Pre-Conference Seminar F
Flash Storage Networking

Rob Davis, Ilker Cebeli, Brian Pan, Abdel Sadek, Rupin Mohan, Steve McQuerry, and Alan Weckel
Why Network Flash Based Storage?

- There are advantages to shared storage
  - Better utilization:
    - capacity, rack space, power
  - Scalability
  - Manageability
  - Fault isolation, and recovery
- Shared storage requires a Network
Agenda

• Networked Flash Storage Overview – 1:00 to 1:20
  • Rob Davis, Mellanox, VP Storage Technology

• Ethernet Networked Flash Storage – ~1:20 to 1:40
  • Steve McQuerry, Pure Storage, Senior Technical Marketing Engineer

• InfiniBand Networked Flash Storage – ~1:40 to 2:00
  • Abdel Sadek, NetApp, Technical Program Manager

• PCIe Networked Flash Storage – ~2:00 to 2:20
  • Brian Pan, H3 Platform, GM

• Fibre Channel Networked Flash Storage – ~2:20 to 2:40
  • Rupin Mohan, HPE, Chief Technologist
Agenda (cont.)

• Conference Break – 2:45 to 3:00
• How Networking Affects Flash Storage Systems – 3:00 to 3:20
  • Ilker Cebeli, Samsung, Sr. Dir. Product Planning
• Flash Storage Networking, How the market is evolving – ~3:20 to 3:40
  • Alan Weckel, 650 Group, Technology Analyst/Co-Founder
• Q/A and Panel Discussion – ~3:40 to 4:00
  • All Presenters
Flash Makes Networking More Difficult

![Bar chart showing access time in microseconds for different storage media technologies: HDD, SSD, PM. The chart indicates a 10,000x improvement in access time for PM compared to HDD.](chart.jpg)
Faster Storage Needs a Faster Network

Only 2 SSDs fill 10GbE
24 HDDs to fill 10GbE
One SAS SSD fills 10GbE

Flash SSDs move the Bottleneck from the Disk to the Network
What is the solution?
Faster Network Wires are Available

Ethernet & InfiniBand – 200Gb, going to 400Gb…
Pcie – Gen3(8Gb/lane), going to Gen4(16Gb/lane)…
FC – 32Gb, going to 128Gb…
Flash Needs Speed and Low Latency

Network hops multiply latency

Averaged Response Time Logarithmic scale

HDDs (ms) 1ms
SATA SSDs (100μs)
100μs
10μs
1μs

Common Switch & Adapter
NVMe SSDs (10μs)
Optane SSDs (<10μs)
Low Latency Switch & Adapter

Request/Response
Faster Network Components Solves Some of the Problem…

Networked Storage
Storage Protocol (SW)
Network

Storage Media
HDD
SSD
NVM

Protocol and Network
Access Time (micro-sec)

HD SSD NVM
FC, TCP RDMA RDMA+

Faster Wire Speeds
More Efficient Protocol

Flash Memory Summit 2019
Santa Clara, CA
Faster Protocols

- NVMe-oF
  - RDMA (RoCE, IB)
  - Fibre Channel
  - PCIe
  - TCP
- RDMA
  - SMB Direct
  - VSAN over RDMA
  - iSER
  - NFSoRDMA
  - Ceph o RDMA
Where best to plug in?
Flash Storage – Closer to Servers
The Solution will often drive the protocol and the network technology

- All technologies support Block
- All technologies do not support File and Object

Technologies can be layered:
- Block behind File

Flash based solutions will almost always improve performance with a faster network and protocol.

**Block storage**
Data stored in fixed-size ‘blocks’ in a rigid arrangement—ideal for enterprise databases

**File storage**
Data stored as ‘files’ in hierarchically nested ‘folders’—ideal for active documents

**Object storage**
Data stored as ‘objects’ in scalable ‘buckets’—ideal for unstructured big data, analytics and archiving
Conclusions

• There are tried and true reasons for networking your storage
• Networking flash requires special considerations
  • Faster Storage needs Faster Networks!
  • And protocols
• For the next few hours this team will present the different options and trade offs
• Then you get to question us
Thank You
Steve McQuerry is a Senior Technical Marketing Engineer on the platform team at Pure Storage. One of his areas of focus is helping customers understand the use cases and best practices for deploying NVMe-oF with FlashArray.

Steve is a CCIE Emeritus (CCIE #6108) is a 20+ year data center veteran. For the last 16 years he has held both field and product positions for storage, networking, and compute OEMs. Steve has published multiple networking books and has been recognized as distinguished speaker at industry events. Steve holds a Bachelor of Science in Engineering Physics from Eastern Kentucky University.
Ethernet Networked Flash Storage

Flash Storage Networking

Steve McQuerry, Sr. TME PureStorage
• Background
• Common Transport Use Cases
• Customer profiles
• Observed Trends
• Pros & Cons
95% Ethernet

- PRIMARY STORAGE
  - Traditional SAN (FC, iSCSI)
  - 20% of capacity

- SECONDARY STORAGE
  - Mostly Ethernet
  - 80% of capacity
  - Rapid growth
  - Diverse data types
  - Scale-out
  - Tiered data
Ethernet is the dominate transport in the Data Center from a port count perspective. Ethernet speeds have increased exponentially over the last 10 years. File and Software Defined Storage are predominantly Ethernet based transports. Ethernet has had challenges when it comes to guaranteed throughput and deterministic latency.
Common Transport Use Cases

- **iSCSI** – Front end block, not typically focused on performance
- **iSER** – Front end block focused on performance leverages RDMA
- **iWARP** – Front end; block, file, and SDS
- **RoCE** – Front/Back end; block, and SDS
- **Other IP** – Front end; block, file and SDS,
Customer Ethernet Transport Mix
Ethernet Performance Roadmap
Observed Trends

- Majority of block storage continues to be on Fibre Channel
- Ethernet is becoming more interesting to customers from an operational standpoint
- File has been predominately IP/Ethernet and that trend continues
- SDS leverages IP/Ethernet and DAS
- Ethernet has become an interesting transport for disaggregating storage from compute
Pros & Cons

- Well understood
- Common Infrastructure
- Multiple Vendors
- Rapidly evolving
- Lower Cost

- Lossy/Unpredictable performance
- QoS can be complex
- Sometimes still a separate infrastructure
Thank You
Abdel Sadek Bio

Abdel has been in the storage industry for over 15 years. Working for LSI then NetApp. He has a wide experience in different SAN protocol including NVMe-oF, IB, FC, SAS and iSCSI.

He’s part of the NetApp E-Series Technical Marketing team focusing on High Performance Computing, data analytics and Media and entertainment.
Agenda

- Background
- Storage types
- Latency vs. Scalability
- InfiniBand solution
- Use Cases
- Pros and Cons
• Storage is moving to be “Memory-like”
  • Economics based on NAND, not on “spinning rust” and mechanicals
  • Supply line is now linked to Fabs and Semiconductor technology
  • Memory-like semantics supported by RDMA
• Software and supporting H/W is evolving to address the moving bottleneck
  • Demands lighter weight (I/O) software stack
  • Network now a critical component
  • Persistent memory becoming a market reality
  • Persistent memory Filesystems are emerging
  • Rapid adoption of In-Memory computing
# Different Types of Storage

<table>
<thead>
<tr>
<th>Layer 0</th>
<th>Layer 1</th>
<th>Layer 2</th>
<th>Layer 3</th>
<th>Layer 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Latency</td>
<td>&lt;20ns</td>
<td>2µs–10µs</td>
<td>10µs–200µs</td>
<td>300µs–10ms</td>
</tr>
<tr>
<td>IOPS</td>
<td>00s millions</td>
<td>0s millions</td>
<td>0s millions</td>
<td>00,000s</td>
</tr>
<tr>
<td>Bandwidth</td>
<td>00s GBps</td>
<td>0s GBps</td>
<td>100s GBps</td>
<td>100s GBps</td>
</tr>
<tr>
<td>Capacity</td>
<td>1s TB</td>
<td>10s TB</td>
<td>100s TB</td>
<td>10s PB</td>
</tr>
<tr>
<td>Cost</td>
<td>Very high</td>
<td>High</td>
<td>Mid</td>
<td>Mid/low</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type</th>
<th>Access</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRAM</td>
<td>Memory bus</td>
</tr>
<tr>
<td>DDR</td>
<td>Memory bus</td>
</tr>
<tr>
<td>SCM</td>
<td>Memory bus</td>
</tr>
<tr>
<td>SSD with SCM cache</td>
<td>Fabric</td>
</tr>
<tr>
<td>HDD with SSD cache</td>
<td>Ethernet/FC</td>
</tr>
<tr>
<td>HDD</td>
<td>SAS</td>
</tr>
<tr>
<td></td>
<td>PCIe/SAS</td>
</tr>
<tr>
<td></td>
<td>Ethernet</td>
</tr>
</tbody>
</table>

Flash Memory Summit 2019
Santa Clara, CA
Challenge: Latency vs. Scalability

Scalability with low latency

Applications

Networking and Transport

Media

Low Latency
In-Memory
IB: SRP, iSER, NVMe/IB
Low Latency
Scalability
NVMe, SAS
Storage Class Memory
InfiniBand Architecture Highlights

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

- High performance databases
- Parallel Filesystems
- AI/ML/DL workloads.
- Financial trading.
- Realtime modeling (metrology, logistics)
Case 1: RDMA Hand Shake with NVMe/IB
Case 2: Parallel Filesystems (HPC)

Client 1 → Client 2 → Client 3 → Client n

Parallel Filesystem Backend

Storage/MetaData Servers

Backend SAN Protocols: SAS, iSER, SRP, NVMe/IB

MetaData Target → Storage Target

InfiniBand or Ethernet Fabric

Parallel Filesystem Backend

Storage/MetaData Servers

Backend SAN Protocols: SAS, iSER, SRP, NVMe/IB

MetaData Target → Storage Target
Public References

- **Simula Research Lab:**

- **Australian National University:**
### Pros and Cons of InfiniBand

<table>
<thead>
<tr>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low Latency</td>
<td>Distance with low latency</td>
</tr>
<tr>
<td>RDMA support</td>
<td>H/W offload adoption limited</td>
</tr>
<tr>
<td>High Bandwidth</td>
<td>Optimization required for IOPs</td>
</tr>
<tr>
<td>Open source drivers</td>
<td>Limited OS support</td>
</tr>
<tr>
<td>Reliable</td>
<td>Different cables/connector for different speeds</td>
</tr>
<tr>
<td>Lossless</td>
<td></td>
</tr>
</tbody>
</table>
Thank You
Brian Pan | H3

GM

huaiyangpan

www.h3platform.com

brian.pan@h3platform.com

+886 2 2698 3800#110
PCIe® Networked Flash Storage

Brian Pan

H3 Platform
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

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

*Source – PCI-SIG PCI Express 3.0 FAQ*
NVMe Through PCIe Fabric

NVMe Host 1
NVMe Host 2
NVMe Host 3
NVMe Host 4

PCIe Switch

NVMe SSD
NVMe SSD
NVMe SSD
NVMe SSD

PCIe Switch

NVMe SSD
NVMe SSD
NVMe SSD
NVMe SSD

Flash Memory Summit 2019
Santa Clara, CA
Features of PCIe Fabric

- NVMe SSD sharing
  - NVMe can be assigned to any connected host in PCIe fabric
  - NVMe SSD can be shared by using NVMe virtual function (The VF can be assigned to VM on the host)

- Namespace and VF mapping
  - Users can create/delete namespace and manage the mapping of VF and namespace
Architecture of NVMe SSD SR-IOV

x86 server
- CentOS 7.5
- VMware

JBOf
- Broadcom ARM mCPU
- Broadcom PCIe switch
- PCIe switch driver

NVMe SSD
- Samsung 1725a
- Samsung PF/VF driver
JBOf Specification

- **PCIe switch**
  - Broadcom 9797 PCIe switch
  - Broadcom ARM 58522 mCPU

- **Host connection**
  - 2x PCIe Gen3 x16 for host connection

- **NVMe SSD**
  - 16x U.2 Samsung 1725b NVMe SSD
Software Features

- **NVMe SSD management**
  - Create namespace of NVMe SSD
  - Create maximum VF of NVMe SSD (15+1 VF)
  - Assign VF to namespace or vise versa

- **Host and NVMe SSD hotplug**
  - Manage surprise removal and plug-in

- **GUI and API**
  - GUI or API for JBOf management
NVMe with SR-IOV. VM talk to VF directly. PF is sit on PCIe switch.

NVMe without SR-IOV. Passthrough model

NVMe without SR-IOV. Hypervisor manage VMs to NVMe SSD

Unprivileged domain (DomU)

Privileged domain (Dom0)
SR-IOV Preparation Process

Detect SRIOV devices

Match the device with compatible list and load driver by autoscript

Successful mounted?

- Yes
  - Launch PF drivers and auto set up the maximum VF quantity

- No
  - Cannot be bootup and show warning on GUI, LED send alert mail,

Show the PF ready and unallocated VF in GUI

End
VF and Namespace Assignment
Namespace on VM

Direct-Attach VF to VM Passthruogh
Dump NS from VM

[root@localhost ~]# nvme list

<table>
<thead>
<tr>
<th>Node</th>
<th>SN</th>
<th>Model</th>
<th>Namespace Usage</th>
<th>Format</th>
<th>FW Rev</th>
</tr>
</thead>
<tbody>
<tr>
<td>/dev/nvme0n1</td>
<td>S3H9NX0J7000013</td>
<td>SAMSUNG MZWLL6T4HMLS-00003</td>
<td>2</td>
<td>53.69 GB / 53.69 GB</td>
<td>GPNA6B3T</td>
</tr>
<tr>
<td>/dev/nvme0n2</td>
<td>S3H9NX0J7000013</td>
<td>SAMSUNG MZWLL6T4HMLS-00003</td>
<td>5</td>
<td>134.22 GB / 134.22 GB</td>
<td>GPNA6B3T</td>
</tr>
</tbody>
</table>
## FIO Test Result – Direct vs SR-IOV

<table>
<thead>
<tr>
<th>Performance</th>
<th>Throughput</th>
<th>Latency (4K R)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Directly attached</td>
<td>3180 MB/s</td>
<td>91 usec</td>
</tr>
<tr>
<td>Aggregate SR-IOV</td>
<td>3058 MB/s</td>
<td>90 usec</td>
</tr>
</tbody>
</table>
Multi-host SR-IOV vs Pass Throught vs Hyper+ SR-IOV

Host 1

VM1

Hypervisor

VF

NVMe SSD

VF

NVMe SSD

VF

NVMe SSD

VF

NVMe SSD

PCIe Switch

PCIe

Host 2

VM2

Hypervisor

PCIe

VM3

Hypervisor

PCIe

VM4

Hypervisor

PCIe

NVMe with SR-IOV. VM talk to VF directly. PF is sit on PCIe switch.

NVMe without SR-IOV. Passthrough model

NVMe without SR-IOV. Hypervisor manage VMs to NVMe SSD

Unprivileged domain (DomU)

Privileged domain (Dom0)
Benefits of Sharing NVMe Through SR-IOV

- **Performance and latency**
  - VF latency is only 1/3 of PF latency in multi-VMs

- **Cost saving**
  - Tens of VFs associated with a single PF, extending the capacity of a device and lowering the hardware cost
  - With better latency and performance, the utilization rate will be higher to further reduce the hardware cost
Benefits of Sharing NVMe Through SR-IOV

- Multi-path IO via PCIe
  - The namespace on NVMe can be accessed by different hosts through PCIe connection

- Flexibility configuration
  - Dynamical control by the mCPU to assign VF or create namespace, users have the flexibility to manage NVMe SSD
## Comparison of Different Interconnect

<table>
<thead>
<tr>
<th>Technology</th>
<th>PCIe</th>
<th>Ethernet</th>
<th>Fiber</th>
<th>Infiniband</th>
</tr>
</thead>
<tbody>
<tr>
<td>Latency (us)</td>
<td>0.25</td>
<td>10</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>Perf. Gbps, single port</td>
<td>256</td>
<td>400</td>
<td>64</td>
<td>200</td>
</tr>
<tr>
<td>Applications</td>
<td>I/O Area Network (IAN)</td>
<td>Local Area Network (LAN)</td>
<td>Storage Area Network (SAN)</td>
<td>I/O Area Network (IAN)</td>
</tr>
</tbody>
</table>
## Comparison of Different Interconnect

<table>
<thead>
<tr>
<th>Technology</th>
<th>PCIe</th>
<th>Ethernet</th>
<th>Fiber</th>
<th>Infiniband</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Scalibility</strong></td>
<td>Low</td>
<td>Very high</td>
<td>Medium</td>
<td>High</td>
</tr>
<tr>
<td><strong>Ease of management</strong></td>
<td>Low</td>
<td>High</td>
<td>Medium</td>
<td>Medium</td>
</tr>
</tbody>
</table>
Performance Results with SR-IOV

Latency
Server_1 access to VF_0 (NS_A)

<table>
<thead>
<tr>
<th>Latency (usec) avg. and distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>4K Read (Random)</td>
</tr>
<tr>
<td>90</td>
</tr>
<tr>
<td>50-100=97.70%, 100-250=2.30%</td>
</tr>
</tbody>
</table>

Server_1 access to VF_0 (NS_A)

<table>
<thead>
<tr>
<th>Latency (usec) avg. and distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>4K Write (Random)</td>
</tr>
<tr>
<td>25</td>
</tr>
<tr>
<td>10-20=85.11%, 20-50=14.84%</td>
</tr>
</tbody>
</table>

Note
The Latency measured using Fio in CentOS 7.5, with queue depth 1 by 1 worker and CPU core allowed 1.
The performance measured using Fio in CentOS 7.5, with queue depth 32 by 16 workers and CPU core allowed 8.
Performance Results with SR-IOV

### Read Performance

<table>
<thead>
<tr>
<th>Tested Functions</th>
<th>4K Read (Random)</th>
<th>4K Read (Sequential)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MB/s</td>
<td>IOPS</td>
</tr>
<tr>
<td>Server1 access to VF0 (NS_A)</td>
<td>1,496</td>
<td>382,990</td>
</tr>
<tr>
<td>Server2 access to VF1 (NS_C)</td>
<td>1,562</td>
<td>400,016</td>
</tr>
<tr>
<td>Server1 access to VF0 (NS_B)</td>
<td>1,494</td>
<td>382,601</td>
</tr>
<tr>
<td>Server2 access to VF1 (NS_B)</td>
<td>1,560</td>
<td>399,519</td>
</tr>
</tbody>
</table>

### Write Performance

<table>
<thead>
<tr>
<th>Tested Functions</th>
<th>4K Write (Random)</th>
<th>4K Write (Sequential)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MB/s</td>
<td>IOPS</td>
</tr>
<tr>
<td>Server1 access to VF0 (NS_A)</td>
<td>924</td>
<td>230,984</td>
</tr>
<tr>
<td>Server2 access to VF1 (NS_C)</td>
<td>979</td>
<td>244,881</td>
</tr>
<tr>
<td>Server1 access to VF0 (NS_B)</td>
<td>953</td>
<td>238,382</td>
</tr>
<tr>
<td>Server2 access to VF1 (NS_B)</td>
<td>988</td>
<td>253,093</td>
</tr>
</tbody>
</table>
Performance Results without SR-IOV

Latency

Linux access to VF_0 (NS_A)

4K Read (Random)

Latency (usec) avg. and distribution

91 50-100=92.42%, 100-250=7.57%

Linux access to VF_0 (NS_A)

4K Write (Random)

Latency (usec) avg. and distribution

18 10-20=96.71%, 20-50=3.24%

Note

The Latency measured using Fio in CentOS 7.5, with queue depth 1 by 1 worker and CPU core allowed 1.
The performance measured using Fio in CentOS 7.5, with queue depth 32 by 16 workers and CPU core allowed 8.
## Performance Results without SR-IOV

### Read Performance

<table>
<thead>
<tr>
<th>Tested Functions</th>
<th>4K Read (Random)</th>
<th>4K Read (Sequential)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MB/s</td>
<td>IOPS</td>
</tr>
<tr>
<td>Linux access to PF</td>
<td>3,180</td>
<td>814k</td>
</tr>
<tr>
<td>Linux access to VF0</td>
<td>3,130</td>
<td>801k</td>
</tr>
<tr>
<td>VM 1 access to VF0 (NS_A)</td>
<td>2,822</td>
<td>722k</td>
</tr>
<tr>
<td>VM 1 access to VF0 (NS_B)</td>
<td>1,611</td>
<td>415k</td>
</tr>
<tr>
<td>VM 2 access to VF1 (NS_B)</td>
<td>1,570</td>
<td>403k</td>
</tr>
</tbody>
</table>

### Write Performance

<table>
<thead>
<tr>
<th>Tested Functions</th>
<th>4K Write (Random)</th>
<th>4K Write (Sequential)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MB/s</td>
<td>IOPS</td>
</tr>
<tr>
<td>Linux access to PF</td>
<td>1,716</td>
<td>439k</td>
</tr>
<tr>
<td>Linux access to VF0</td>
<td>1,783</td>
<td>457k</td>
</tr>
<tr>
<td>VM 1 access to VF0 (NS_A)</td>
<td>1,782</td>
<td>456k</td>
</tr>
</tbody>
</table>
Thank You
Rupin Mohan is a Director of R&D and CTO of Storage Networking (SAN) at HPE Storage. Rupin leads a global engineering team responsible for development of Storage Networking products. Rupin has filed 30+ patents at HPE. He is a Board Member and Marketing Chairman for FCIA. Rupin completed his MBA from MIT Sloan School of Management as a Sloan Fellow. He also holds a MS in Engineering from Tufts University and BE in Computer Engineering from Delhi Institute of Technology.
NVMe over Fibre Channel

Rupin Mohan
Director R&D, Chief Technologist (SAN)
HPE Storage
Fibre Channel: Timeline

1988: Work begins on FC protocol

1994: FCA and FCLC Formed

1997: 1Gb FC

1999: FCIA Formed

2001: 2Gb FC

2005: 4Gb FC

2008: 8Gb FC

2009: FCOE

2012: 16GFC

2015: Gen6 32GFC

2019: Gen7 64GFC

Arbitrated Loop

Virtualization NPIV

NVMe Fabrics Fibre Channel

Converged Networks

Fabric Services

2015: Converged Networks
2019 Data Center Storage Trends

Workloads are being brought back from the public cloud

• Lack of understanding performance or sensitivity requirements
• Last year 41% of business brought at least one workload back*

Rapid growth of all-flash arrays, NVMe & NVMe-oF

• NVMe over FC expected to outpace NVMe over Ethernet starting this year*
• NVMe over FC is easy to deploy, high-performance, extremely reliable

Analytics/Artificial Intelligence Requirements

• Requires great volumes of data & fast access to it. Storage has become disaggregated due to public cloud & edge. Finding data takes too long. Lost time = diminished value of analytics and AI initiatives.

*ESG, 1/7/19, 2019 Data Storage Predictions: More Cloud Missteps, FC Is Back, and Finding Data Holds Back AI
According to Dell’Oro Group’s latest Storage Area Networks (SAN) 5-Year Forecast Report (final data will be available in Dell’Oro Group’s 4Q18 report):

- Total Fibre Channel SAN port shipments (Fibre Channel switch and Fibre Channel adapter) for 2018 is expected to approach 7.7 million, up more than 11 percent over 2017 (6.8 million)
- Total Fibre Channel SAN revenue (Fibre Channel switch and Fibre Channel adapter) is expected to approach $2.5 billion in 2018, up nearly 22 percent over 2017
**FC-NVMe-2**

**WHATS NEW?**

- Refinements to existing FC-NVMe standard
- Enhances error recovery to more granular level
- Improves performance predictability
- Prevents connection disruption

- Sequence level error recovery – SLER - (exchanges are given more time before getting terminated, re-transmissions, new commands)

- A method to respond/correct minor error conditions to avoid NVMe subsystem disconnect/reconnect

- Will auto-negotiate for compatibility with FC-NVMe connections

**HOW**

- Now a published standard.
NVMe-oF deployment (FC)

NVMe storage attached in the backend

NVMe end to end using FC
The landscape today….  

<table>
<thead>
<tr>
<th>Protocol</th>
<th>Latency</th>
<th>Scalable</th>
<th>Performance</th>
<th>Enterprise Footprint</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fibre Channel</td>
<td>Lower</td>
<td>Yes</td>
<td>High</td>
<td>Reliable Storage Fabric</td>
</tr>
<tr>
<td>RoCEv2</td>
<td>Lowest</td>
<td>Yes</td>
<td>High</td>
<td>Negligible</td>
</tr>
<tr>
<td>iWARP</td>
<td>Medium</td>
<td>Yes</td>
<td>Medium</td>
<td>Negligible</td>
</tr>
<tr>
<td>TCP</td>
<td>High</td>
<td>Yes</td>
<td>Medium</td>
<td>Medium with iSCSI</td>
</tr>
<tr>
<td>InfiniBand</td>
<td>Lowest</td>
<td>Limited</td>
<td>High</td>
<td>None</td>
</tr>
</tbody>
</table>
TABLE OF CONTENTS

Foreword ...........................................................................................................5

FCIA President Introduction ........................................................................6

The State of Fibre Channel by Storage Switzerland ......................8

Fibre Channel New Technologies: FC-NVMe-2.............................9

The 2019 Fibre Channel Roadmap .............................................................10

Fibre Channel’s Future is Bright in Media and Entertainment ..........12

Securing Fibre Channel SANs with End-to-End Encryption ..........14

Download: https://fibrechannel.org/  Hardcopy: FCIA booth
Thank You
Ilker Cebeli

Ilker is a Senior Director of Product Planning at Samsung. He is responsible for leading the Emerging memory, SSD, and All-Flash-Array related storage solutions and technologies. He has spent 25 years in enterprise computing, storage, and networking working in various roles. Prior to joining to Samsung, Ilker worked at Micron, and was leading and directing emerging memory projects in memory division. Ilker also spent 15 years at Intel and he was responsible for Intel's Xeon™ product planning and server platform architecture definition.
How Networking Affects Flash Storage Systems

Ilker Cebeli
Senior Director of Product Planning, Samsung
This presentation and/or accompanying oral statements by Samsung representatives collectively, the “Presentation”) is intended to provide information concerning the SSD and memory industry and Samsung Electronics Co., Ltd. and certain affiliates (collectively, “Samsung”). While Samsung strives to provide information that is accurate and up-to-date, this Presentation may nonetheless contain inaccuracies or omissions. As a consequence, Samsung does not in any way guarantee the accuracy or completeness of the information provided in this Presentation.

This Presentation may include forward-looking statements, including, but not limited to, statements about any matter that is not a historical fact; statements regarding Samsung’s intentions, beliefs or current expectations concerning, among other things, market prospects, technological developments, growth, strategies, and the industry in which Samsung operates; and statements regarding products or features that are still in development. By their nature, forward-looking statements involve risks and uncertainties, because they relate to events and depend on circumstances that may or may not occur in the future. Samsung cautions you that forward looking statements are not guarantees of future performance and that the actual developments of Samsung, the market, or industry in which Samsung operates may differ materially from those made or suggested by the forward-looking statements in this Presentation. In addition, even if such forward-looking statements are shown to be accurate, those developments may not be indicative of developments in future periods.
NVMe Technology – Background

- Optimized for flash
  - Traditional SCSI designed for disk
  - NVMe bypasses unneeded layers
  - Dramatically reducing latency
NVMe Design Advantages

- Lower latency
- Direct connection to CPU’s PCIe lanes
- Higher bandwidth
- Scales with number of PCIe lanes
- Best in class latency consistency
- Lower cycles/IO, fewer cmds, better queueing
- Lower system power
- No HBA required
NVMe Technology – Background

- NVMe outperforms SATA SSDs
  - 6X-7X more bandwidth,
  - 40-50% lower latency
  - Up to 8x more IOPS
What is NVM Express Over Fabrics?

- A protocol interface to NVMe that enable operation over other interconnects (e.g., Ethernet, InfiniBand™, Fibre Channel).
- Shares the same base architecture and NVMe Host Software as PCIe.
- Enables NVMe Scale-Out and low latency (<10µS latency) operations on Data Center Fabrics.
- Avoids protocol translation (avoid SCSI).
Some of the use cases for NVMe Over Fabrics

1. Software-Defined Storage (SDS)
2. Hyper-Converged JBOF Storage
   - Direct Attached JBOF
   - SAS DAS Replacement
   - PCIe
Performance Test Configuration – 2016

• **1x NVMe-oF target**
  - 24x NVMe 2.5” SSDs
  - 2x 100GbE NICs
  - Dual x86 CPUs

• **4x initiator hosts**
  - 2x25GbE NICs each

• **Open Source NVMe-oF kernel drivers**
Local vs. Remote Latency Comparison – 2016

<table>
<thead>
<tr>
<th></th>
<th>Read Gap</th>
<th>Write Gap</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>~17 us</td>
<td>~9 us</td>
</tr>
</tbody>
</table>
Performance Test Configuration – 2017

- **1x NVMe-oF target**
  - 36x NF1 SSDs
  - 2x 100GbE NICs, 2x 50GbE NICs
  - Dual x86 CPUs

- **6x initiator clients**
  - 2x25Gb/s each

- **Open Source NVMe-oF kernel drivers**
  - Ubuntu Linux 16.04/4.9 on Target
Local vs. Remote Latency Comparison - 2017

**Flash Memory Summit**

### Read Gap vs. Write Gap

<table>
<thead>
<tr>
<th></th>
<th>Read Gap</th>
<th>Write Gap</th>
</tr>
</thead>
<tbody>
<tr>
<td>2016</td>
<td>~17 us</td>
<td>~9 us</td>
</tr>
<tr>
<td>2017</td>
<td>~14 us</td>
<td>~10 us</td>
</tr>
</tbody>
</table>

**Random IO Latency Chart**

- **Local**: 79.05, 16.37
- **Remote**: 92.81, 26.05
SSDs Will Continue to get Faster

- **2016 Tests**
  - Read Gap: ~14 us
  - Write Gap: ~10 us

- **2017 Tests**
  - Read Gap: ~17 us
  - Write Gap: ~9 us
Thank You
Alan Weckel is Technology Analyst/Co-Founder at 650 Group, where he is in charge of Ethernet switch, Cloud and data center research. He has written many articles for the trade and technical press, and is frequently quoted in such leading publications as Bloomberg, Businessweek, Forbes, Network World, and the Wall Street Journal. Before co-founding 650 Group, he was VP/analyst at Dell’Oro Group and had engineering and software development experience at Raytheon, General Electric Power Systems, and Cisco. He holds a BSEE and an MS in Management from Rensselaer Polytechnic Institute.
Flash Storage Networking,
How the market is evolving

Alan Weckel (alan@650group.com)
Trends changing how compute and storage are consumed
Storage: How and Where We Store Data is Changing

• Enterprise Storage Systems Market is Shrinking
  • Enterprises continue to buy systems
  • Enterprise market for converged and hyperconverged is growing

• Cloud Market is Growing
  • Hyperscalers buy components
  • Hyperscalers build their own software

• Areas of growth in Storage Systems Market
  • Cloud
  • All Flash Arrays
  • Hyperconverged
Server Shipments: Shipments into the Cloud

- Cloud servers will dominate compute
  - Higher-end processor
  - Smart NIC
  - Better software
  - Different type of storage

- Enterprise servers are being deployed in colocation facilities

- East/West traffic is no longer limited to one data center
  - Ethernet Based Architectures
  - Large amounts of data being moved across the world
Workloads: Installed Base by Deployment

- Enterprise workloads continue to grow
  - More workloads per server
  - Type of application is changing
  - Colocation becoming common

- Cloud workload growth exploding
  - All types of applications are growing
  - IoT will be a major driving of workload growth
  - Edge Computing and AI changing DC design
Server and Smart NICs:
Server Installed Base

- Cloud is the new leader in technology transitions
  - Entire Telco market is smaller than Amazon
  - Cloud is moving from 2-3 to 3-4 technology generations ahead of the enterprise

- Tier 2 and 3 Clouds are increasingly riding on top of Tier 1 Cloud Infrastructure

- Clouds use different architecture and buys different equipment then the enterprise
Ethernet Switch – Data Center
Ethernet Switch – Data Center: Total Market Revenue
US Top 5 Cloud Providers: Amazon, Apple, Facebook, Google, Microsoft
Chinese Tier 1 Cloud Providers: Alibaba, Baidu, Tencent
Fibre Channel

- **FC Switches:**
  - Shipments increased 15% in 2018 following six consecutive years of decline

- **FC HBAs:**
  - Shipments increased 10% in 2018

- **Reasons for Fibre Channel switch and HBA growth:**
  - Very strong economic growth for most of 2018 with Enterprise IT spending following suit.
  - Server and external storage array market recovery
  - Focus and execution following acquisitions
  - Some advance buying

- **2019 forecast assumes low single-digit shipment increase.**
InfiniBand

- 2018 Shipments
  - Up 8% Y/Y to ~700K ports
    - Following steep 2017 decline

- 2018 Revenue
  - Up 6% Y/Y to $217M

- HDR/200Gbs HCA shipments very small in 2018
  - But starting to ramp very strongly for Switches
    - 3Q18: ~1K ports
    - 4Q18: ~20K ports
### Merchant Silicon’s product cycles accelerating in the Cloud

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.3 Tbps</td>
<td>10 Gbps</td>
<td>10 Gbps</td>
<td>10/40 Gbps</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.8 Tbps</td>
<td>25 Gbps</td>
<td>25 Gbps</td>
<td>100 Gbps</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.2 Tbps</td>
<td>25 Gbps</td>
<td>25/50 Gbps</td>
<td>100 Gbps</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.4 Tbps</td>
<td>25 Gbps</td>
<td>25/50 Gbps</td>
<td>100/200 Gbps</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.2 Tbps</td>
<td>100 Gbps</td>
<td>100 Gbps</td>
<td>400 Gbps</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12.8 Tbps</td>
<td>50 Gbps</td>
<td>50/100 Gbps</td>
<td>200/400 Gbps</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12.8 Tbps</td>
<td>100 Gbps</td>
<td>100 Gbps</td>
<td>400 Gbps</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25.6 Tbps</td>
<td>100 Gbps</td>
<td>100 Gbps</td>
<td>800 Gbps</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Two waves of 400 Gbps
  - 8 X 50 Gbps
  - 4 X 100 Gbps

- Pace of Innovation Increasing
  - Four major silicon cycles in five years
  - Some technologies will get orphaned
Conclusion

• Speed of technology advancement is more rapid
• Ethernet is expanding into the Storage connectivity and Data Center transport markets at a rapid pace
• Cloud customers have different architectures and use different equipment then the enterprise
• 2019 will usher in Smart NICs and 200/400 Gbps which will expand the market for Ethernet
Thank You
Panel Q/A

Rob Davis, Ilker Cebeli, Brian Pan, Abdel Sadek, Rupin Mohan, Steve McQuerry, and Alan Weckel
Thank You