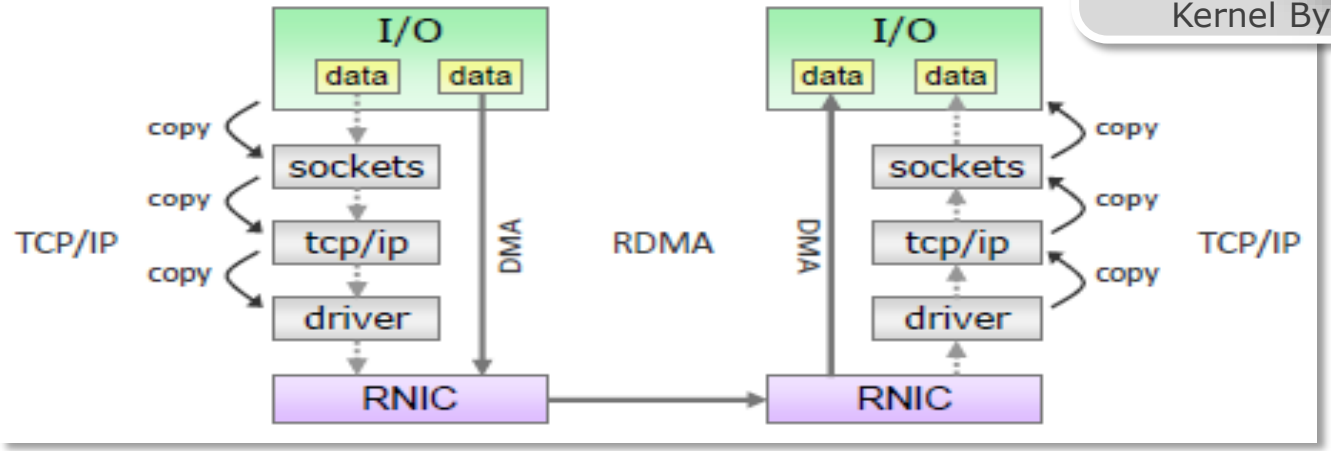
- Separate flow control per Virtual Lanes provides:
  - Alleviation of head-of-line blocking
  - Virtual Fabrics – Congestion and latency on one VL does not impact traffic with guaranteed QOS on another VL even though they share the same physical link



Remote Direct Memory Access RDMA

- Transport offload
- Kernel bypass

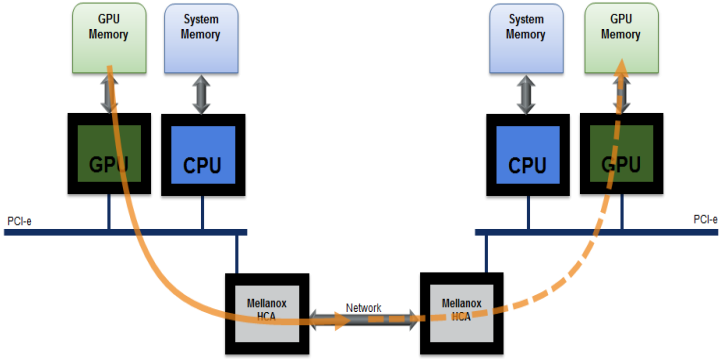


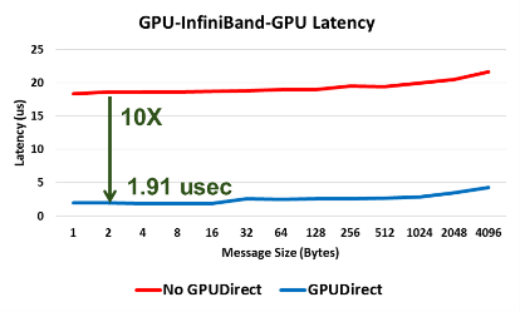


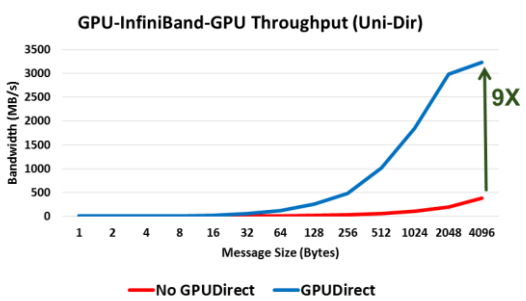
10X Better Performance with GPUDirect™ RDMA

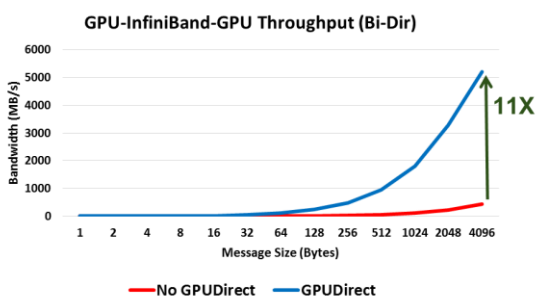
- Purpose-built for Acceleration of Deep Learning
- Lowest communication latency for acceleration devices
- No unnecessary system memory copies and CPU overhead
- Enables GPUDirect™ RDMA and ASYNC, ROCm and others
- InfiniBand and RoCE

GPUDirect™ RDMA, GPUDirect™ ASYNC









Scaling HPC and ML with GPUDirect over InfiniBand on vSphere 6.7

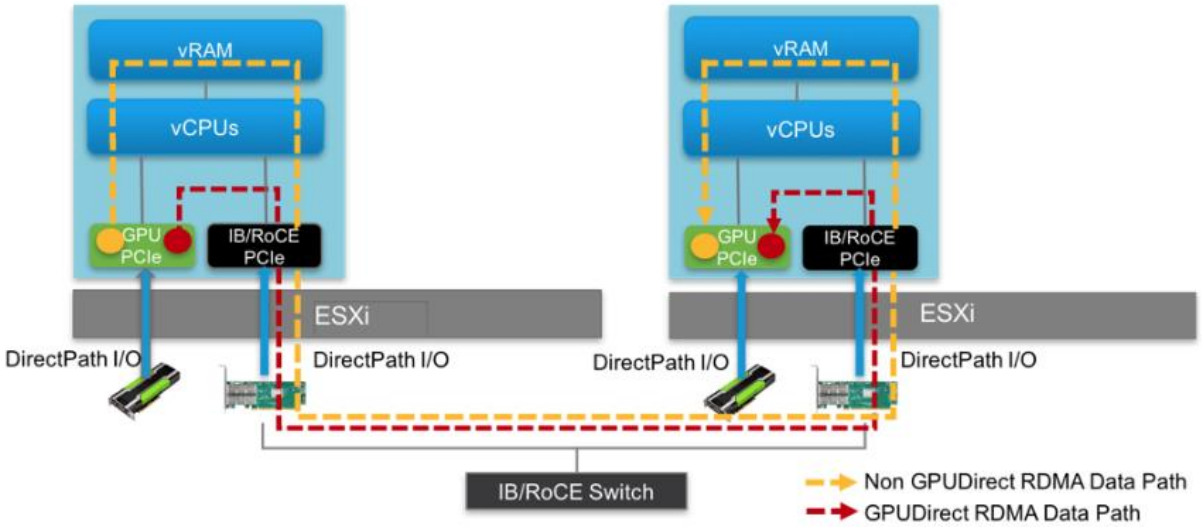
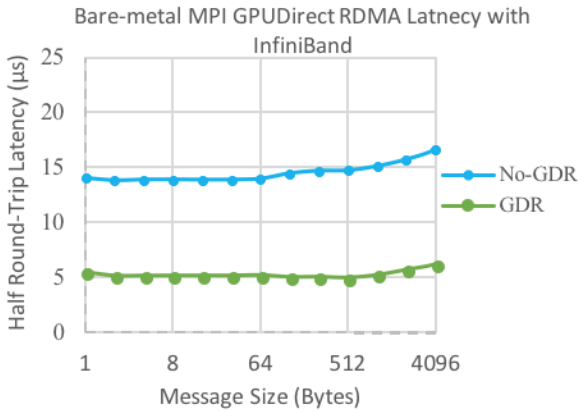
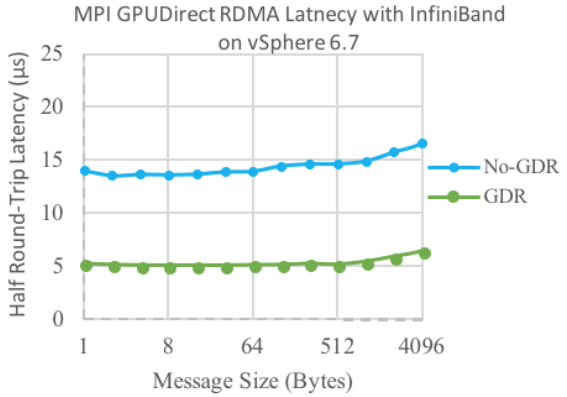


Figure 3: Testbed virtual cluster architecture showing the no-GPUDirect RDMA vs. GPUDirect RDMA data path with DirectPath I/O on vSphere 6.7





Source: Scaling HPC and ML with GPUDirect RDMA on vSphere 6.7

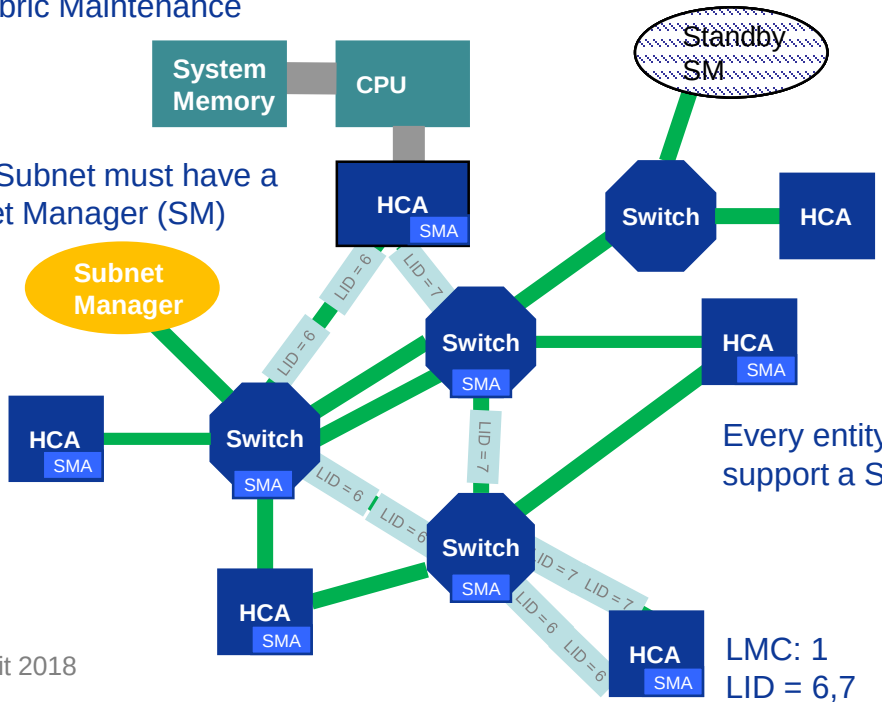
Subnet Management

Topology Discovery
Fabric Initialization
Fabric Maintenance

Initialization uses
Directed Route MADs:

Management use unreliable datagrams (MAD)

Each Subnet must have a Subnet Manager (SM)



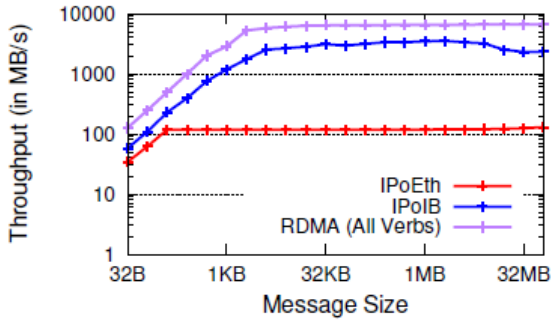
Every entity (HCA, Switch or Router) must support a Subnet Management Agent (SMA)

LMC: 1
LID = 6,7

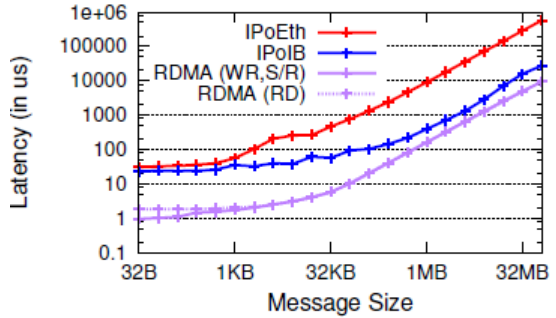
Multipathing: LMC Supports Multiple LIDS

InfiniBand Superior Performance*

Network Throughput and Latency

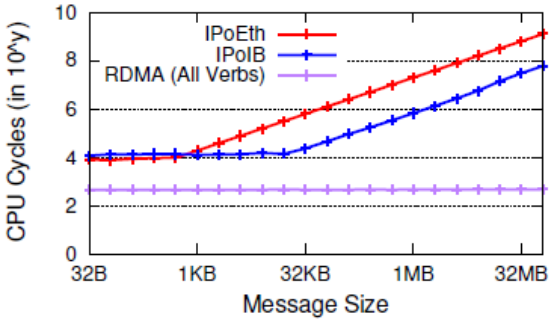


(a) Throughput

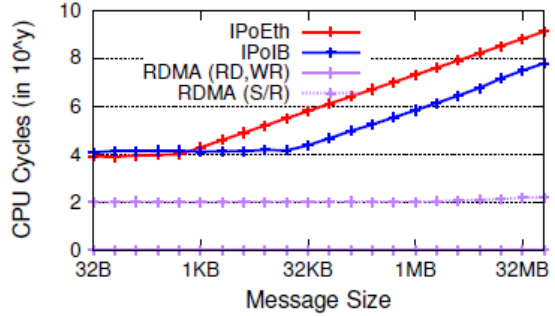


(b) Latency

CPU Overhead for Network Operations



(a) Client

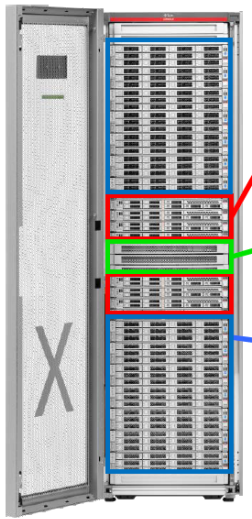


(b) Server

* Source: Brown University Research: "The End of Slow Networks: It's Time for a Redesign"

InfiniBand Enables Most Cost Effective Database Storage

Exadata X5-2 Product Components



- **Scale-Out Database Servers**
  - **Two 18-core x86 Processors (36 cores)**
  - Oracle Linux 6
  - Oracle Database Enterprise Edition
  - Oracle VM (optional)
  - Oracle Database options (optional)
- **Fastest Internal Fabric**
  - 40 Gb/s InfiniBand
  - Ethernet External Connectivity
- **Scale-Out Intelligent Storage**
  - **High-Capacity Storage Server**
  - **Extreme Flash Storage Server**
  - **Exadata Storage Server Software**

X5-2 Database Server

36 cores per server
256 – 768 GB DRAM

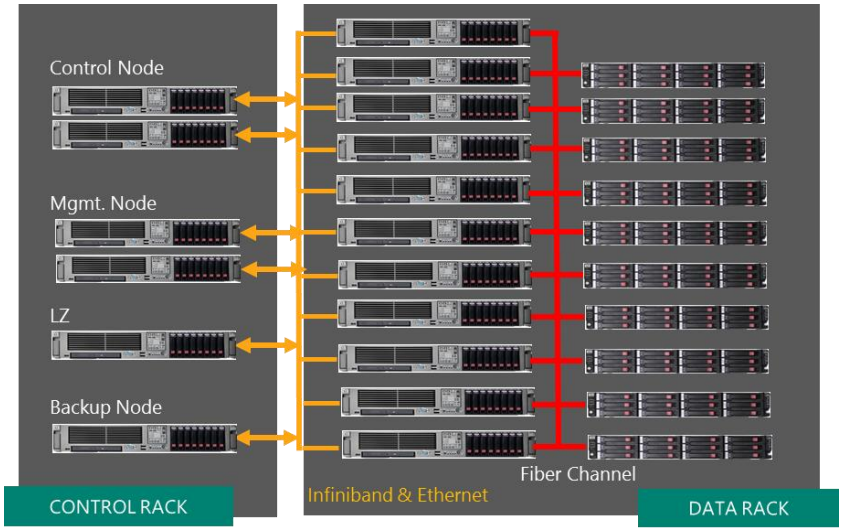
High-Capacity Storage Server

Extreme Flash Storage Server

Source: Oracle

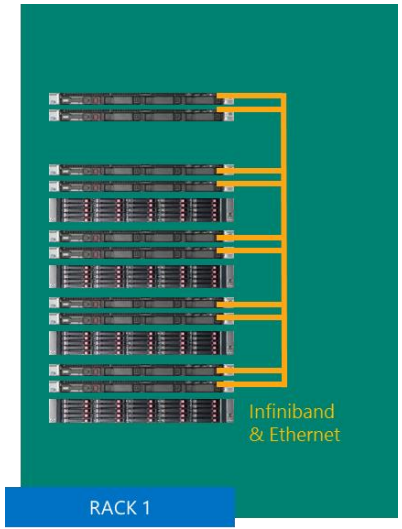
InfiniBand Networking Storage enables Higher Efficiency

PDW* V1 Reference: The Basic Full Rack



Per RACK details
- 160 cores on 10 compute nodes
- 1.28 TB of RAM on compute
- Up to 30 TB of temp DB
- Up to 150 TB of user data

Parallel Data Warehouse
10X Faster & Lower Capital Cost

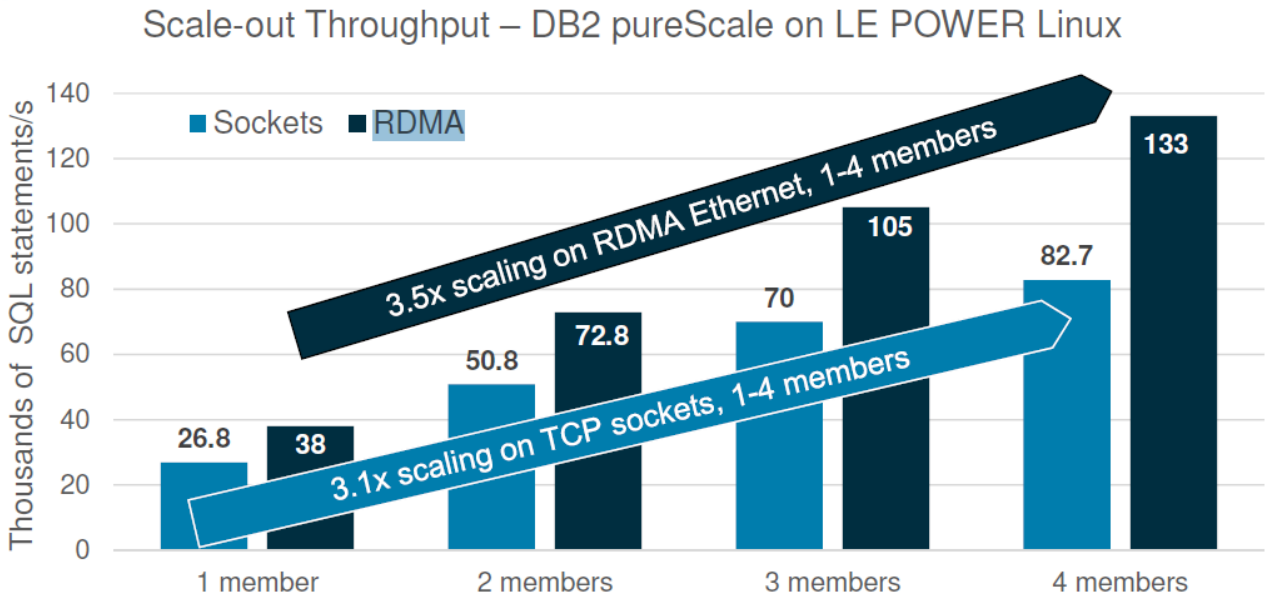


Per RACK Details
- 128 cores on 8 compute nodes
- 2TB of RAM on compute
- Up to 168 TB of temp DB
- Up to 1PB of user data

*Parallel Data Warehouse

Source: Big Data Integration with SQL Server PDW 2012

RDMA enables Higher Scalability with IBM DB2 pureScale

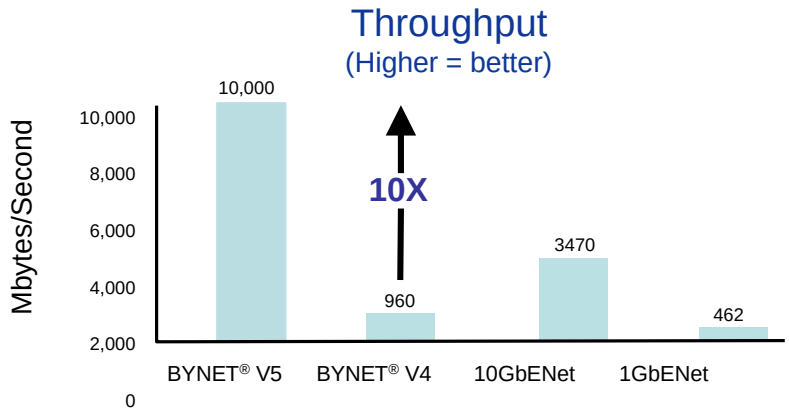


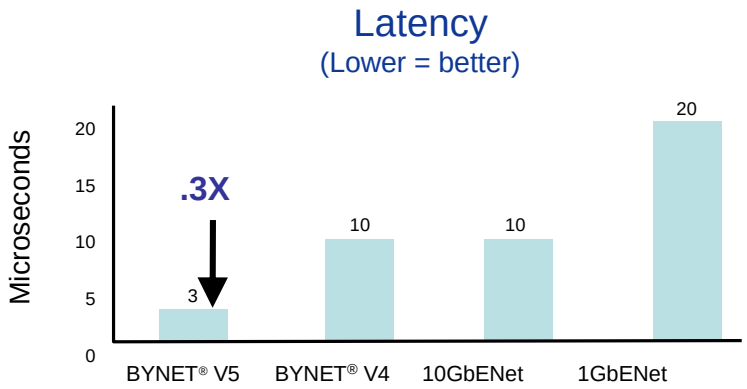
- 80% read / 20% write OLTP workload
- POWER8 4c/32t, 160 GB LBP
- 10 Gb RoCE RDMA Ethernet / 10 Gb TCP sockets

Source: IBM

Teradata BYNET® V5 Performance

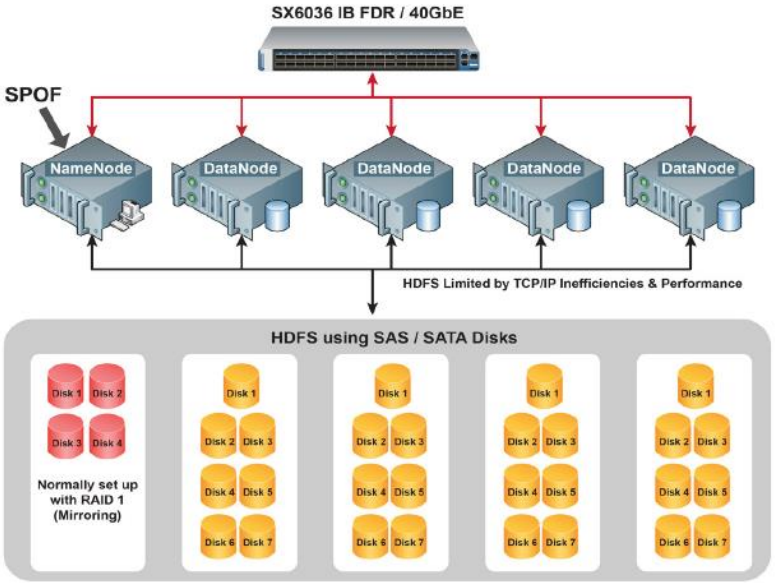
- BYNET's basic link performance enhanced with InfiniBand
  - Dual InfiniBand links provide 10GB per second
  - 10X higher than previous BYNET®
- Message delays decreased
  - Latency in interconnect reduced by 2/3



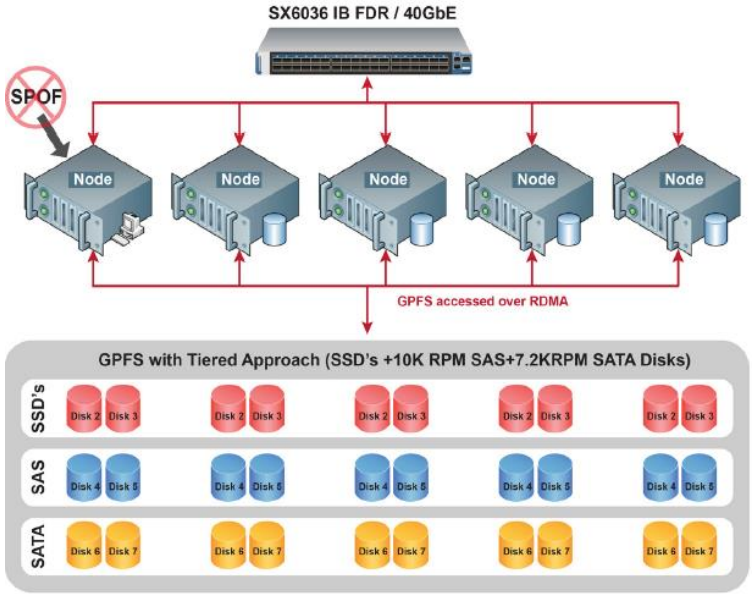


InfiniBand Unleashed the Power of Flash

Hadoop HDFS Architecture

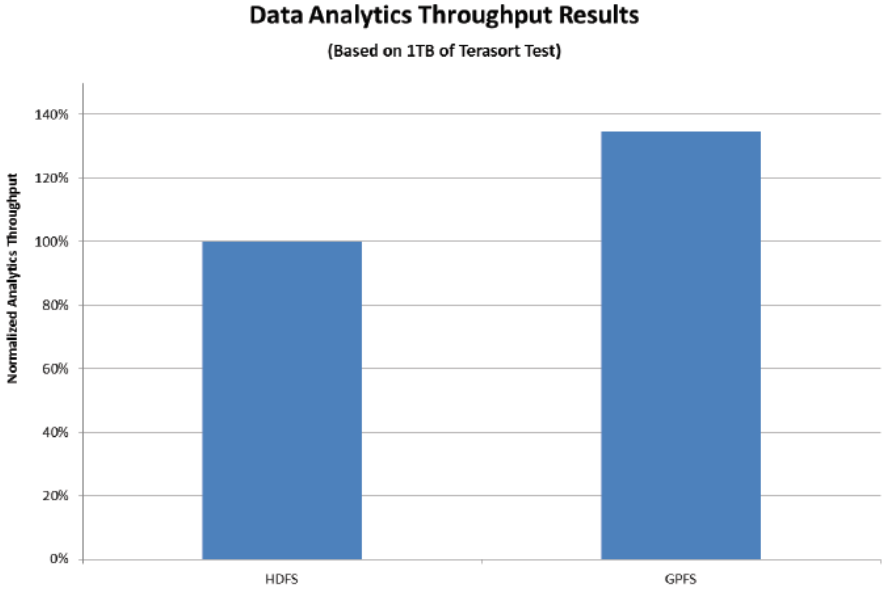


Hadoop GPFS Architecture

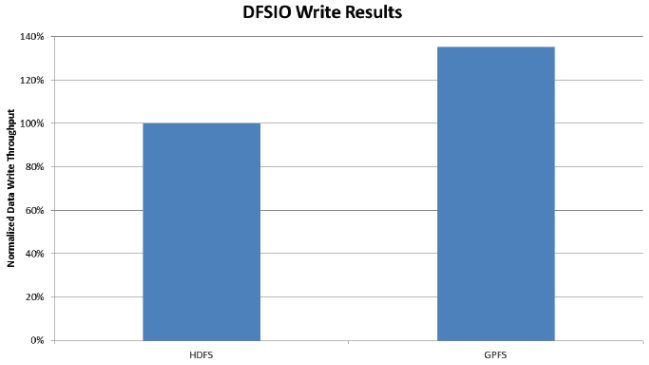


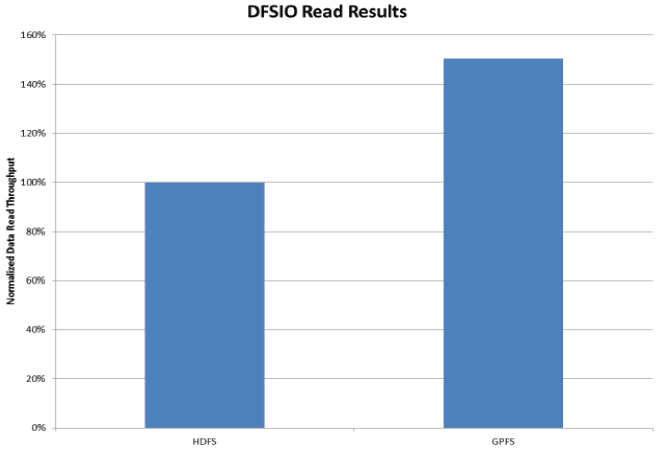
Source: Driving IBM BigInsights Performance Over GPFS Using InfiniBand+RDMA

InfiniBand Accelerate Big Data Analytics

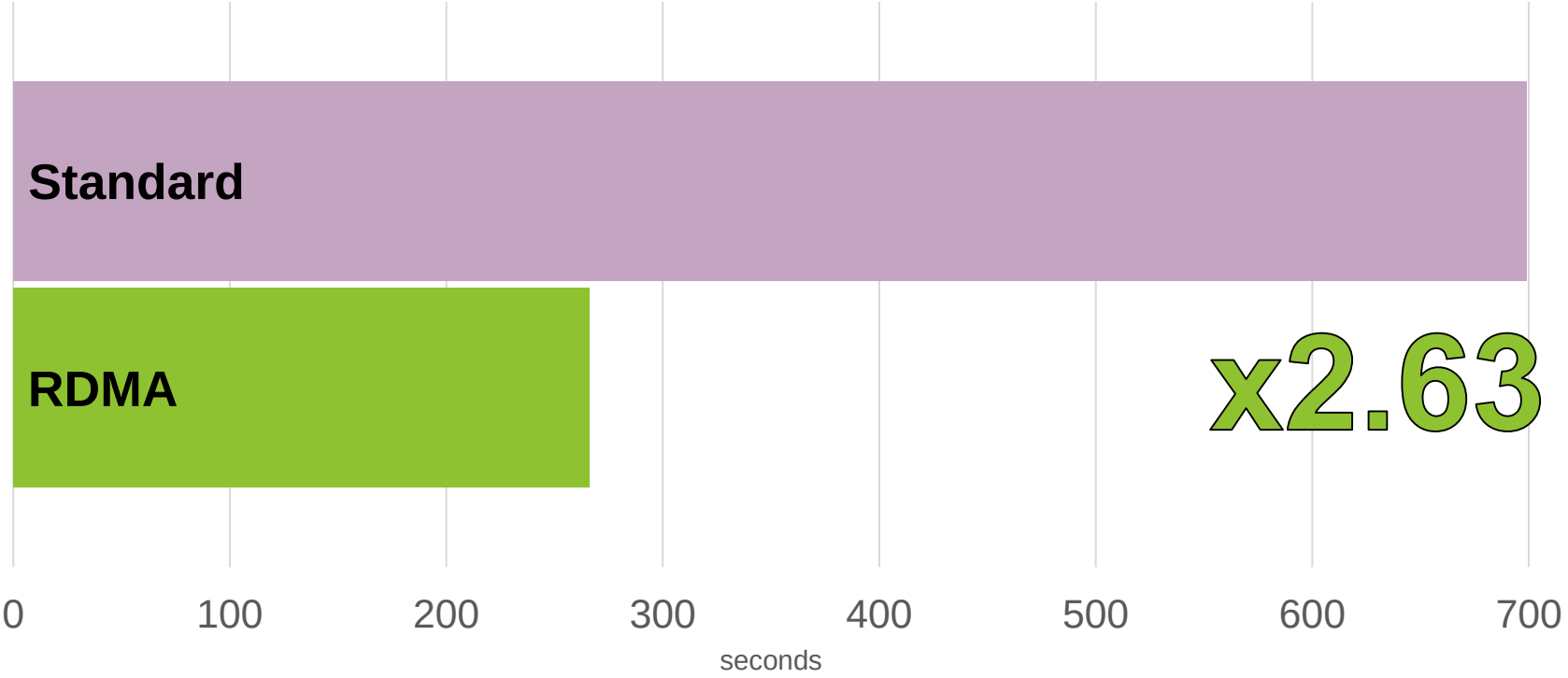


Source: Driving IBM BigInsights Performance Over GPFS Using InfiniBand+RDMA



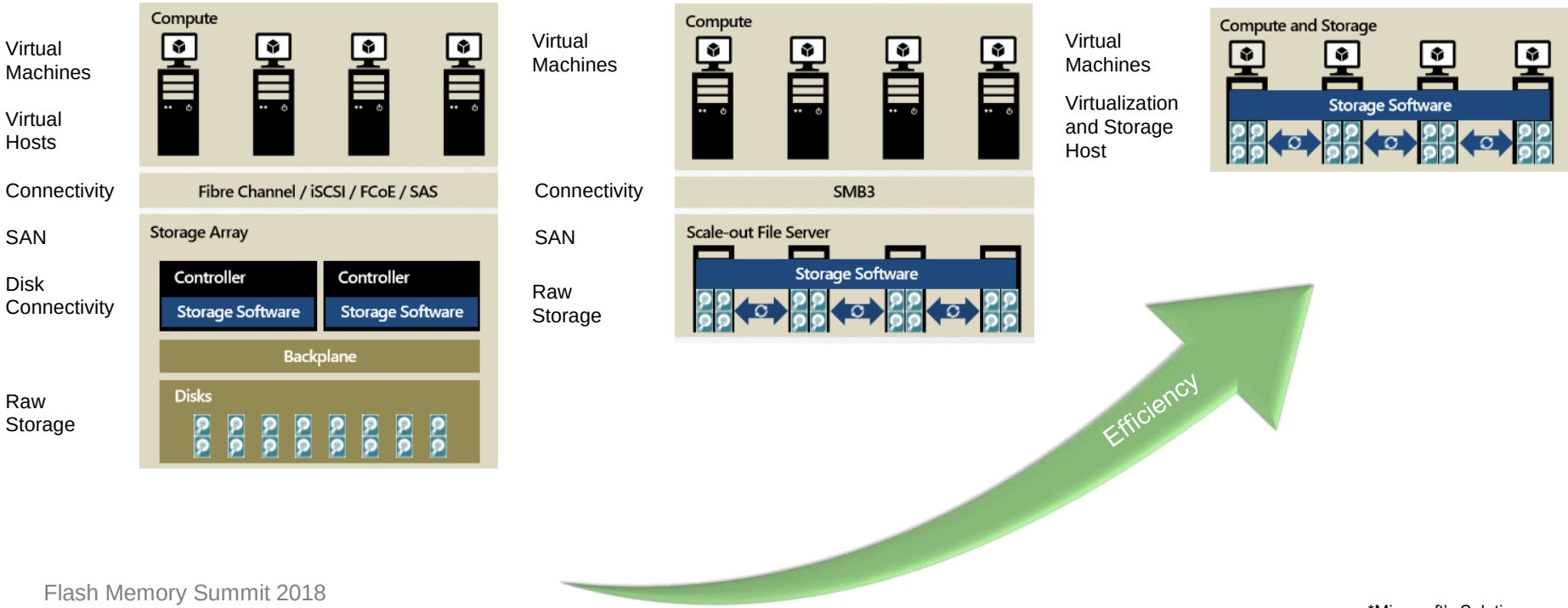


TeraSort - Performance Results



Source: #HWCSAIS18

RDMA Enables Higher Performance SDS Solutions

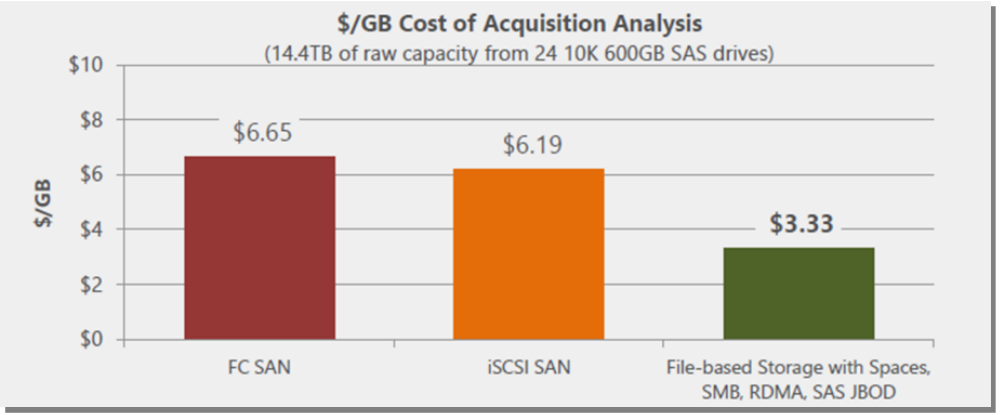


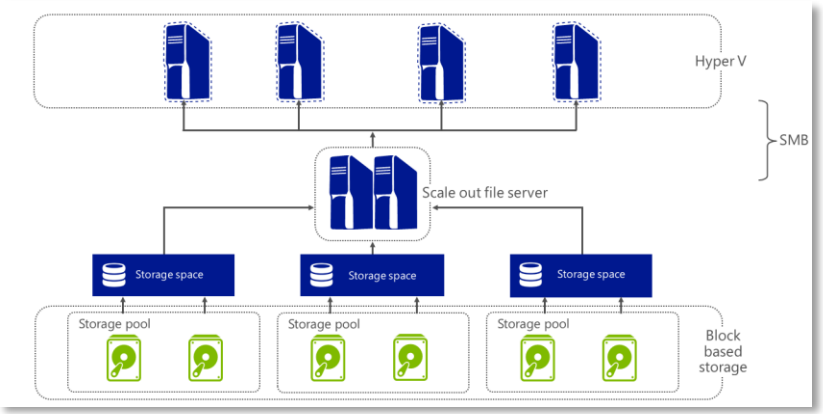
*Microsoft's Solutions

InfiniBand Cuts SAN Cost by 50%

- Delivers SAN-like functionality from the Windows Stack
  - Using SMB Direct (SMB 3.0 over RDMA)

- Utilize inexpensive, industry-standard, commodity hardware
  - Eliminate the cost of proprietary hardware and software from SAN solutions

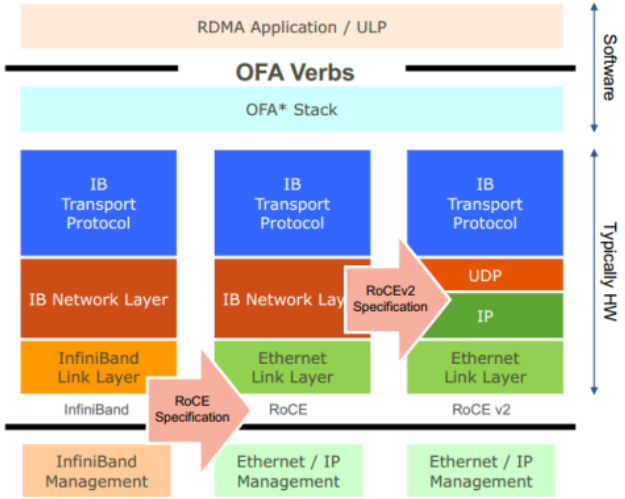


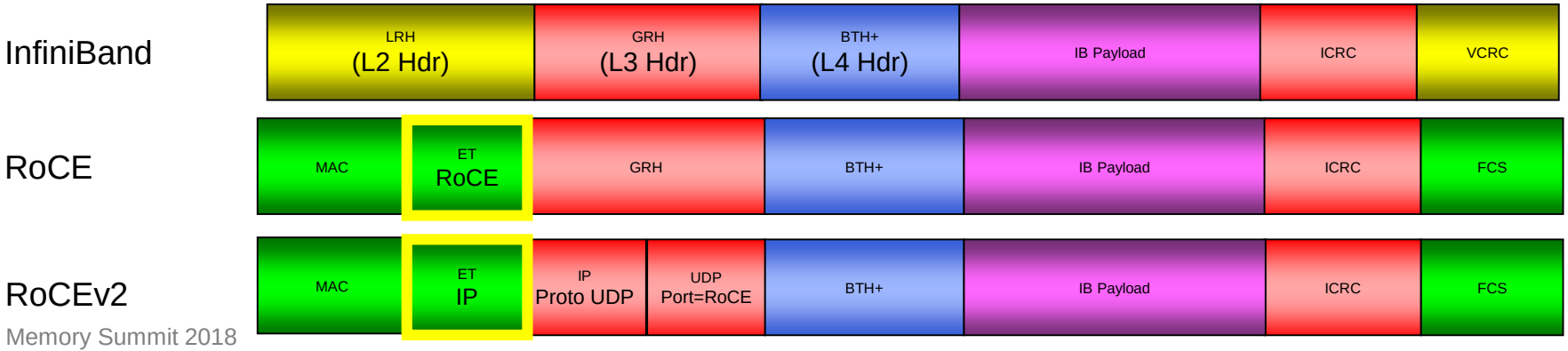


Source: Microsoft

RoCE – RDMA (InfiniBand) over Converged Ethernet

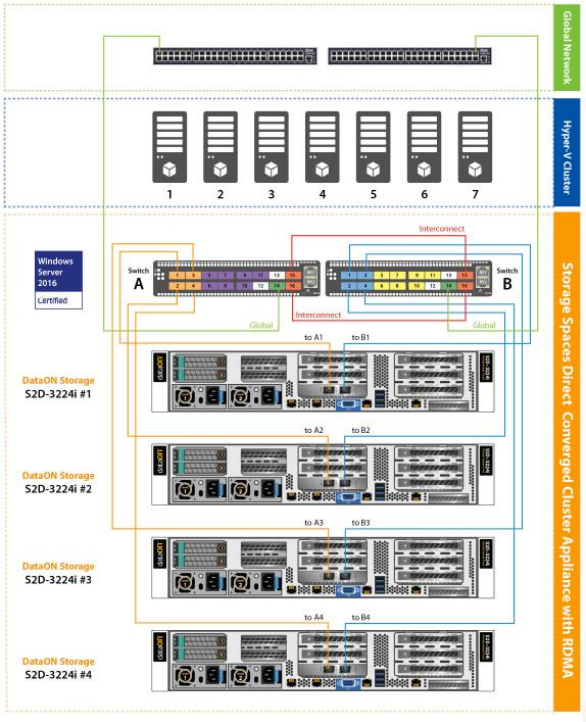
- InfiniBand transport over Ethernet
- API Compatible
- Efficient, light-weight transport, layered directly over
  - Ethernet – RoCE
  - UDP – RoCEv2
- Takes advantage of DCB Ethernet
  - PFC, ETS, and QCN





DataON WSSD* Hyper-Converged Infrastructure

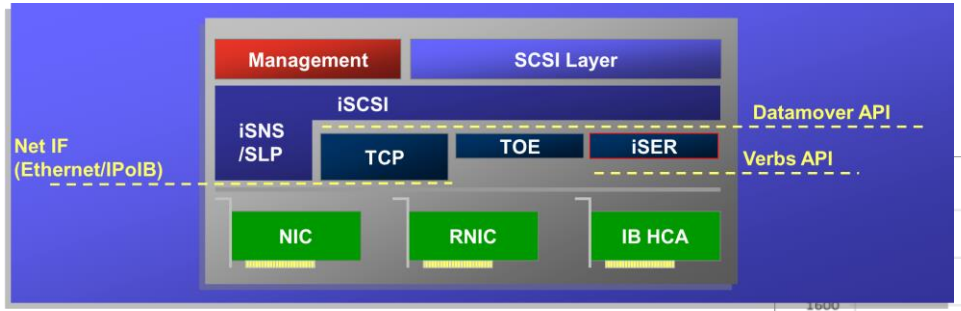
- Microsoft's WSSD Certified
- RoCE networking
- Increased efficiency
  - 30X** vs. previous solution

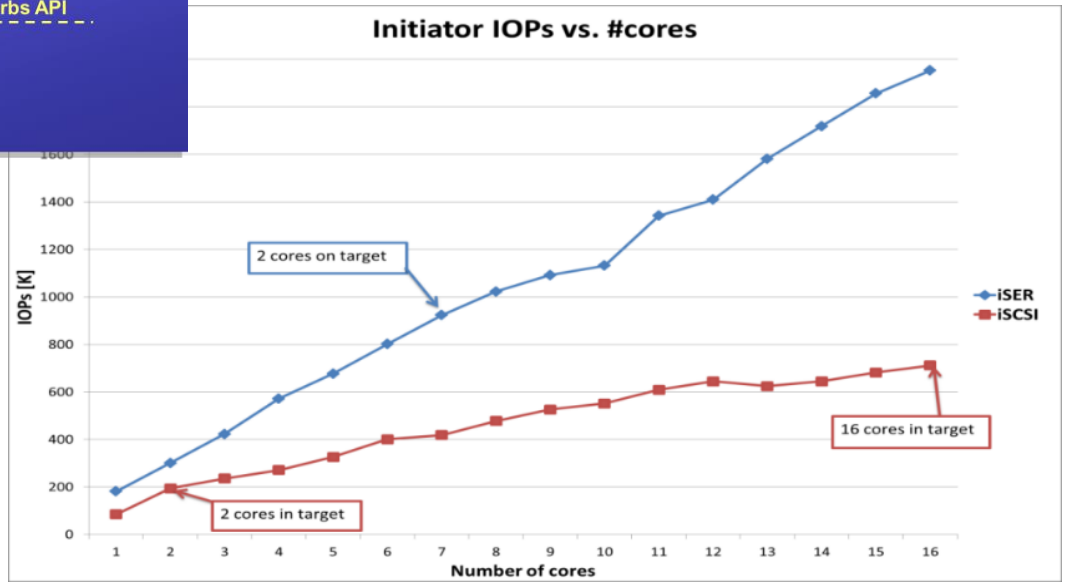


** Source: DataON

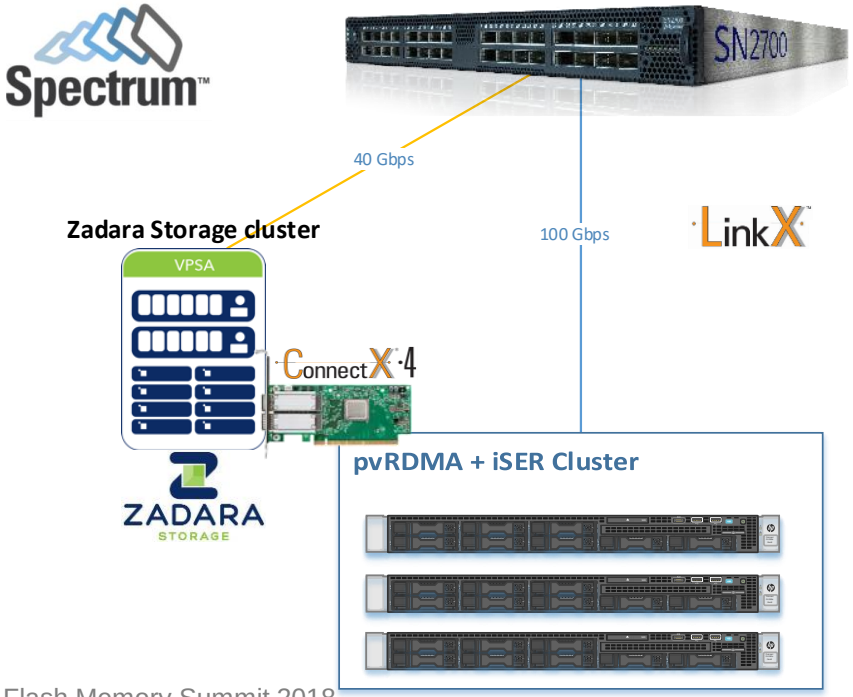
*Windows Server Software-Defined

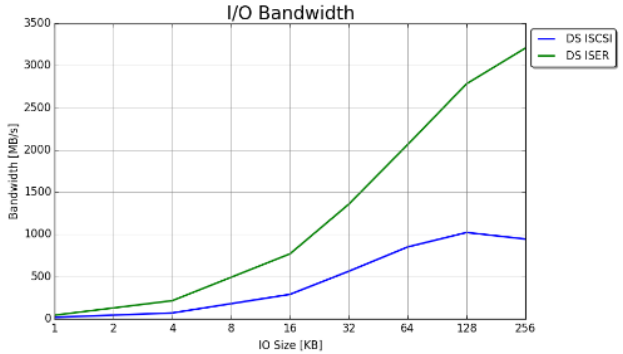
iSER – iSCSI with RDMA Extensions

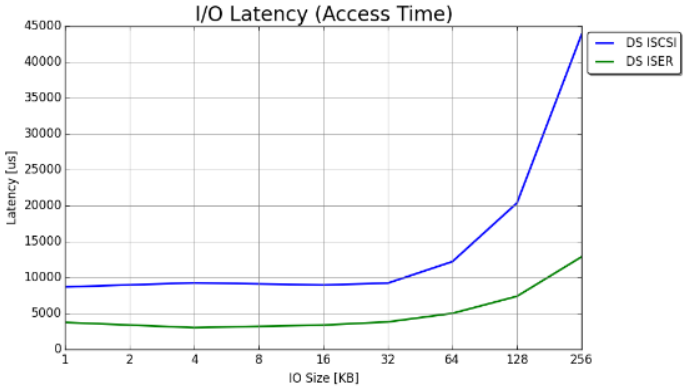




iSER Delivers 3X Higher Efficiency vs. iSCSI







Flash Memory Summit

RDMA enabled Networking Powers Modern Storage Platforms

DataDirect NETWORKS

dataON DataONSTORAGE

DELL EMC

FUJITSU

Hewlett Packard Enterprise

ORACLE

Microsoft

Micron

NetApp

NIMBUS DATA

IBM xiv tms

SEAGATE

TOSHIBA

TERADATA

Western Digital

VIRIDENT

Higher Performance, Higher Efficiency and Higher Scalability

Peter Onufryk

- Peter is a Fellow in the Data Center Solutions Business Unit. where he is responsible for architecture and validation of storage products. He received a Ph.D. in Electrical and Computer Engineering from Rutgers University, has been granted over 40 patents

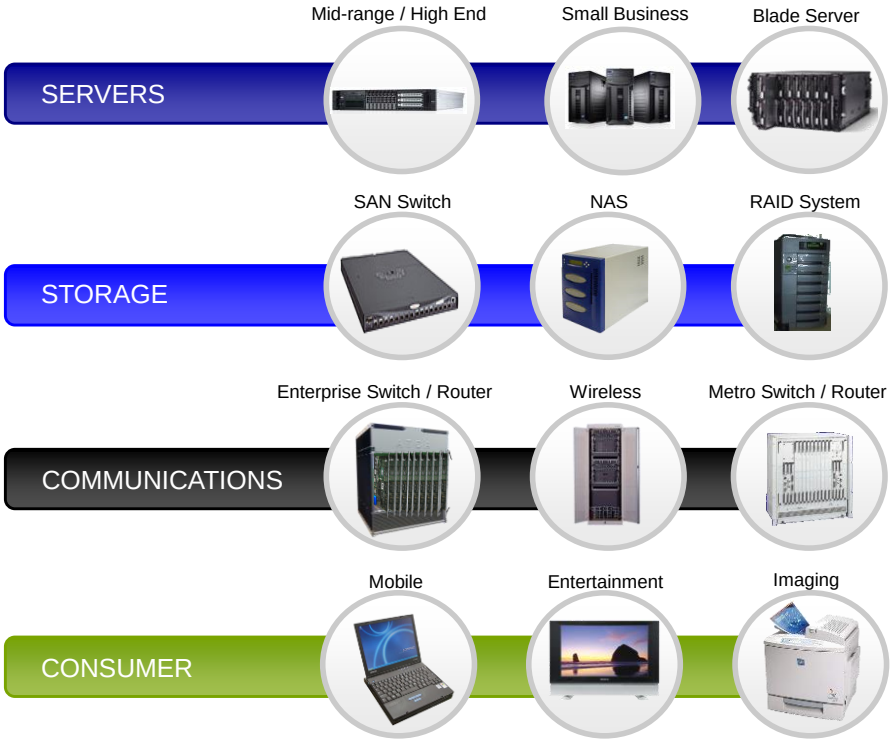
NVM PCIe® Networked ~~Flash~~ Storage

Peter Onufryk

Microsemi Corporation

PCI Express® (PCIe®)

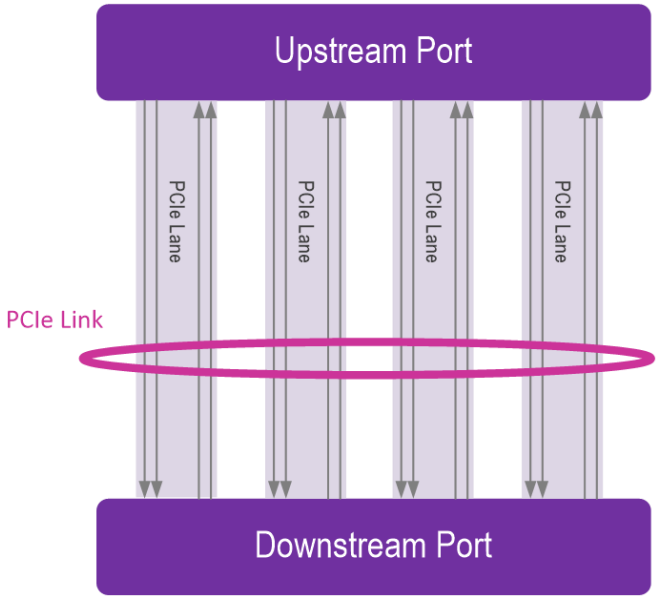
- Specification defined by PCI-SIG®
  - www.pcisig.com
- Packet-based protocol over serial links
  - Software compatible with PCI and PCI-X
  - Reliable, in-order packet transfer
- High performance and scalable from consumer to Enterprise
  - Scalable link speed (2.5 GT/s, 5.0 GT/s, 8.0 GT/s, 16 GT/s, 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

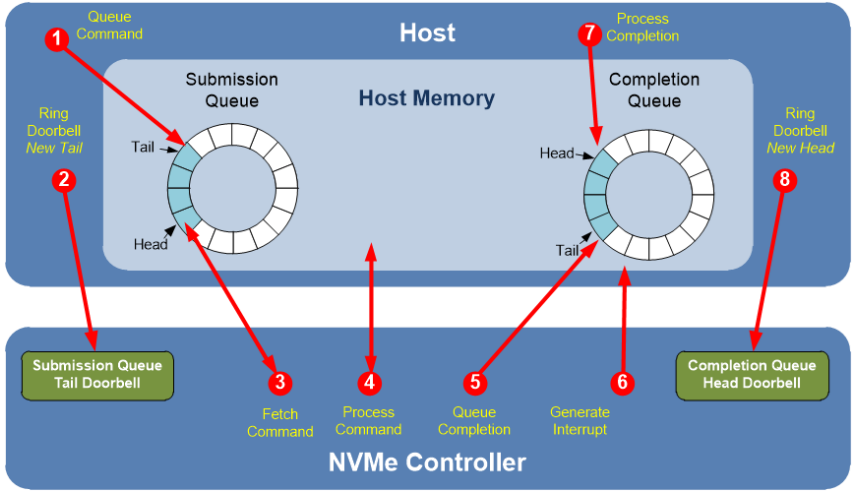
Generation	Raw Bit Rate	Bandwidth Per Lane Each Direction	Total x16 Link Bandwidth
Gen 1*	2.5 GT/s	~ 250 MB/s	~ 8 GB/s
Gen 2*	5.0 GT/s	~500 MB/s	~16 GB/s
Gen 3*	8 GT/s	~ 1 GB/s	~ 32 GB/s
Gen 4	16 GT/s	~ 2 GB/s	~ 64 GB/s
Gen 5	32 GT/s	~4 GB/s	~128 GB/s



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

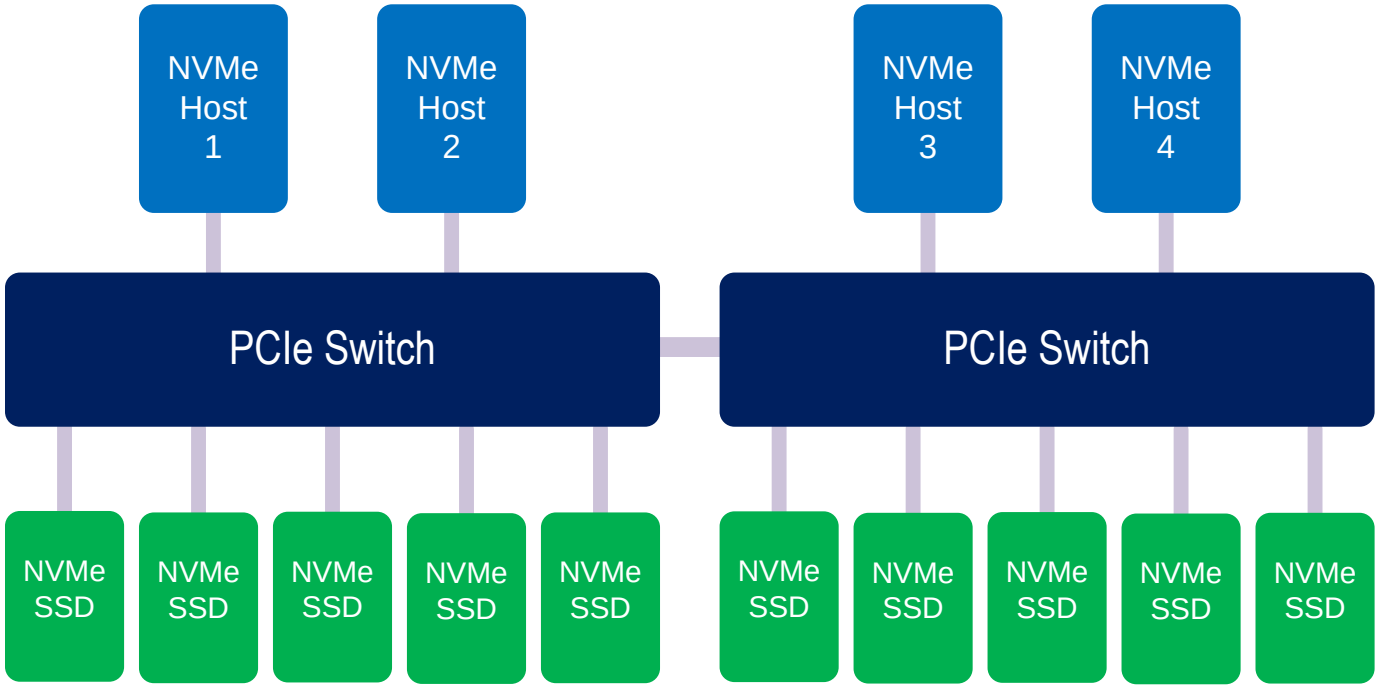


PCIe SSD = NVMe SSD

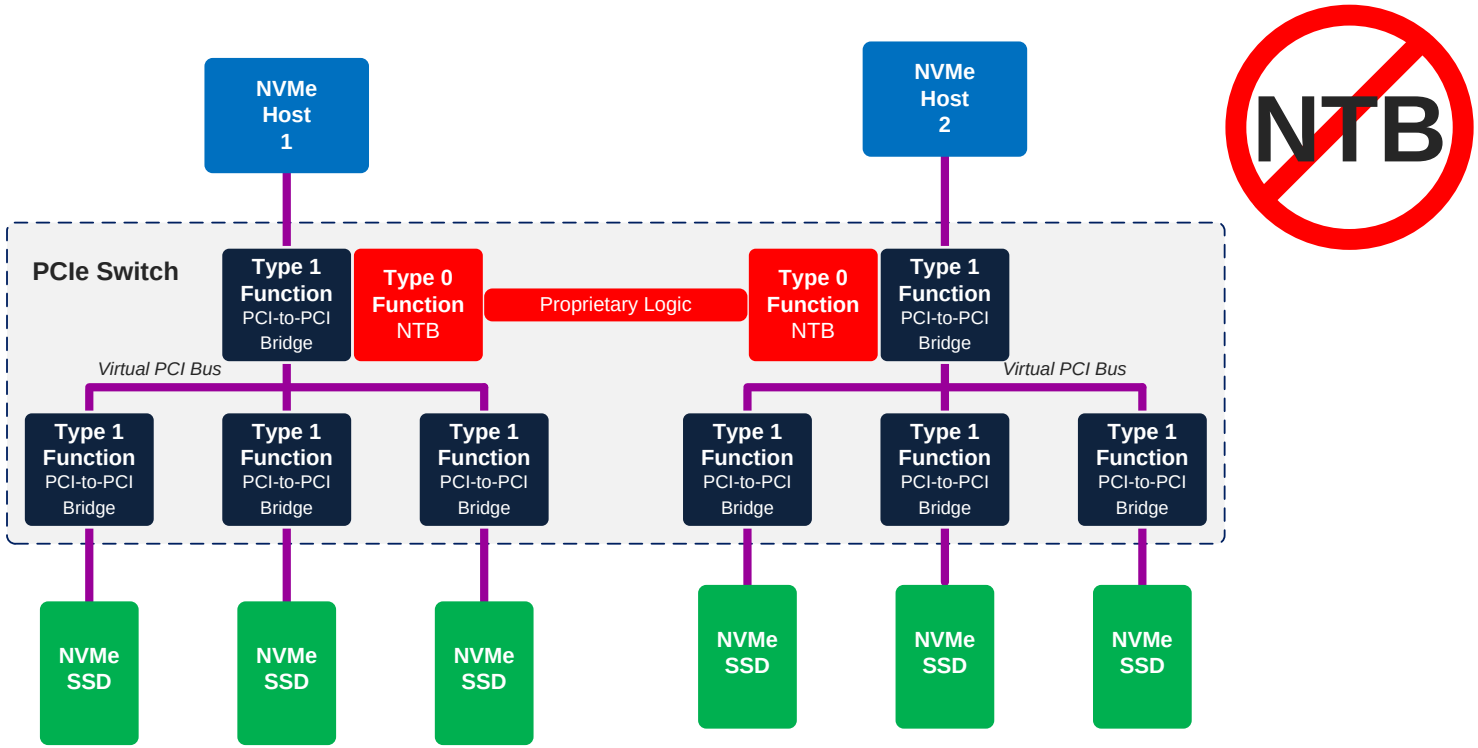
Ideal NVM Fabric

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

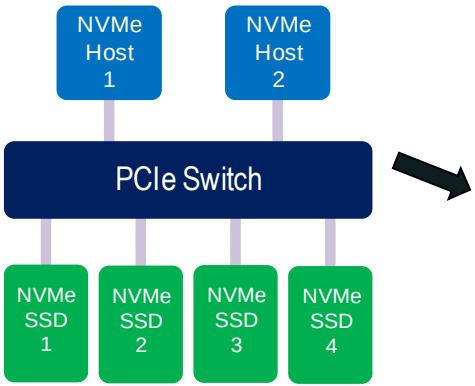
PCIe Fabric


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

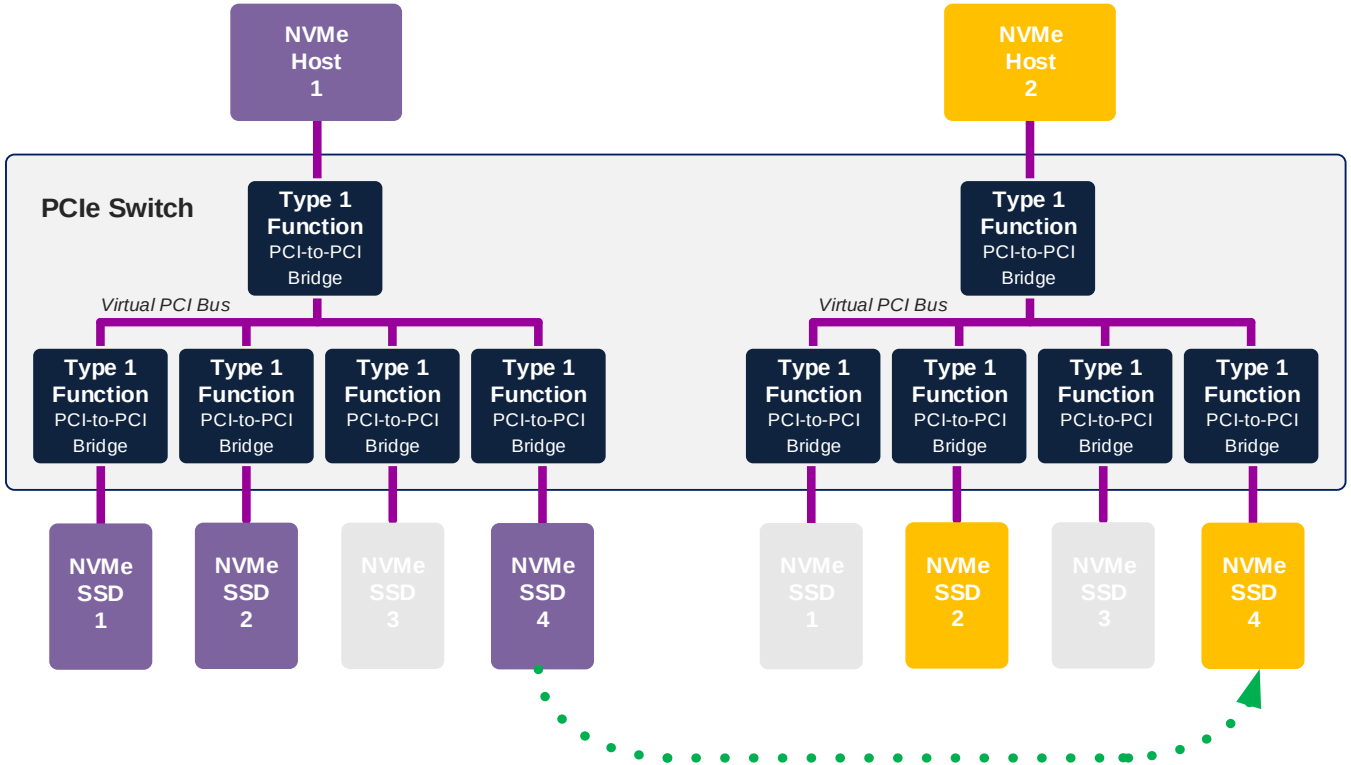
Non-Transparent Bridging (NTB)



Dynamic Partitioning

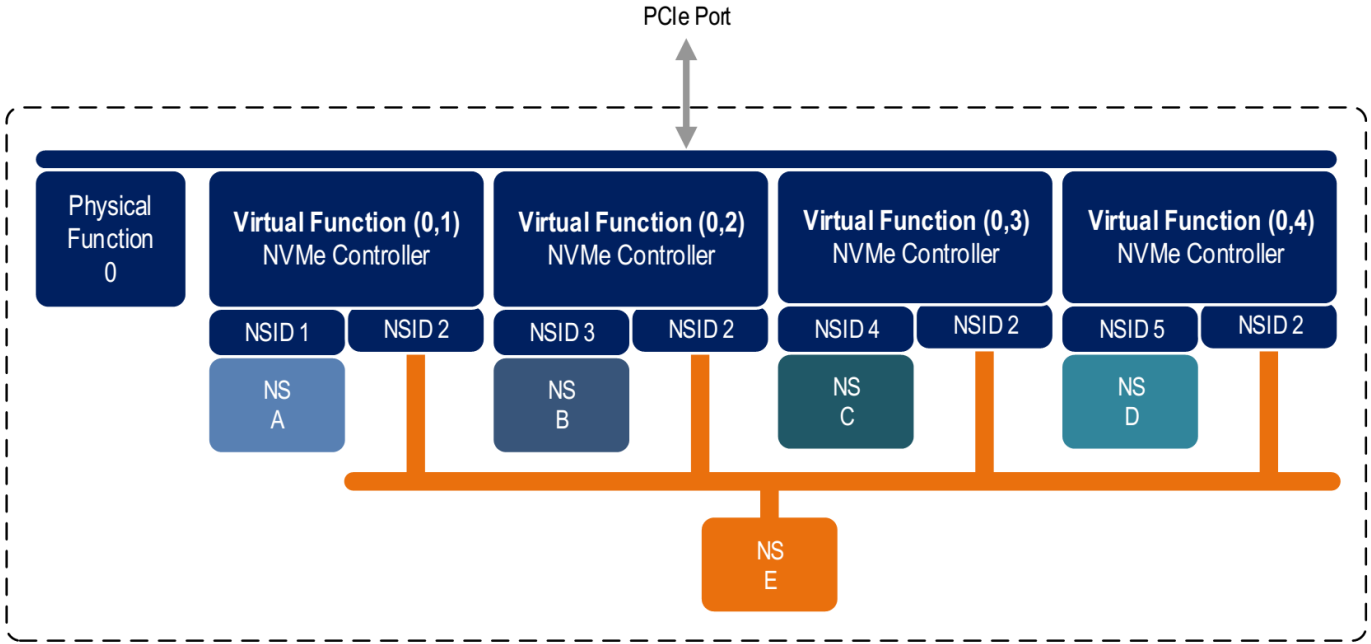


Physical View

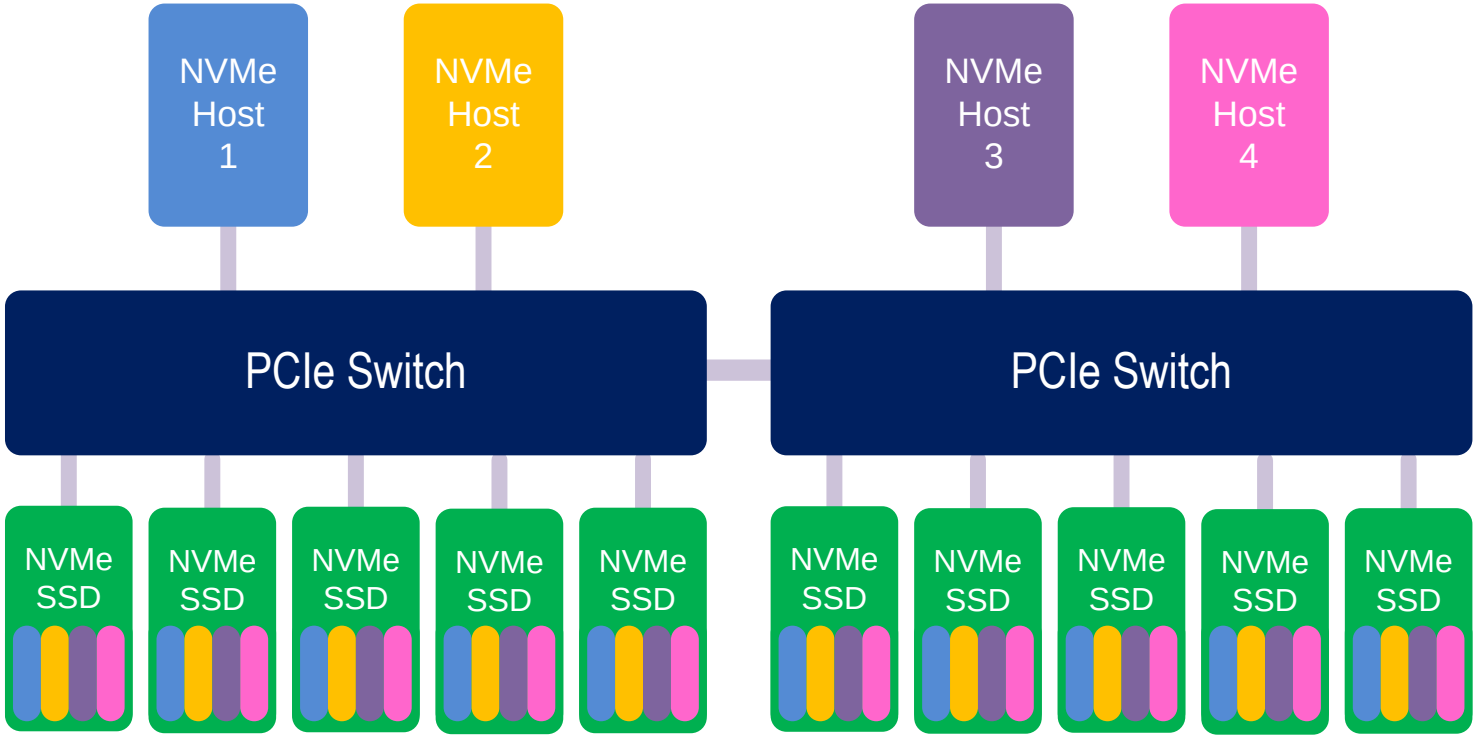


Functional View

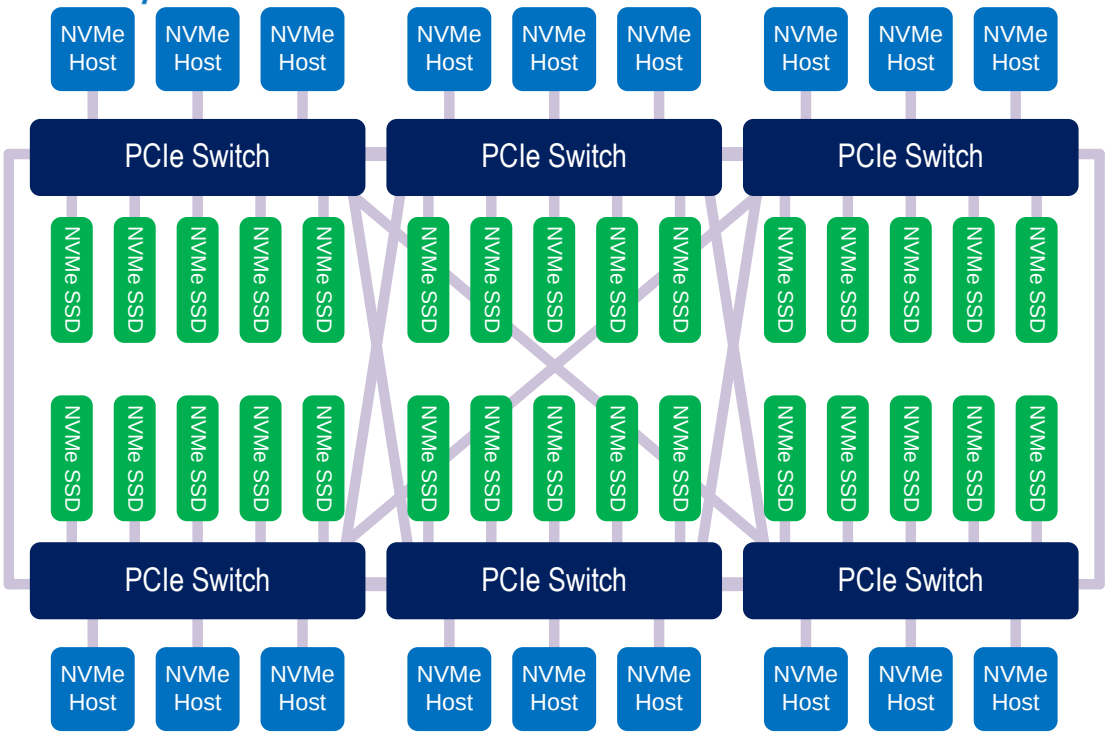
NVMe SR-IOV


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

Multi-Host I/O Sharing



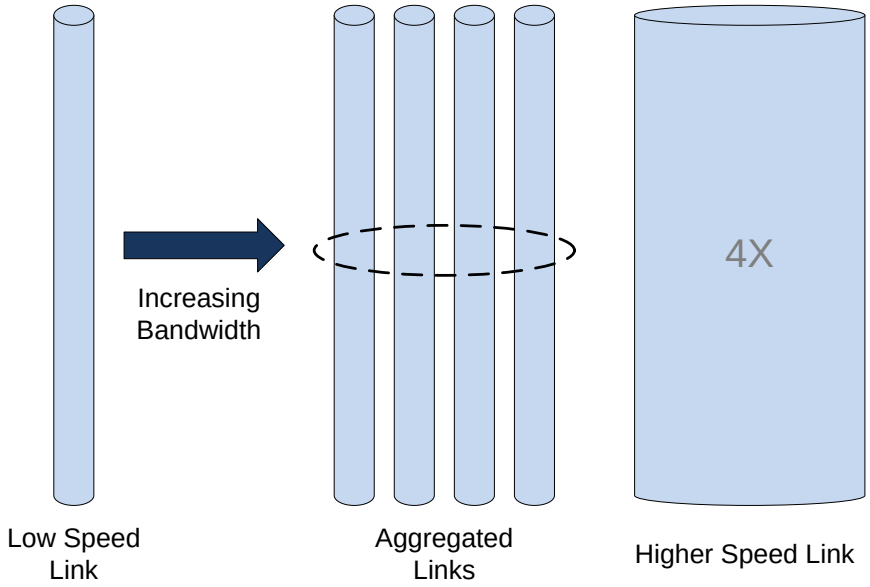
PCIe Fabric



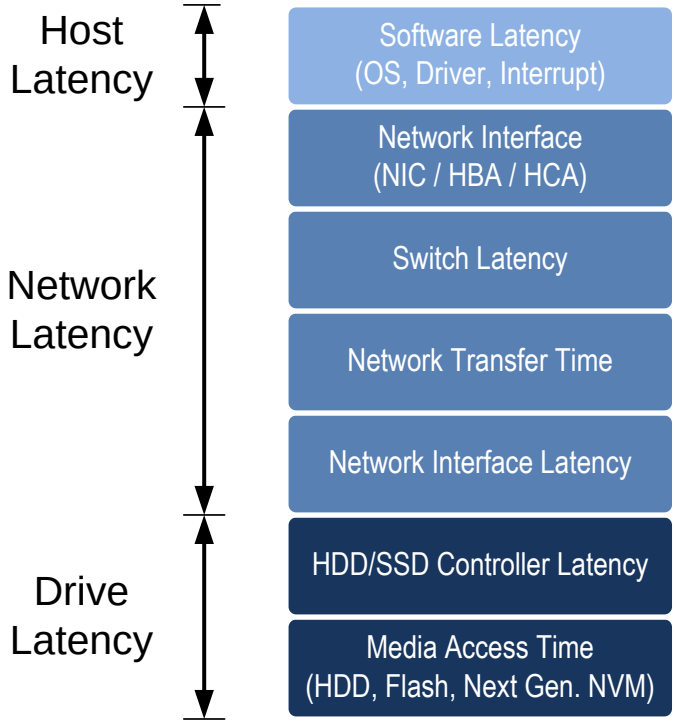
- Storage Functions
  - Dynamic partitioning (drive-to-host mapping)
  - NVMe shared I/O (shared storage)
  - Ability to share other storage (SAS/SATA)
- Host-to-Host Communications
  - RDMA
  - Ethernet emulation
- Manageability
  - NVMe controller-to-host mapping
  - PCIe path selection
  - NVMe management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

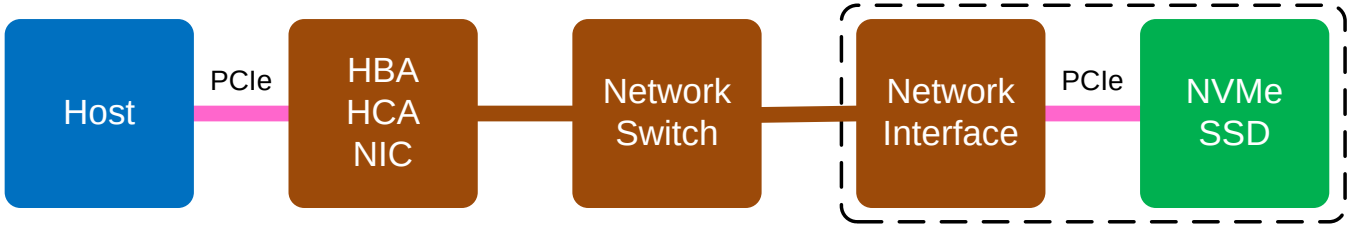


Latency



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

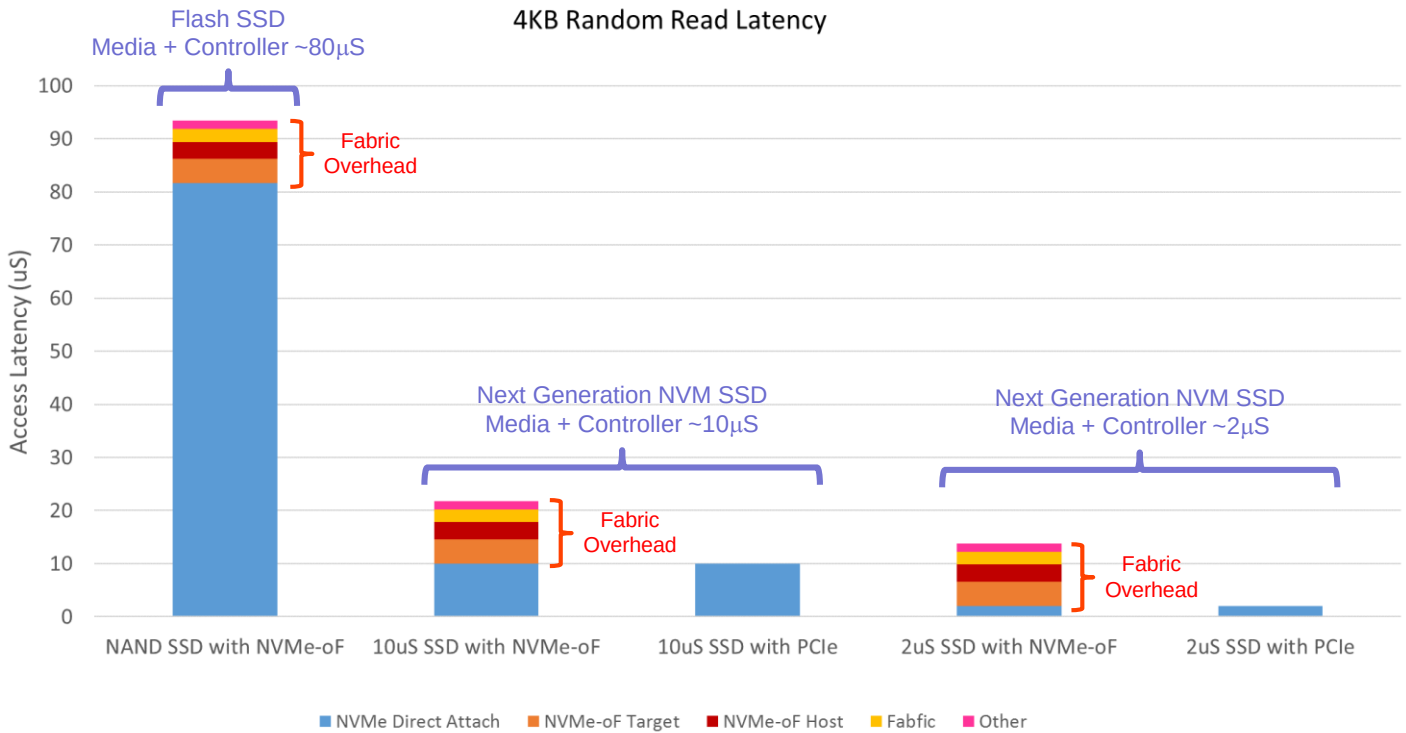
The PCIe Advantage



Other Flash Storage Networks

PCIe Fabric

The PCIe Latency Advantage

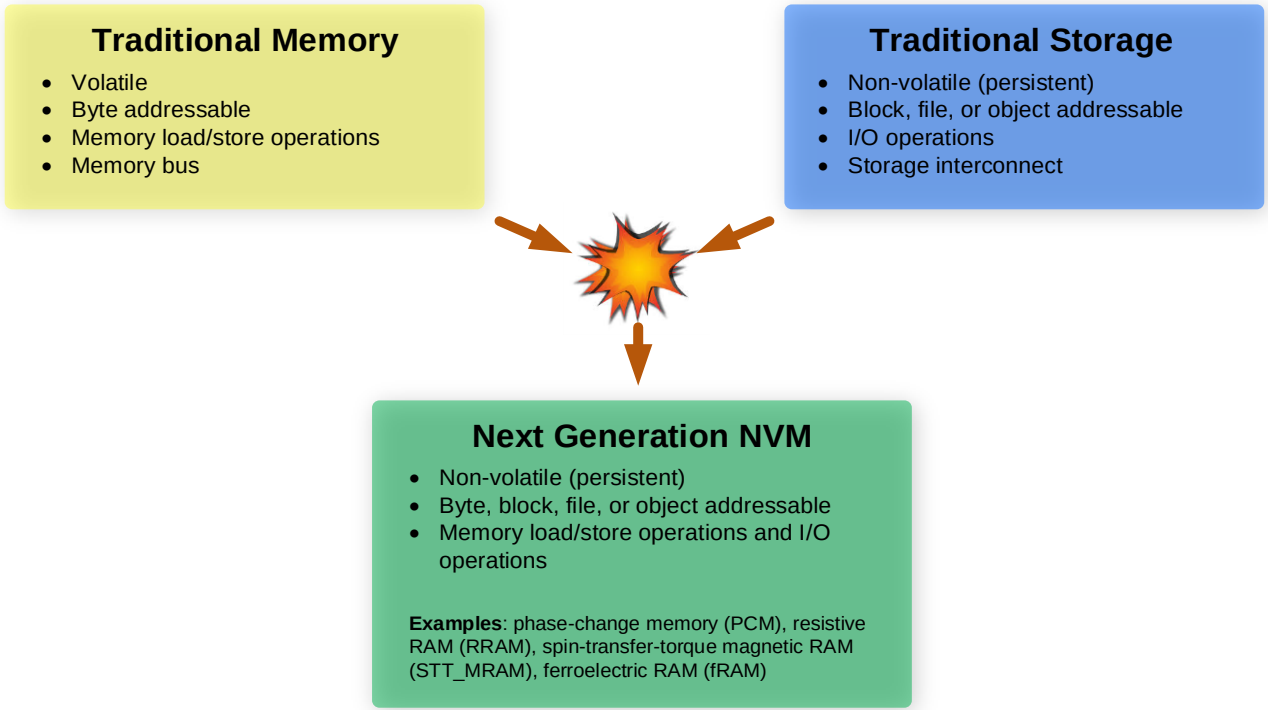


Latency data from Z. Guz et al., "NVMe-over-Fabrics Performance Characterization and the Path to Low-Overhead Flash Disaggregation" in SYSTOR '17

PCIe Fabric Characteristics

Property	Ideal Characteristic	PCIe Fabric	Notes
Cost	Free	Low	• PCIe built into virtually all hosts and NVMe drives
Complexity	Low	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)

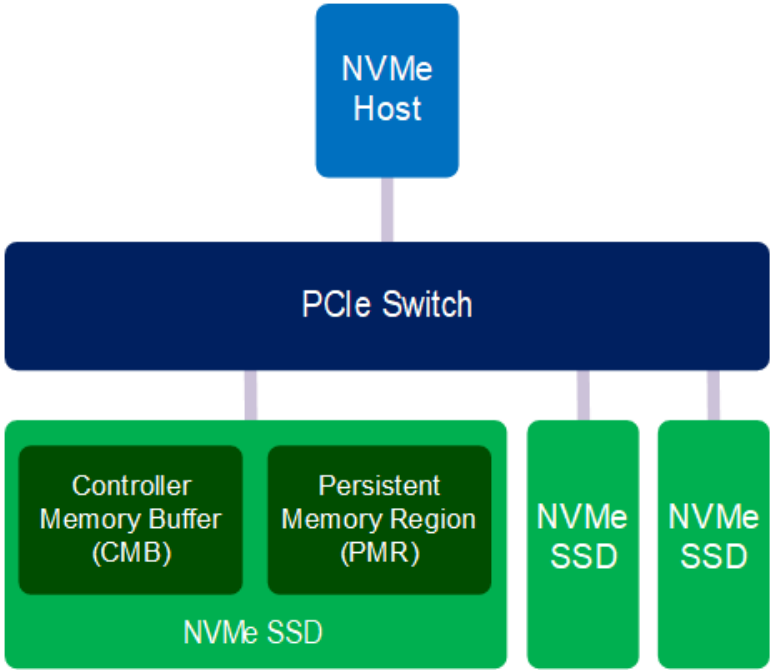
Persistent Memory & Next Gen. NVM



NVMe and Memory Operations

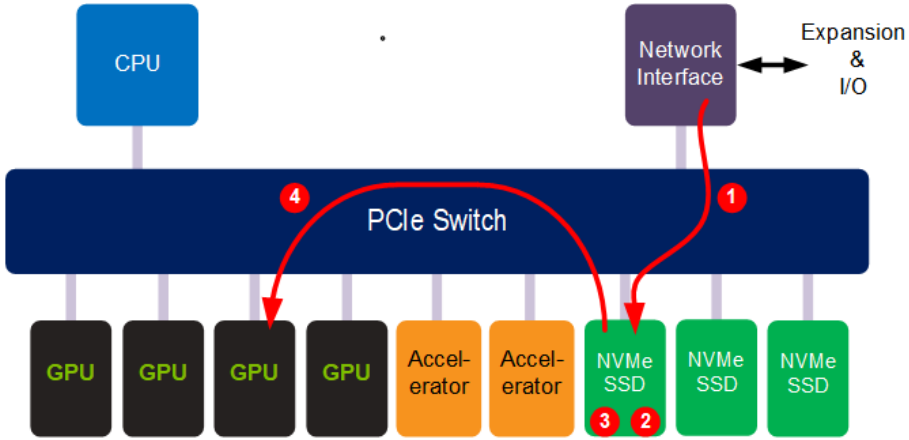
- Controller Memory Buffer (CMB)
  - PCI memory space exposed to host (byte addressable)
  - May be used to store commands & data
  - Contents **do not** persist across power cycles and resets

- 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 allow direct network to storage and accelerator to storage communications



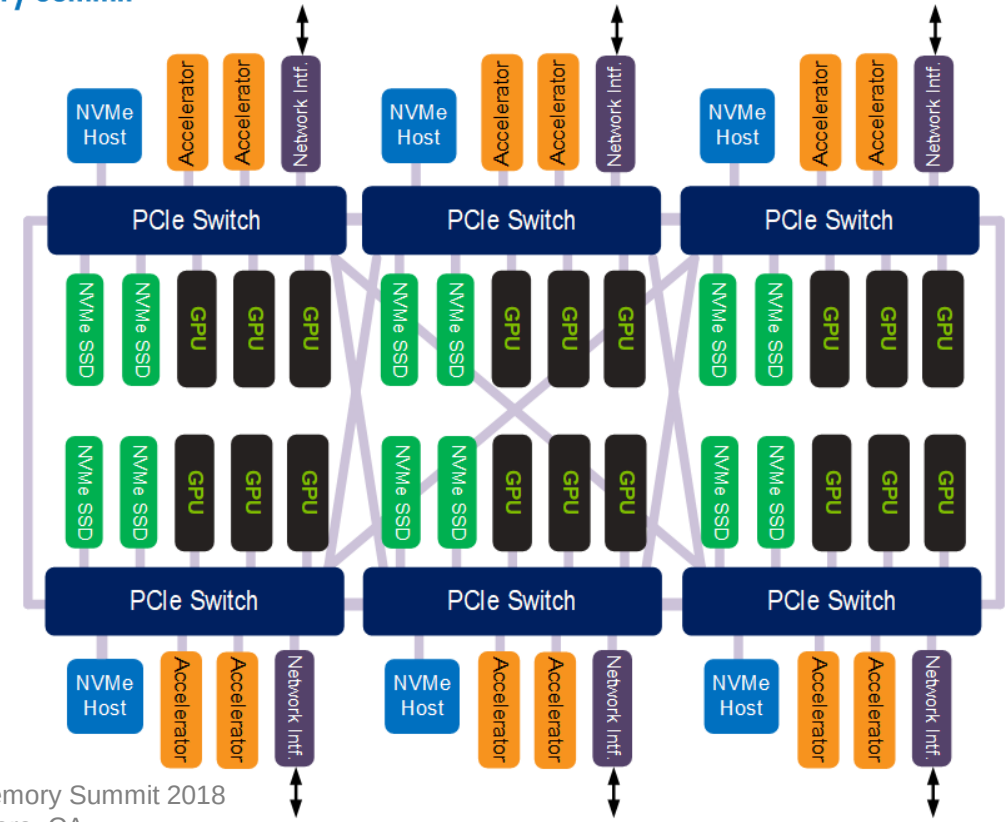
Example:

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