SDC
STORAGE DEVELOPER CONFERENCE
SNIA ■ SANTA CLARA, 2015

Nonvolatile Memory (NVM), Four Trends in the Modern Data Center, and the Implications for the Design of Next Generation Distributed Storage Platforms

David Cohen, System Architect, Intel
Brian Hausauer, Hardware Architect, Intel

INFINIBAND℠
TRADE ASSOCIATION

Abstract

There are four trends unfolding simultaneously in the modern Data Center: (i) Increasing Performance of Network Bandwidth, (ii) Storage Media approaching the performance of DRAM, (iii) OSVs optimizing the code path of their storage stacks, and (iv) single processor/core performance remains roughly flat. A direct result of these trends is that application/workloads and the storage resources they consume are increasingly distributed and virtualized. This, in turn, is making Onload/Offload and RDMA capabilities a required feature/function of distributed storage platforms. In this talk we will discuss these trends and their implications on the design of distributed storage platforms.

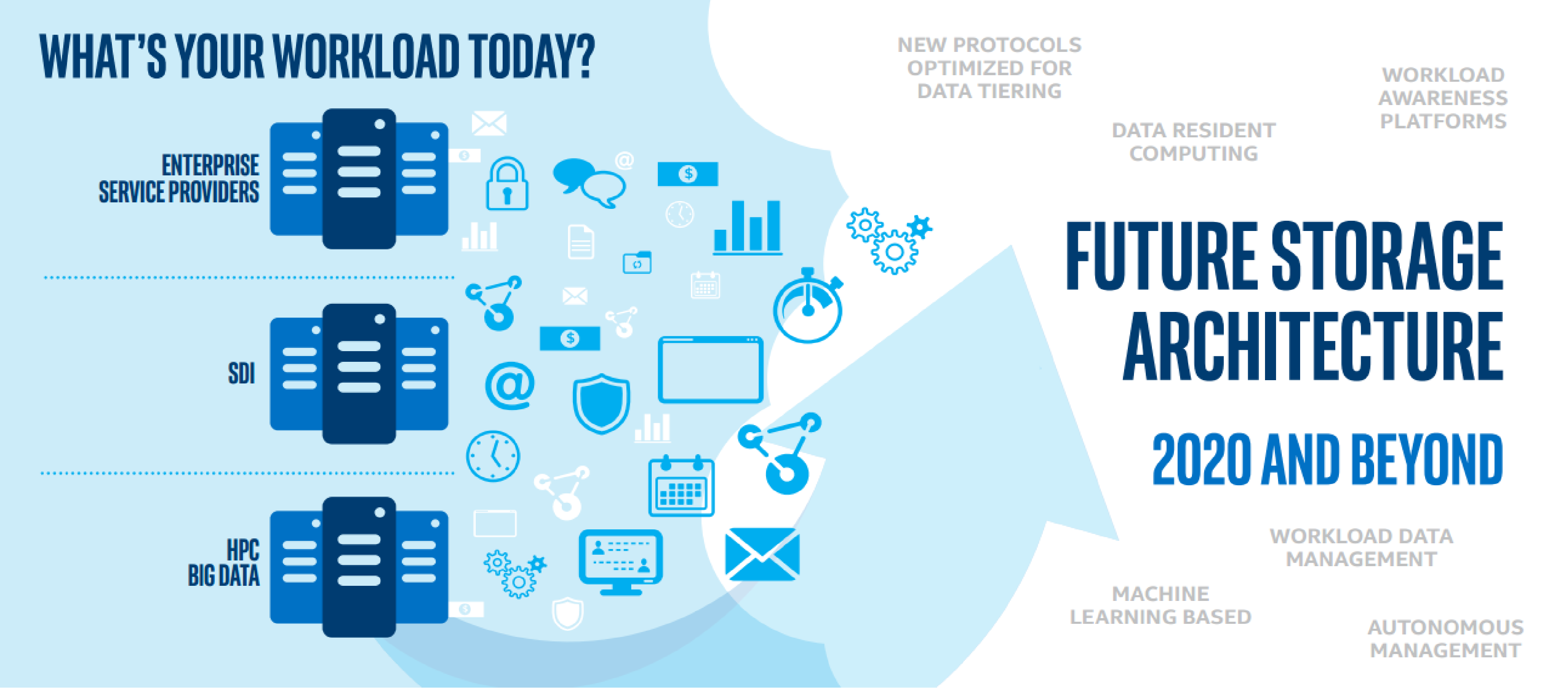
Learning Objectives

- Highlight the four trends unfolding in the data center
- Elaborate on the implication of these trends on design of modern distributed storage platforms
- Provide details on how onload/offload mechanisms and RDMA become feature/function requirements for these platforms in the near-future

The Emergence of Cloud Storage

A new storage architecture has emerged in the modern data center that is predicated on data-center-wide connectivity to support application mobility.

WHERE STORAGE IS HEADED?



WHAT'S YOUR WORKLOAD TODAY?

ENTERPRISE SERVICE PROVIDERS

SDI

HPC BIG DATA

NEW PROTOCOLS OPTIMIZED FOR DATA TIERING

WORKLOAD AWARENESS PLATFORMS

DATA RESIDENT COMPUTING

FUTURE STORAGE ARCHITECTURE

2020 AND BEYOND

WORKLOAD DATA MANAGEMENT

MACHINE LEARNING BASED

AUTONOMOUS MANAGEMENT

Architectural Convergence on the Cloud Storage Platform

Application Mobility & Scale Requires Data Center-Wide Connectivity

1. **Disaggregation : Enables Mobility**
   - Disaggregated (i.e., horizontally-scaled) Compute and Storage has delivered massive increases in capacity
2. **Job Scheduling : Increases Efficiency**
   - Realizing the benefits of these increases in capacity has been predicated on increasingly sophisticated scheduling and job placement
3. **System Balance : Optimization function**
   - The move to scale-out networking has been key to delivering sufficient end-to-end bandwidth to not stall computations (i.e., maintain system balance)

Disaggregation is about Application Mobility and Scale

Disaggregation to Enable Application Deployment Anytime/Anywhere

1. A modern Data Center is designed so that ***an application can be deployed anywhere/anytime***. This is achieved by deploying services:
   - ❒ Disaggregated (i.e., horizontally-scaled) Compute and Storage has delivered massive increases in capacity

2. **Job Scheduling : Increases Efficiency**
   - ❒ Realizing the benefits of these increases in capacity has been predicated on increasingly sophisticated scheduling and job placement

3. **System Balance : Optimization function**
   - ❒ The move to scale-out networking has been key to delivering sufficient end-to-end bandwidth to not stall computations (i.e., maintain system balance)

Application Deployment Anytime/Anywhere Requires Scale-Out, Server-based Storage

1. Does Disaggregation (as defined above) mean that Compute and Storage services run on different hardware?

   No. IaaS, PaaS, and SaaS based applications consume storage services indirectly via IP-based networking services. This means the application can be running over the same hardware/servers (aka “Hyperconverged”) or on different servers.

2. Can Disaggregation be supported by traditional, external storage appliances (aka “scale-up” storage)?

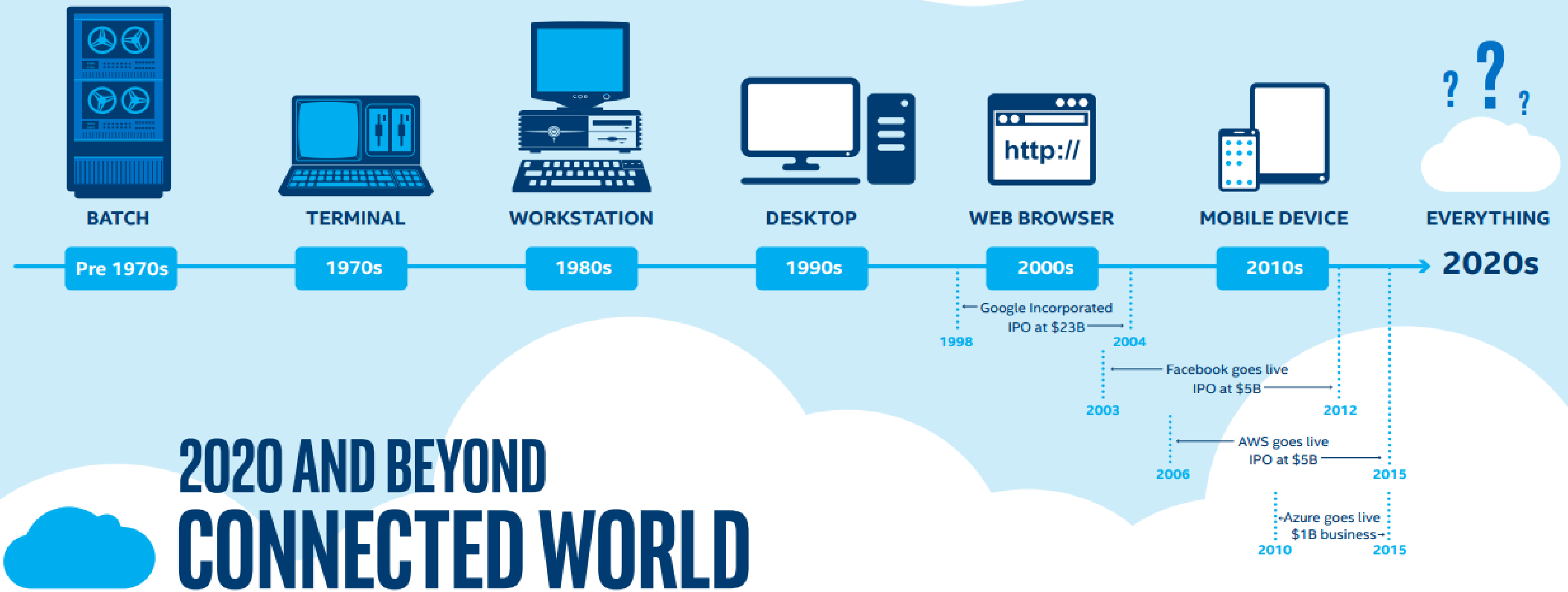
   No, this is frequently referred to as “Converged Infrastructure (as opposed to Hyperconverged).” However, each appliance is a silo with limited capacity and performance. The scope of the deployment is constrained to some number of servers interconnected to the storage appliance via a shared network. If the appliance runs out of capacity (or performance) data (and/or applications) need to be migrated to a new deployment of servers/storage-appliance. This significantly hampers the design goal of enabling “an application can be deployed anywhere/anytime.”

3. How does Server-base Storage (SBS) address this constraint?

   SBS delivers a storage service via the IP-network. While SBS supports a model identical to the storage appliance, the alternative scale-out model runs the storage service over many physical servers. Capacity and performance is scale by adding servers over which to run the service. In the case of hyperconverged infrastructure the application and the storage service operate over the same servers.

Who is driving this innovation?

EMERGENCE OF THE CLOUD MARKET



BATCH — Pre 1970s

TERMINAL — 1970s

WORKSTATION — 1980s

DESKTOP — 1990s

WEB BROWSER — 2000s

MOBILE DEVICE — 2010s

EVERYTHING — 2020s

1998 Google Incorporated IPO at $23B 2004

2003 Facebook goes live IPO at $5B 2012

2006 AWS goes live IPO at $5B 2015

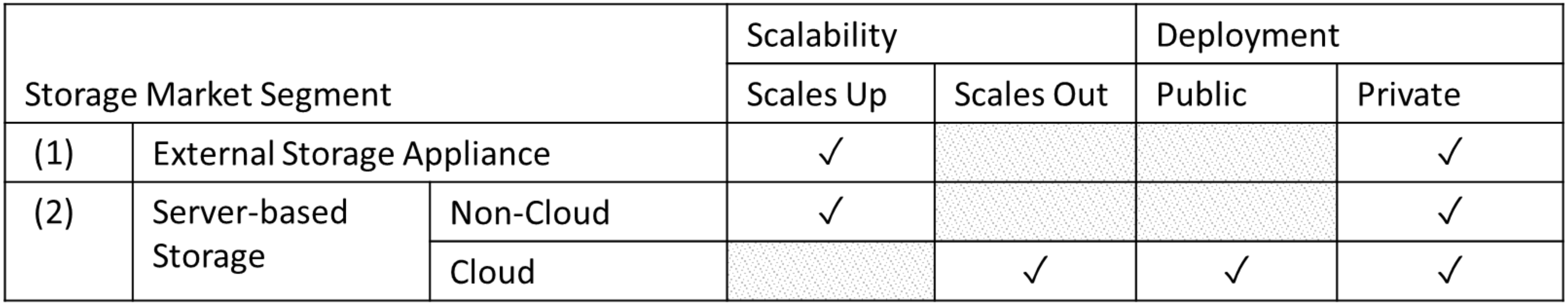
2010 Azure goes live $1B business 2015

2020 AND BEYOND
CONNECTED WORLD

“Cloud” based Storage Scales-Out to support Application Mobility

Server-based Storage

a simple taxonomy



Storage Market Segment			Scalability		Deployment	
			Scales Up	Scales Out	Public	Private
(1)	External Storage Appliance		✓	▨	▨	✓
(2)	Server-based Storage	Non-Cloud	✓	▨	▨	✓
		Cloud	▨	✓	✓	✓

Scaling out for Application Mobility shift the burden to the Network

Four Trends in Modern Data Centers

1. Operating System Vendors (OSV) optimizing the code path of their network and storage stacks

2. Increasing Performance of Network Bandwidth

3. Storage Media approaching the performance of DRAM

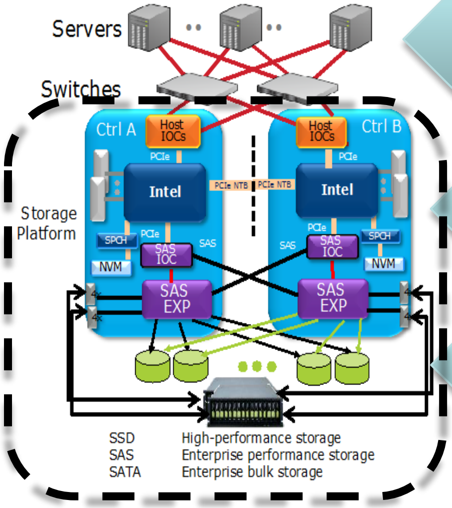
4. Single processor/core performance not increasing at the same rate as network and storage

The Shift from Appliance-based to Cloud-based Storage

The new storage architecture reliance on data-center-wide connectivity is increasingly focused on latency.

Evolving Storage ARCHITECTURE Landscape

1. Traditional Storage Architecture



Over Provisioned Redundant Networks

Replication internal to storage appliance

No SPoF

In Traditional Storage architectures:

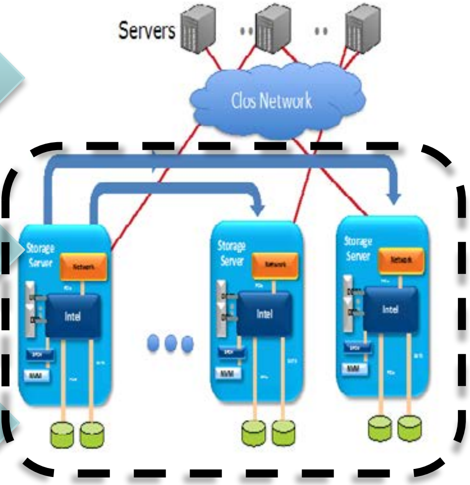
- Analysts predict little to no growth
- Availability is built into appliance because network bandwidth is expensive
- Network is redundant & over provisioned
- Limited in scale and dependent LAN connectivity
- Custom HW & SW
- Chassis is the appliance

2. Cloud Storage Architecture

Flat Networks Resiliency in Clos Fabric

Rack is Failure Domain Data Center **is** Appliance

Direct Attached Storage (DAS)

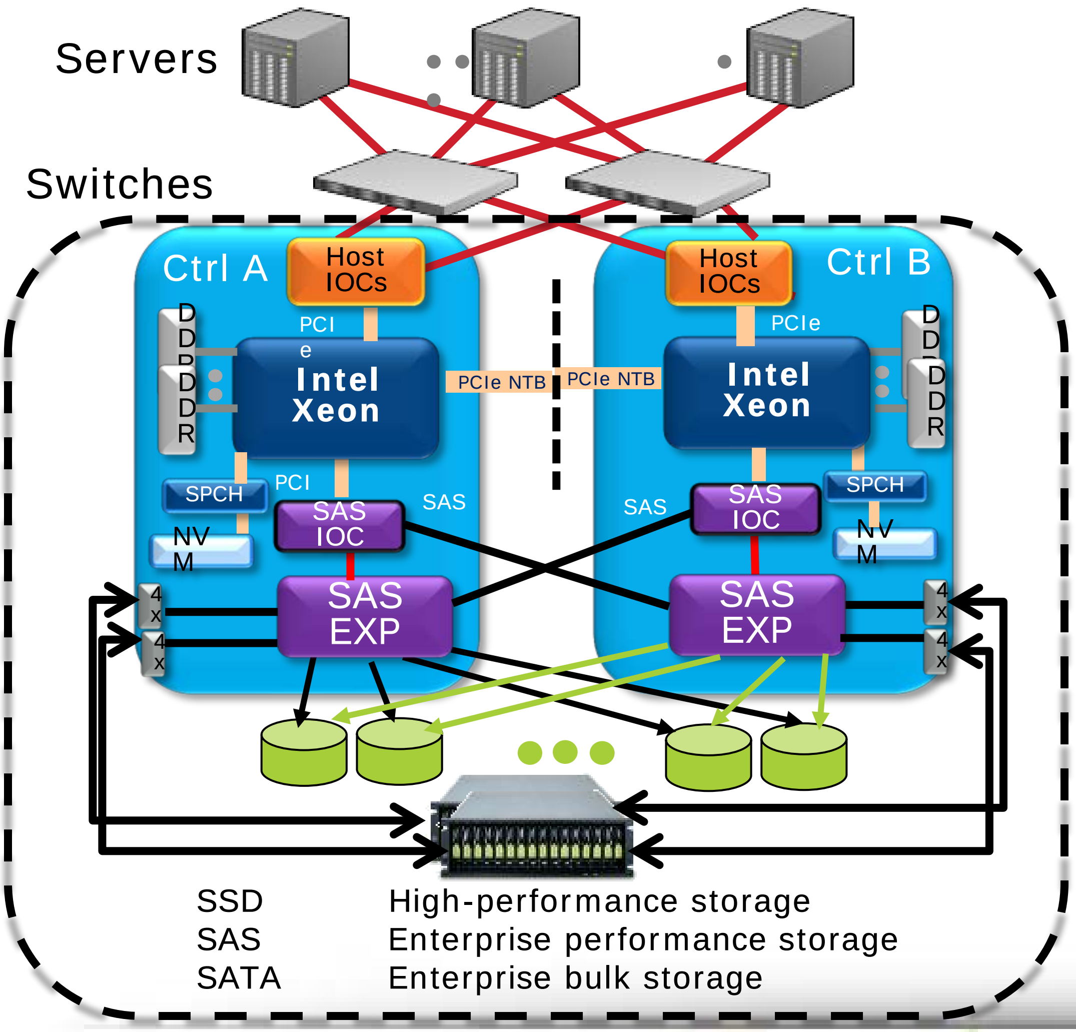


In Cloud Storage architectures:

- Analysts predict substantial growth
- Use of high BW fabric enables consumption of large amounts of NVM
- Per GB costs are order magnitude less by leveraging tiering
- Storage is an application on commodity HW
- Data Center (or ‘zone’) is the appliance

The Traditional and Cloud Storage Architectures Differ on How they Deal with Availability

Traditional Storage Architecture Dataflow for a Write Operation



SSD High-performance storage
SAS Enterprise performance storage
SATA Enterprise bulk storage

1. Write Data from Application arrives; copy (log) placed in non-volatile memory region
2. Data is made durable by replicating in partner non-volatile memory area (no SPoF)
3. Application write is acknowledged, program proceeds
4. Data written at leisure to disk

Enables:

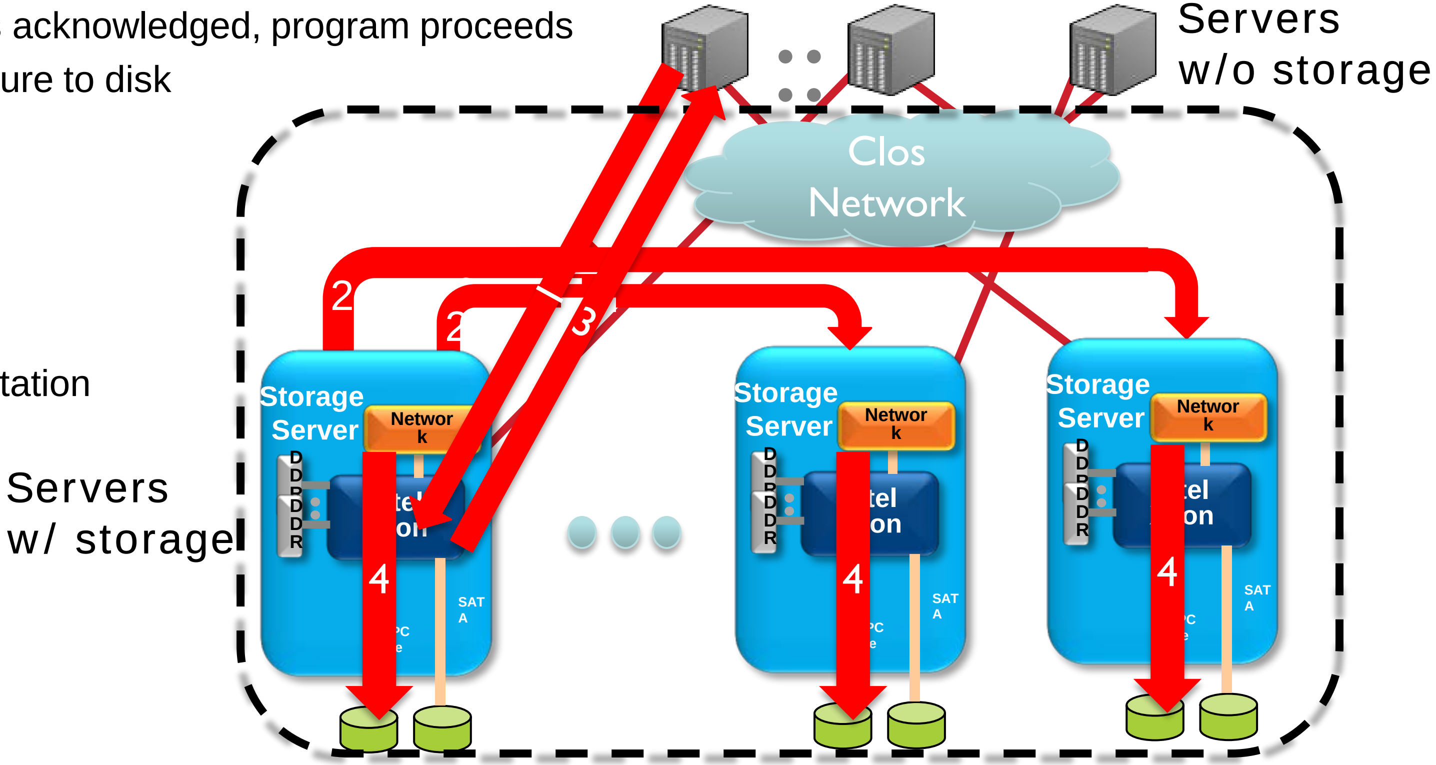
- No SPoF
- Minimal network use
- Minimum response time
- Tiering

Cloud Storage Architecture Dataflow for a Write Operation

1. Write Data from Application arrives; copy (log) placed in non-volatile memory region
2. Data is made durable by replicating in partner non-volatile memory area (no SPoF)
3. Application write is acknowledged, program proceeds
4. Data written at leisure to disk

Enables:

- no SPoF
- greater network use
- Lower cost implementation
- Larger fan out



The Cloud Storage Architecture's Tiering model is also a key differentiator

Storage Tiering

Cloud Storage uses a tiering model based on Hot, Warm, and Cold data. A relatively small amount of higher performance storage is used to service application I/O requests. As data is accessed less frequently it is moved to less expensive media.

Storage Tiering in the Cloud Storage Architecture

Hot Tier (Performance):

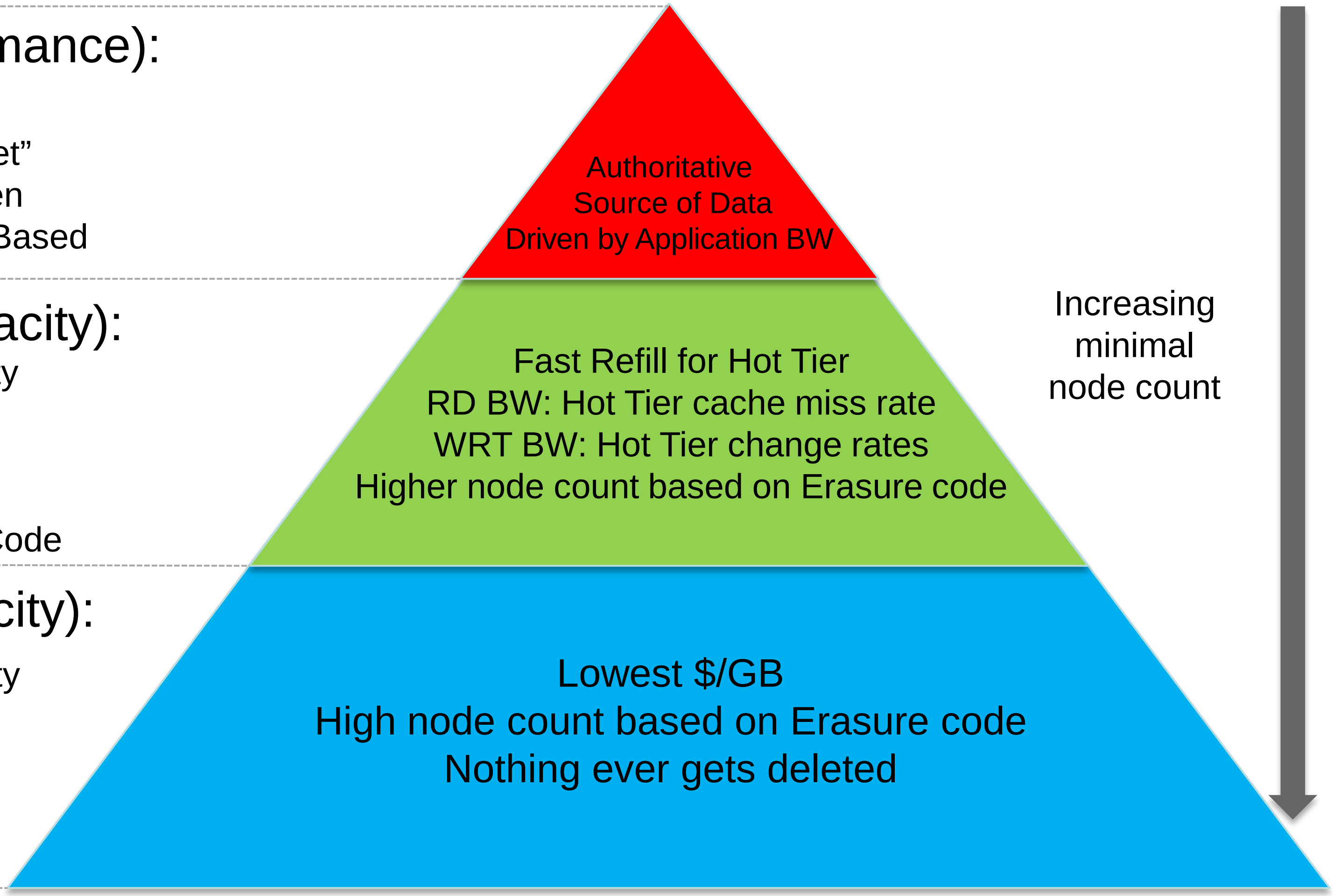
- 5-10% of Capacity
- “Active Working Set”
- Performance Driven
- Local Replication Based

Warm Tier (Capacity):

- 15-30% of Capacity
- “Data < 1 year”
- Cost/Performance Driven
- Medium Erasure Code

Cold Tier (Capacity):

- 60-75% of Capacity
- Cost Driven
- Maximum Erasure Code
- Future Multi-Site



Authoritative
Source of Data
Driven by Application BW

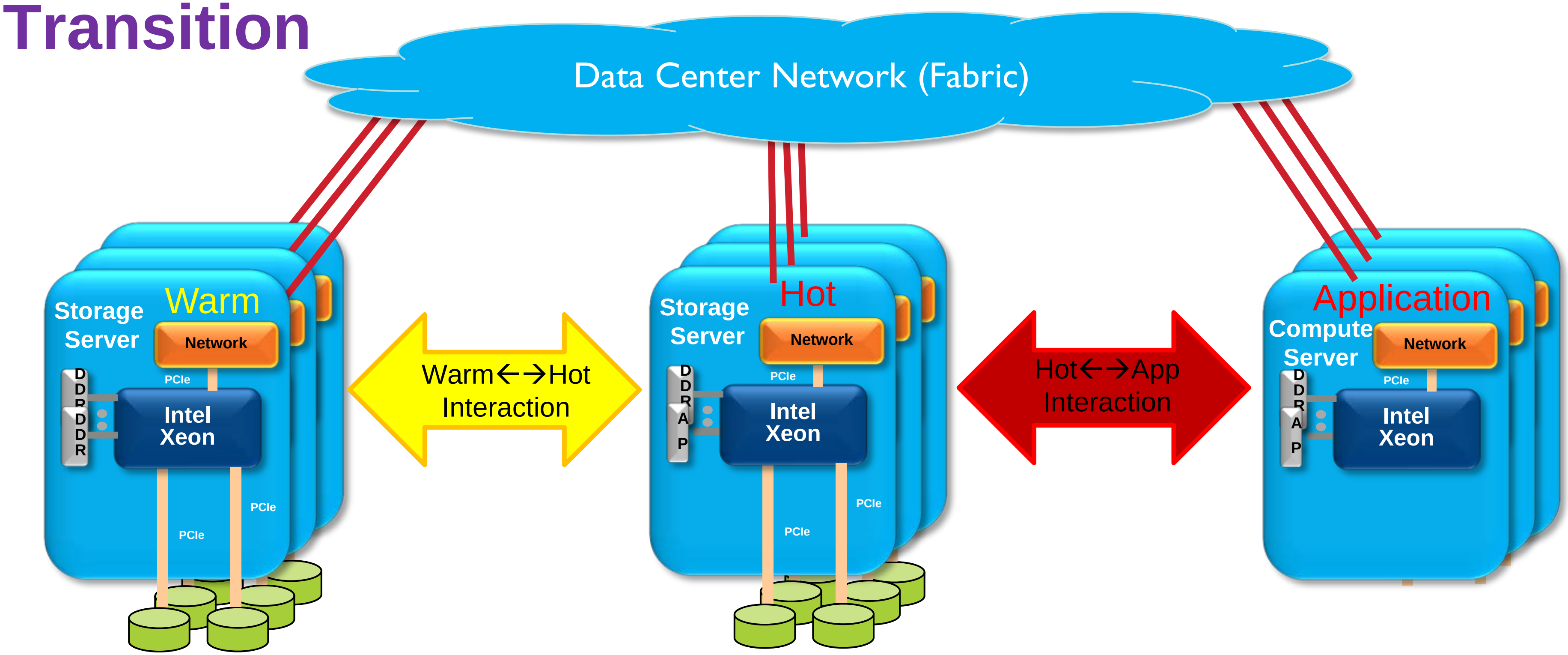
Fast Refill for Hot Tier
RD BW: Hot Tier cache miss rate
WRT BW: Hot Tier change rates
Higher node count based on Erasure code

Lowest $/GB
High node count based on Erasure code
Nothing ever gets deleted

Increasing
minimal
node count

Storage Tiering enables the use of higher performance, more expensive storage media

End-to-End Storage I/O – Positioning the Media Transition



Rotational/SATA → 3D-NAND
(SAS) (NVMe)

Warm Tier

Storage

Hot Tier

Compute Tier

The transition from rotational to solid-state media shifts focus to low latency network I/O

The Shift in Focus to Latency

Workloads in Cloud deployments are concerned per-operation elapsed time (aka “latency”) and the “tail” of the distribution as measured across many of these operations.

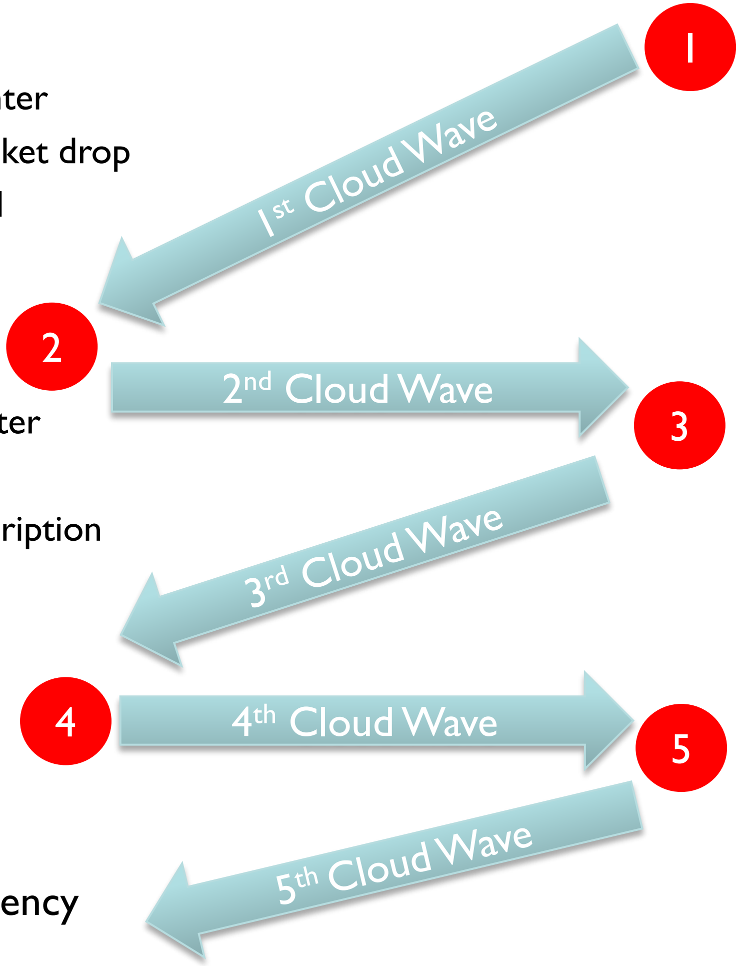
Flash Accelerates the Data Center, Drives Innovation in the Network

- Traditional Network
  - ~10ms across data center
  - Highly Buffered no packet drop
  - Highly Oversubscribed
  - 1Gbs to Host

- Flat Network
  - ~10uS across data center
  - No Buffering
  - Much lower oversubscription
  - 10Gbs to Host

- 25/40/50Gbs to Host

- 100Gbs to Host, Low Latency Messaging, RDMA



- SAS HDD
  - ~200 IOPs @ ~5mS
  - ~100MB/s streaming

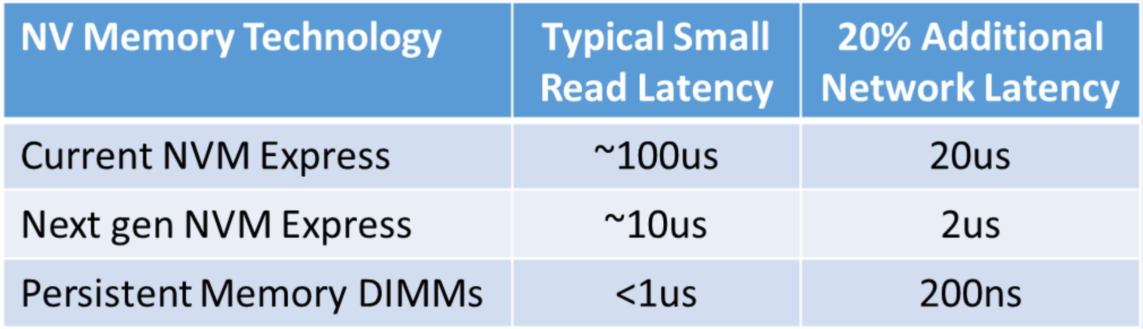
- NVM Express™ SSD
  - ~400,000 IOPs @ ~100uS
  - 2GB/s

- 3D-Xpoint
  - ~ <???> IOPs @ ~ <???>uS
  - <???>GB/s

The next wave of innovations will focus on addressing Latency

Network Latency requirements

- Table shows typical small read latency for several current and next gen NV Memory technologies
- Last column shows a "rule of thumb", that 20% of additional network latency is acceptable
- For many block and file access protocols, the network latency includes separate command, data transfer, and status messages, requiring up to two network round trip times (RTTs).



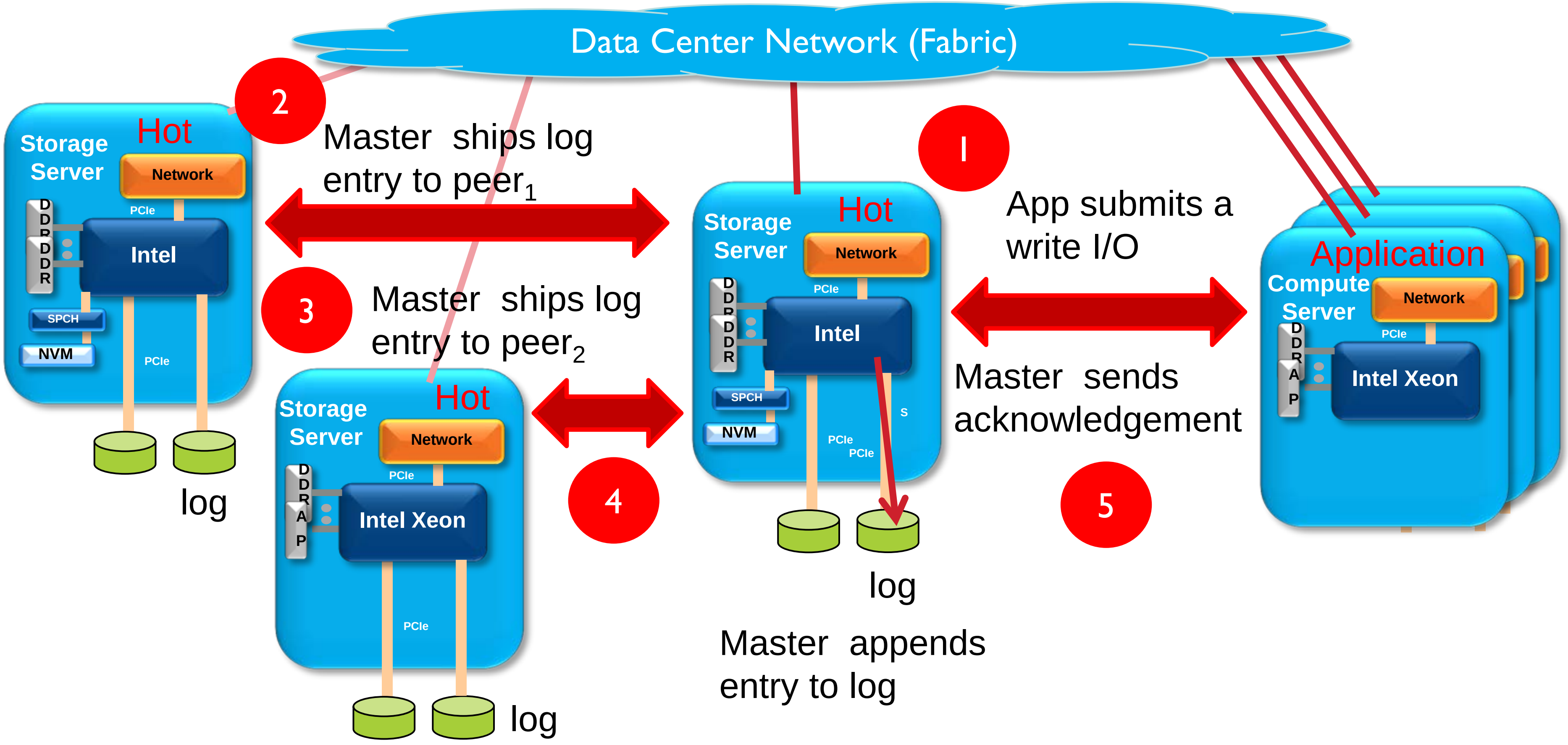
NV Memory Technology	Typical Small Read Latency	20% Additional Network Latency
Current NVM Express	~100us	20us
Next gen NVM Express	~10us	2us
Persistent Memory DIMMs	<1us	200ns

Conclusions

- It is difficult to achieve the 'Next gen NVM Express' network latency goal without RDMA
- It is very difficult to achieve the 'Persistent Memory DIMM' network latency goal with as-is block or file network protocols. A new or enhanced protocol needs:
    - Reduced number of network messages per IOP and max single RTT
    - No per-IOP CPU interaction on the target

Write I/O Operations, Availability, and Distributed Logs

Recall from Slide 14, a Write I/O Operation must be appended to the Master's log along with the logs of at least two peers before an acknowledgement is returned to Application. This is a synchronous operation.



The response time of Write I/O Operations is a challenge due to synchronous replication

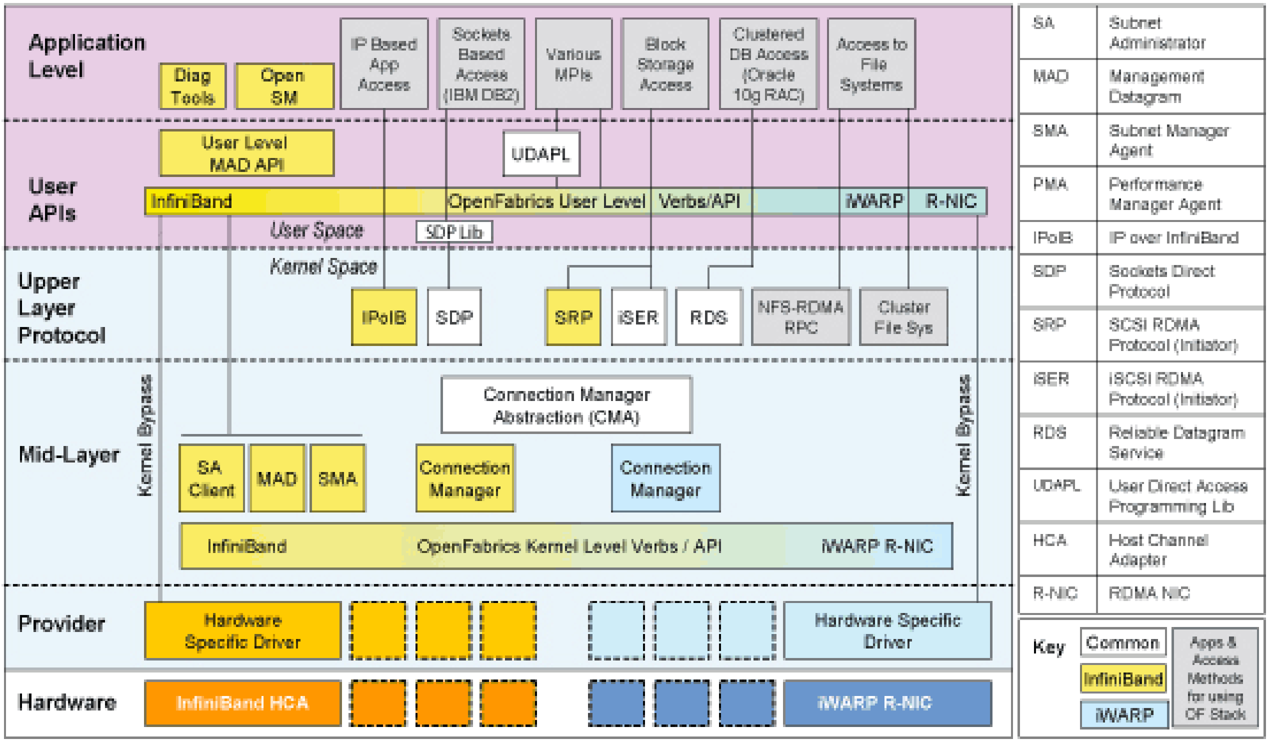
Satisfying the Requirements of Cloud Storage

RDMA and High Performance Storage Networking

Server OSes Already Support the I/O Infrastructure Needed to address Cloud Scale-Out Storage

Example: Linux OpenFabrics Software Stack for RDMA

- Supports a range of block, network filesystem, and distributed parallel filesystem protocols, with both initiator and target implementations
- Uses RDMA for low latency and high message rate
- Supports cloud scale-out storage when used with IP-based RDMA



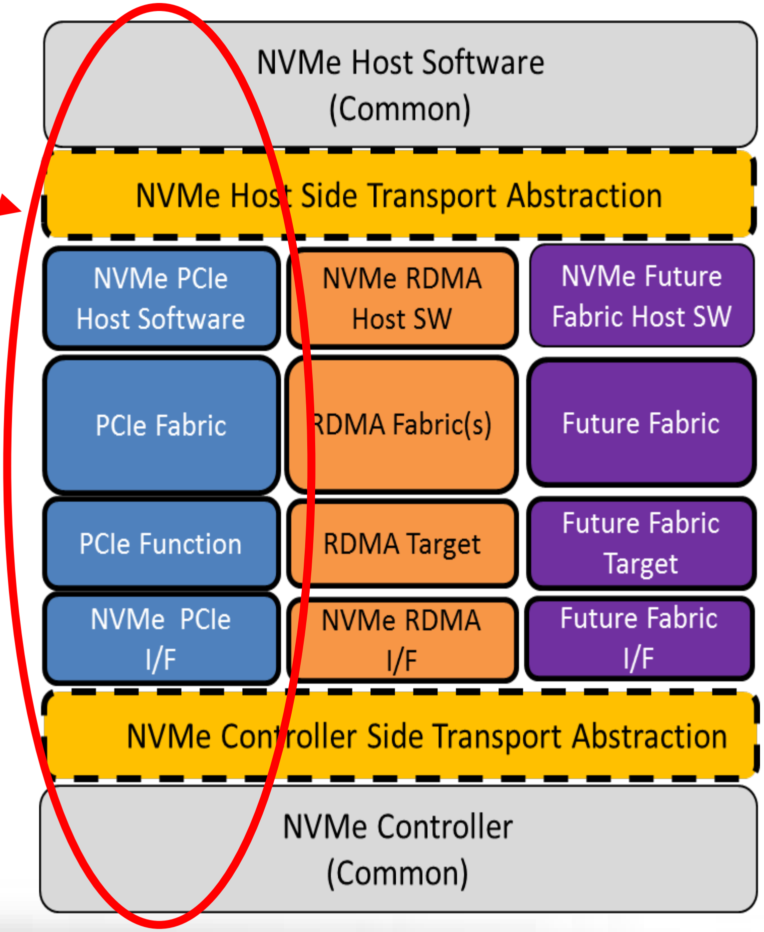
SA	Subnet Administrator
MAD	Management Datagram
SMA	Subnet Manager Agent
PMA	Performance Manager Agent
IPoIB	IP over InfiniBand
SDP	Sockets Direct Protocol
SRP	SCSI RDMA Protocol (Initiator)
iSER	iSCSI RDMA Protocol (Initiator)
RDS	Reliable Datagram Service
UDAPL	User Direct Access Programming Lib
HCA	Host Channel Adapter
R-NIC	RDMA NIC

Protocols are Evolving to Better Address Cloud Storage Opportunities

Example: NVMe over Fabrics block storage

Evolving from a PCIe Fabric-connected solution…

- Very low latency and high message rate
- But with scale-out limitations due to PCIe fabric



Protocols are Evolving to Better Address Cloud Storage Opportunities

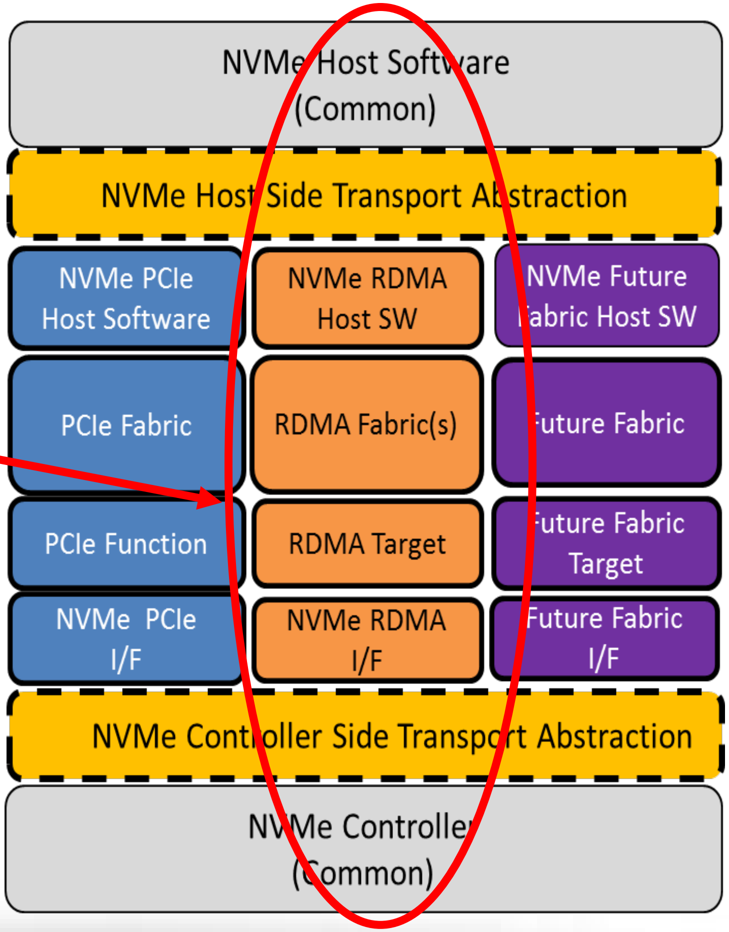
Example: NVMe over Fabrics block storage

Evolving from a PCIe Fabric-connected solution…

- Very low latency and high message rate
- But with scale-out limitations due to PCIe fabric

…to include an RDMA Fabric-connected solution

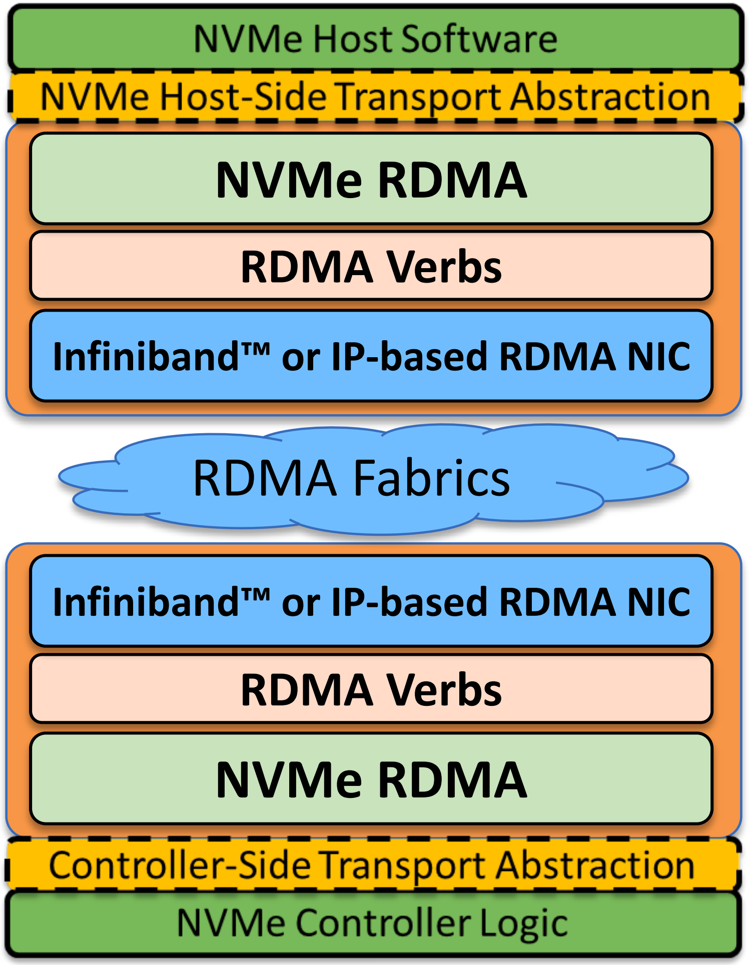
- Preserves the latency and message rate characteristics of the original
- Solves the scale-out limitations



How does NVMe over Fabrics preserve the latency and message rate characteristics of NVMe?

By direct mapping of the NVMe programming model to RDMA Verbs

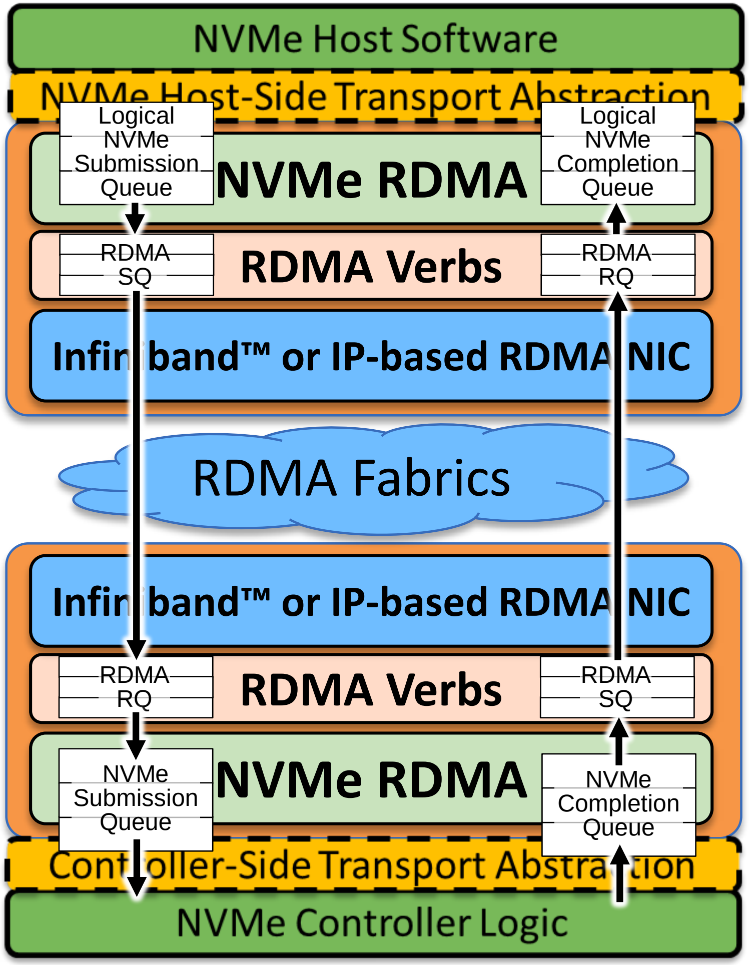
- Maintains the NVMe PCIe operational model and NVMe descriptors
- Simple mapping of NVMe IOQ to RDMA QP


NVMe Host Software
NVMe Host-Side Transport Abstraction
NVMe RDMA
RDMA Verbs
Infiniband™ or IP-based RDMA NIC
RDMA Fabrics
Infiniband™ or IP-based RDMA NIC
RDMA Verbs
NVMe RDMA
Controller-Side Transport Abstraction
NVMe Controller Logic

How does NVMe over Fabrics preserve the latency and message rate characteristics of NVMe?

By direct mapping of the NVMe programming model to RDMA Verbs

- ❐ Maintains the NVMe PCIe operational model and NVMe descriptors
- ❐ Simple mapping of NVMe IOQ to RDMA QP
- ❐ Simple translation of NVMe DMA operations to RDMA operations
- ❐ Simple mapping and translations enable low latency, high message rate, and very simple conversion from RDMA- to NVMe- semantics in the Controller

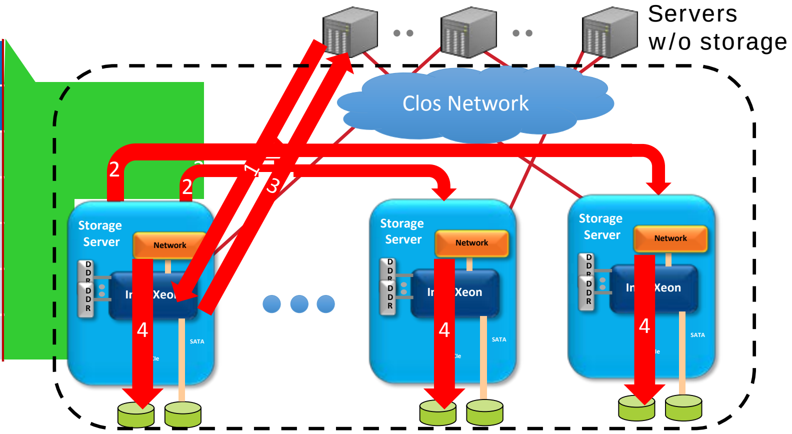

NVMe Host Software
NVMe Host-Side Transport Abstraction
Logical NVMe Submission Queue
NVMe RDMA
Logical NVMe Completion Queue
RDMA SQ
RDMA Verbs
RDMA RQ
Infiniband™ or IP-based RDMA NIC
RDMA Fabrics
Infiniband™ or IP-based RDMA NIC
RDMA RQ
RDMA Verbs
RDMA SQ
NVMe Submission Queue
NVMe RDMA
NVMe Completion Queue
Controller-Side Transport Abstraction
NVMe Controller Logic

Cloud Storage Dataflow - Network Performance Requirements

How many NVMe drives are required to saturate various network link speeds using *NVMe over Fabrics* protocol at the Storage Server Home Node?

- Each NVMe drive is capable of 2GB/s sustained writes
- Traffic pattern is 100% write, 3x replication, 4KB block i/o

Link Speed (Gb/s)	# NVMe drives to saturate link	4KB block 100% Wr, 3x replication	
		KIOPs	Pkt Rate (Mp/s)
25	0.8	346	4.15
50	1.6	691	8.30
100	3.1	1383	16.59
200	6.3	2765	33.19
400	12.5	5531	66.37



- In this Cloud Storage Architecture, a modest number of NVMe devices can
  - generate massive network bandwidth
  - drive packet rates high enough to make RDMA/offload solutions very attractive

Call to Action

- We highlighted the four trends unfolding in the data center
  - Increasing performance of Network Bandwidth
  - Storage Media such as Intel's 3D-XPoint that is approaching the performance of DRAM
  - Single processor/core performance is not increasing at the same rate as network and storage, placing an emphasis on scaling workloads out over available cores and exploiting RDMA to offload cycles related to network processing.
  - In anticipation of these first three trends, OSVs are optimizing the code path of their storage stacks to take advantage of the increased network and storage performance
- We elaborated on the implication of these trends on design of modern distributed storage platforms
- We provided details on how onload/offload mechanisms and RDMA become feature/function requirements for these platforms in the near-future with a focus on NVMe-over-Fabrics.

InfiniBand Trade Association

Global member organization dedicated to developing, maintaining and furthering the InfiniBand specification

- Architecture definition
  - RDMA software architecture
  - InfiniBand, up to 100Gb/s and 300Gb/s per port
  - RDMA over Converged Ethernet (RoCE)
- Compliance and interoperability testing of commercial products
- Markets and promotes InfiniBand/RoCE
  - Online, marketing and public relations engagements
  - IBTA-sponsored technical events and resources

Steering committee members

CRAY THE SUPERCOMPUTER COMPANY · EMULEX An Avago Technologies Company · hp · IBM · intel · Mellanox TECHNOLOGIES · Microsoft · ORACLE · QLOGIC

For More Information

INFINIBAND℠
TRADE ASSOCIATION

www.infinibandta.org

RoCE™

www.roceinitiative.org

© InfiniBand Trade Association

Speaker Bios

- **David Cohen**, System Architect, Intel
  - Dave is a System Architect and Senior Principal Engineer in Intel's Data Center Group where he focuses on the system implications of the intersection of networking and storage in the modern data center.
- **Brian Hausauer**, Hardware Architect, Intel
  - Brian is a Hardware Architect and Principal Engineer in Intel's Data Center Group with focus on Ethernet RDMA engine architecture, and the application of RDMA to emerging storage use cases.