Ultra-Fast NVMe Storage Networks for Next Generation Flash Arrays

Sponsored by the Fibre Channel Industry Association (FCIA)

Moderator – Mark Jones FCIA President and Director Tech Marketing Broadcom Inc.
Fibre Channel – Legacy of Interoperability, reliability and robustness
- Products have been shipping in volume for more than 20 years – 120M ports shipped, 43M in current use.

Key Factors to Fibre Channel’s success
- Industry-wide participation in FC standards – INCITS T11
- ~Avg of Two Plugfests per year (39 total) that ensure vendors conform to industry standards
FCIA FC-NVMe Plugfest

- July 23, 2018 - 4th FC-NVMe plugfest
  - 13 Companies/products tested
    - HBAs, Switches, Storage Arrays, Analyzers/Jammers
  - Key Accomplishments
    - Testing of End-end commercial available products
    - Multi vendor interoperability, standards conformance
    - Data Integrity validation over switch multi-hop fabrics
    - Error injection to validate correct FC-NVMe and FC recovery
    - Concurrent FC-NVMe + FC over same Initiator, fabric, target ports
    - “Big Build” overnight stress testing of all of the above.
How to Participate with the FCIA

• [Https://fibrechannel.org](https://fibrechannel.org)
  • 2018 FC Solutions Guide
  • FC Roadmaps
  • Plugfest Information
  • FC Education links

• **FC Education**
  • [www.Brightalk.com](http://www.Brightalk.com) - FCIA

• **Social Media**
  • LinkedIn, Twitter @FCIAnews
Agenda

• Curt Beckman – The New Normal in Storage Latency
• Rupin Mohan – NVMe: A New Language for Storage
• Craig Carlson – FC-NVMe Status and Updates
• Dennis Martin – FC-NVMe Test Results
• David Rodgers – Fibre Channel Test and Measurement
• Q&A
The New Normal in Storage Latency

Curt Beckmann
Product Architect
BSN, Broadcom
How Flash and NVMe are changing storage latency

- Many elements in networked storage latency
  - Flash and NVMe change which elements matter
- In HDD era, arrays could use cache and spindle count
  - Array acceleration hid SW feature overhead
  - Availability “zero cost” in performance
- Fast SSDs make protocol, feature costs visible
  - Networked storage still has benefits, of course
  - But some apps may prefer speed over features
Elements in application storage latency

Server I/O Stack

SAN

Target Services and Acceleration

Media

APP

OS

HYPER

SCSI

FCHBA

SCSI HDD

SCSI SSD

NVMe SSD

(Future NVMe fast path)

(Cache)

Controller Services & Acceleration

SAN FABRIC

(Raw access)
How Flash and NVMe are changing SAN Storage Latency

<table>
<thead>
<tr>
<th>Media Eras</th>
<th>Fast HDD (estimate)</th>
<th>SCSI Flash (estimate)</th>
<th>NVMe o FC (anticipated)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Server IO Stack</td>
<td>40 µs</td>
<td>30 µs</td>
<td>7 µs</td>
</tr>
<tr>
<td>SAN (no queuing)</td>
<td>*6 µs</td>
<td>3 µs</td>
<td>3 µs</td>
</tr>
<tr>
<td>Avg Services</td>
<td>250 µs</td>
<td>150 µs</td>
<td>20 µs</td>
</tr>
<tr>
<td>Avg Acceleration</td>
<td>-200 µs</td>
<td>0 µs</td>
<td>0 µs</td>
</tr>
<tr>
<td>Media</td>
<td>3 ms</td>
<td>50 µs</td>
<td>10 µs</td>
</tr>
<tr>
<td>Raw Access Total</td>
<td>3.0 ms</td>
<td>(?) 83 µs</td>
<td>20 µs</td>
</tr>
<tr>
<td>Services Total</td>
<td>3.1 ms</td>
<td>233 µs</td>
<td>40 µs</td>
</tr>
</tbody>
</table>

In HDD Era, very little incentive to use raw media with networked storage.

In “anticipated” NVMe SSD era, use of raw media may be justified… but many applications may want both!

*16 GFC
Concurrent Enterprise / Raw Media Use Case

Analytics on Active DB

- Using sensitive data for ML is an effective mechanism to:
  - Increase revenue
  - Build customer loyalty
- ML is data intensive, and want results soon as possible
  - But active DB needs protection (adds latency)
  - Hammering active DB with ML slows both down
  - What do do?
- Separate the problem:
  - Protect DB master on full-featured volume (as now)
  - Regularly snapshot / Clone the DB to Raw Media
  - Use Raw Media reference copy for ML
  - Do both on the same infrastructure you use today
NVMe: A new language for storage

Rupin Mohan
Director R&D, CTO SAN
HPE Storage
NVMe – Technology Disruption

Traditional Storage Arrays
1. Storage Controller runs SCSI
2. Front end FC/iSCSI
3. Backend SAS/SATA
4. Software Feature Rich

Hybrid Storage Arrays
1. Storage Controller runs SCSI with upgraded back end – Controller does SCSI-NVMe translation with NVMe drives in the backend
2. 3D Cross Point for Metadata stores on NVMe stack
3. Front end, FC-NVMe
4. Software Feature Rich

NVMe Storage Arrays
1. Storage Controller only runs NVMe
2. Backend NVMe Drives (PCIe, NVMe over Ethernet/Infiniband)
3. Frontend NVMe (FC-NVMe, NVMe over Ethernet)
4. Software Features - low
NVMe over Fabrics Use Cases
NVMe-oF deployment (FC)

NVMe storage attached in the backend

NVMe end to end using FC

FCIA, www.fibrechannel.org
Flash Memory Summit 2018
Santa Clara, CA
NVMe-oF deployment (Ethernet)

- Host NVMe
- Ethernet Fabric
- Storage Array
- PCIe
- JBOF
- NVMe-oF in the backend

RDMA or TCP/IP
Ethernet
RDMA or TCP/IP
Ethernet

RDMA - RocEv1, RocEv2 and iWARP
(Infiniband not included in pictures)

NVMe end to end using Ethernet

FCIA, www.fibrechannel.org
Flash Memory Summit 2018
Santa Clara, CA
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 / Mature Storage Fabric</td>
</tr>
<tr>
<td>RoCEv2</td>
<td>Lowest</td>
<td>Yes</td>
<td>Higher</td>
<td>Negligible</td>
</tr>
<tr>
<td>iWARP (Intel)</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>
FC-NVMe Status and Update
Craig W. Carlson
Marvell
FC-NVMe is real

- FC-NVMe (Fibre Channel over NVMe)
  - First revision of standard completed in 2016
  - Products are now available
    - Based on existing trusted hardware/software platforms
Future development

- FC-NVMe-2 under development now
  - Major new feature is Enhanced Error Recovery
    - Allows for transport level recovery of lost or corrupted commands
      - Occurrence of this is rare, but not impossible
  - Adds additional reliability to already reliable FC SANs
Fibre Channel

- Ratification of 64GFC serial and 256GFC parallel is under way
- Work started on 128GFC serial with 512GFC parallel following
### FCIA Roadmap

<table>
<thead>
<tr>
<th>Product Naming</th>
<th>Throughput (Mbytes/s)</th>
<th>Line Rate (Gbaud)</th>
<th>T11 Specification Technically Complete (Year)*</th>
<th>Market Availability (Year)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>1GFC</td>
<td>200</td>
<td>1.0625</td>
<td>1996</td>
<td>1997</td>
</tr>
<tr>
<td>2GFC</td>
<td>400</td>
<td>2.125</td>
<td>2000</td>
<td>2001</td>
</tr>
<tr>
<td>4GFC</td>
<td>800</td>
<td>4.25</td>
<td>2003</td>
<td>2005</td>
</tr>
<tr>
<td>8GFC</td>
<td>1,600</td>
<td>8.5</td>
<td>2006</td>
<td>2008</td>
</tr>
<tr>
<td>32GFC</td>
<td>6,400</td>
<td>28.05</td>
<td>2013</td>
<td>2016</td>
</tr>
<tr>
<td>128GFC</td>
<td>25,600</td>
<td>4x28.05</td>
<td>2014</td>
<td>2016</td>
</tr>
<tr>
<td>64GFC</td>
<td>12,800</td>
<td>28.9 PAM-4 (57.8Gb/s)</td>
<td>2017</td>
<td>2019</td>
</tr>
<tr>
<td>256GFC</td>
<td>51,200</td>
<td>4x28.9 PAM-4 (4x57.8Gb/s)</td>
<td>2017</td>
<td>2019</td>
</tr>
<tr>
<td>128GFC</td>
<td>25,600</td>
<td>TBD</td>
<td>2020</td>
<td>Market Demand</td>
</tr>
<tr>
<td>256GFC</td>
<td>51,200</td>
<td>TBD</td>
<td>2023</td>
<td>Market Demand</td>
</tr>
<tr>
<td>512GFC</td>
<td>102,400</td>
<td>TBD</td>
<td>2026</td>
<td>Market Demand</td>
</tr>
<tr>
<td>1TFC</td>
<td>204,800</td>
<td>TBD</td>
<td>2029</td>
<td>Market Demand</td>
</tr>
</tbody>
</table>
FC-NVMe Test Results

Dennis Martin
President, Demartek
Independent Test Lab and Analyst

Demartek
About Demartek

- Industry Analysis and ISO 17025 accredited test lab
- Lab includes enterprise servers, networking & storage (6/12Gb SAS, 10/25/40/100GbE, 8/16/32GFC)
- We prefer to run real-world applications to test servers and storage solutions (databases, Hadoop, VMware, etc.)
- Demartek is an EPA-recognized test lab for ENERGY STAR Data Center Storage testing
- Website: https://www.demartek.com/Testlab/
Storage Interface Comparison

- Free reference page on demartek.com
  - https://www.demartek.com/Storage-Interface-Comparison/
  - Search for “storage interface comparison” in your favorite search engine
- Popular page – includes interactive PDF for download
- Provides comparison of storage interfaces
  - FC, FCoE, IB, iSCSI, NVMe, PCIe, SAS, SATA, Thunderbolt, USB
  - Transfer rates, encoding schemes, history, roadmaps, cabling, connectors
- *We’re not a product vendor – we use these technologies in our lab*
FC-SCSI vs. FC-NVMe

- May 2018 Demartek Evaluation: **Performance Benefits of NVMe™ over Fibre Channel – A New, Parallel, Efficient Protocol**

https://www.demartek.com/ModernSAN/
The Test

- Comparison of FC-SCSI to FC-NVMe
- Same hardware, different protocol

https://www.demartek.com/ModernSAN/
Results: Random Read 4KB

Random Read 4KB
Latency vs. IOPS

At least 34% lower latency

53% higher IOPS at 450 µs

54% higher IOPS at 2300 µs

Note: all measurements taken on a single-node A700s. Standard implementations are dual-node.

https://www.demartek.com/ModernSAN/
Zoom-in: RR 4KB

Random Read 4KB
Latency vs. IOPS (zoom in)

Note: all measurements taken on a single-node A700s. Standard implementations are dual-node.

IOPS
- SCSI FCP
- NVMe/FC

At least 34% lower latency

https://www.demartek.com/ModernSAN/
Oracle 80-20 8KB

Oracle 80-20 8KB
Latency vs. IOPS

Note: all measurements taken on a single-node A700s. Standard implementations are dual-node.

https://www.demartek.com/ModernSAN/
Test & Measurement in Fibre Channel
From Inception to Support

Protocol Awareness is Required from Initial Phy Development for Successful NVMe Application Support

David J. Rodgers
High-Speed Fabric Designs and Protocol Analysis

**Basic Premise:**
Mission Critical Storage demands, i.e. NVMe/oF, are fueling the exponential growth of Fibre Channel speeds, protocols, port counts and densities. The challenge to meet the demands of users and applications requires adaptation and evolution of test and measurement tools and practices.

Specific to Phy Layer Designs, link interfaces have evolved to include improved communications schemes and adapted corresponding high-speed transmitter training and equalization practices.

*The impact on hardware designs requires protocol awareness beginning with initial design, through validation, and in the field after deployment.*
Universal T&M Considerations

- Common to all stages of Fibre Channel Fabric Development, Deployment, and Support for NVMe
  - What issue(s) are we trying to understand and correct?
  - When and How does the issue manifest?
  - Is the issue reproducible?
  - Can root cause be definitively determined?
  - What are the curative measures?
  - Can you test the ‘fix’?
  - What are the Cost considerations to vendors, customers?
The physical communications “Channel” must be stable
- Minor Imperfections, once considered ‘routine’ and unremarkable are no longer “minor”

Vendor Interoperability is required!
- From Switch to HBA to Interconnect options, Vendor offerings must work together

Specification Conformance
- FC Physical/Communications Layer
- NVMe iterations

Line-rate Capture/Analysis Tools Needed
- The ability to use a “neutral” observer
Fibre Channel T&M Future

- New Tools and Processes for PHY Testing
  - High Speed Real Time and/or Sampling Scopes
    - Up to 100GHz Today!
- Specialized Traffic Generation Capabilities Supporting:
  - Physical Coding Sublayer – 8b/10b, 64b/66b, 256b/257b encoding
  - Speed-Negotiation, Transmitter Training Sequences
- New Line Rate analysis capabilities supporting:
  - “Pass Through” tapping
  - Bit-level Capture
- The “Channel” must be “smart”
  - Protocol is inherent in the physical layer!
Thank You!

Q&A

Fibrechannel.org
BIOs
Curt Beckman BIO

- Principal Architect – Broadcom Inc
- Author – NVMe over Fibre Channel Dummies Guide
Criag Carlson BIO

- Co-Chair INCITS T11
- Chair of T11 FC-NVMe working group
- FCIA Board of Directors
- NVM Express Board of Directors
- SNIA Technical Council
Rupin Mohan BIO

- Rupin Mohan is a Director of R&D and CTO of Storage Networking (SAN) at HPE Storage. Rupin leads a global engineering and product management teams responsible for development of Storage Networking products. Rupin has filed for 30+ patents at HPE. He is a Board Member of 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.
Dennis Martin BIO

- Dennis Martin is the founder and President of Demartek, a computer industry analyst organization with its own ISO 17025 accredited test lab. Demartek focuses on lab validation testing and performance testing of storage and related hardware and software products and is recognized by the EPA as an official test lab for ENERGY STAR Data Center Storage. Dennis has been working in the Information Technology industry since 1980.
David Rodgers Bio

- David J. Rodgers is Senior Product Marketing Manager with Teledyne LeCroy Protocol Solutions Group, where he focuses on defining, designing, deploying, and marketing a broad range of high-speed serial analysis test and measurement products for Ethernet and Fibre Channel SANs and LANs.

- With over 30 years of leading-edge computer industry experience, Rodgers represents Teledyne LeCroy Test and Measurement solutions in various industry standards groups, including participation on the Board of Directors for the Ethernet Alliance and the Fibre Channel Industry Association, and in the IEEE and T11 standards bodies, and involvement in the USB, PCIe and NVMe special interest groups. He began his career in the test and measurement space as an original member of the USB Implementers Forum and pioneer marketer of the first protocol specific analyzers for the USB specifications.