Expand the Power of Flash with FC-NVMe

Live Webcast
November 1, 2018
9:00 am PT
Today’s Panelists

Mark Jones
Director, Technical Marketing and Performance
Broadcom

Marcus Thordal
Principal Solution Architect
Broadcom

Rupin Mohan
Director, R&D, Head of Development, CTO SAN
HPE

Craig Carlson
Senior Technologist
Marvell Semiconductor

Dennis Martin
Senior Analyst
Principled Technologies

David Rodgers
Sr/ Product Marketing Manager
Teledyne LeCroy
Fibre Channel Industry Association

• 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 - 4\textsuperscript{th} 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
  • Library of on-demand webcasts
  • Available at FCIA BrightTALK channel [https://www.brighttalk.com/channel/14967](https://www.brighttalk.com/channel/14967)

• Social Media
  • LinkedIn, Twitter @FCIAnews
Agenda

• Marcus Thordal – 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

Marcus Thordal
Principal Solution Architect
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
  - Some applications may prefer speed over features
Elements in Application Storage Latency

Server I/O Stack

SAN

Target Services and Acceleration

Media

SAN FABRIC

Controller Services & Acceleration

Future NVMe fast path

(Raw access)

CACHE

SCSI HDD

SCSI SSD ...

NVMe SSD

FCBA

Controller Services & Acceleration

Future NVMe fast path

(Raw access)

CACHE

SCSI HDD

SCSI SSD ...

NVMe SSD

Server I/O Stack

SAN

Target Services and Acceleration

Media
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>FC-NVMe (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!
Use Case:
Concurrent Enterprise / Raw Media Storage

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
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
NVMe-oF Deployment (Ethernet)

NVMe end to end using Ethernet

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

NVMe-oF in the backend
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 Semiconductor
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
Senior Analyst, Principled Technologies
On September 17, 2018, Demartek was acquired by Principled Technologies

Combined companies have larger lab, graphics and video production capabilities

We prefer to run real-world applications to test servers and storage solutions (databases, Hadoop, VMware, etc.)

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
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

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.

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
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?
Fibre Channel T&M Today

• 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