SCSI Standards and Technology Update

Marty Czekalski
President, SCSI Trade Association
Interface and Emerging Architecture Program Manager - Seagate Technology
SNIA Legal Notice

- The material contained in this tutorial is copyrighted by the SNIA unless otherwise noted.

- Member companies and individual members may use this material in presentations and literature under the following conditions:
  - Any slide or slides used must be reproduced in their entirety without modification
  - The SNIA must be acknowledged as the source of any material used in the body of any document containing material from these presentations.

- This presentation is a project of the SNIA Education Committee.

- Neither the author nor the presenter is an attorney and nothing in this presentation is intended to be, or should be construed as legal advice or an opinion of counsel. If you need legal advice or a legal opinion please contact your attorney.

- The information presented herein represents the author's personal opinion and current understanding of the relevant issues involved. The author, the presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.

NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.
Abstract

SCSI Standards and Technology Update

SCSI continues to be the backbone of enterprise storage deployments and has rapidly evolved by adding new features, capabilities, and performance enhancements.

This talk will include an up-to-the-minute recap of the latest additions to the SAS standard and roadmaps. It will focus on the status of 12Gb/s SAS staging, advanced connectivity solutions such as MultiLink SAS™ and cover SCSI Express, a new transport of SOP (SCSI over PCIe). Presenters will also provide updates on new SCSI feature such as atomic writes, remote copy, and initial work on 24Gb/s SAS.
SCSI Adapts for Solid State Storage

- Optimized Solid State SCSI Initiative
  - SCSI Express
- New SCSI Features for FLASH and Performance
- Express Bay
- Beyond 12Gb/s SAS

SCSI SF Optics
SCSI Express Overview

- **What is SCSI Express?**
  - Proven SCSI protocol combined with PCIe creating an industry standard path to PCIe-based storage

- **Why do we need SCSI Express?**
  - Delivers proven enterprise storage for PCIe-based storage devices in a standardized ecosystem
  - Takes advantage of lower latency PCIe to improve performance
  - Offers unified management and programming interface
SCSI Express Components

- The SCSI storage command set
- Packages SCSI for a PQI queuing layer
- Flexible, high-performance queuing layer
- Accommodates PCIe, SAS, and SATA drives
- Leading server I/O interconnect
SCSI Express Controllers
- Supports SOP-PQI driver functionality on the controller to the target device on the PCIe lanes

SCSI Express Drive/Device
- SOP-PQI protocol
  - Connects to SFF-8639
  - PCIe up to x4 interface

SCSI Express Driver
- Driver supplied by storage OEMs, IHVs or OSVs
  - Open Source Linux driver and IHV drivers available

See STA website for additional SCSI Express nomenclature (www.scsita.org)
**Simple Express Devices**

- **SSDs, etc.**
  - Usually just a single logical unit with LUN 0
  - Any SCSI device type is possible
  - SSD, tape drive, optical drive (CD/DVD/BluRay), etc.
**SOP/PQI Bridges**

- **Host Bus Adapter (HBA)**
  - Bridges from PCI Express to another interconnect supporting SCSI
  - Maps SCSI target devices one-for-one
  - Usually referred to only by the back-end interconnect e.g. “SAS HBA”
  - Manage with SOP bridge management functions

<table>
<thead>
<tr>
<th>Interconnect</th>
<th>SCSI transport protocol</th>
</tr>
</thead>
<tbody>
<tr>
<td>Serial Attached SCSI (SAS)</td>
<td>Serial SCSI Protocol (SSP)</td>
</tr>
<tr>
<td>Fibre Channel (FC)</td>
<td>Fibre Channel Protocol (FCP)</td>
</tr>
<tr>
<td>Ethernet</td>
<td>Internet SCSI (iSCSI)</td>
</tr>
<tr>
<td>Universal Serial Bus (USB)</td>
<td>USB Attached SCSI (UAS)</td>
</tr>
<tr>
<td>InfiniBand</td>
<td>SCSI RDMA Protocol (SRP)</td>
</tr>
<tr>
<td>PCI Express</td>
<td>SCSI over PCI Express (SOP)</td>
</tr>
</tbody>
</table>
RAID Controllers

- Indirectly bridges from PCI Express to another interconnect supporting SCSI
- Not a one-to-one mapping of SCSI target devices
- Presents logical drives over PCI Express created from physical drive
- Manage with standard SCSI commands
- REPORT LUNS reports the logical units that have been created
SOP expects a queuing layer over PCI Express (PQI) to define

- Inbound Queues (IQs) to transfer IUs from SOP initiator port to SOP target port
- Outbound Queues (OQs) to transfer IUs from SOP target port to SOP initiator port
Circular Queue Basics

- **Element array**
  - Fixed size elements (e.g., 64 bytes)

- **Producer index (PI)**
  - Location to which producer writes elements
  - Write to element array [PI++]
  - Wrap at size of the element array

- **Consumer index (CI)**
  - Location from which consumer reads elements
  - Read from element array [CI++]
  - Wrap at size of the element array
Queue Types

- **Administrator queues**
  - Created via PQI device registers
  - Located in PQI device memory space
  - Single administrator IQ and administrator OQ
  - IUs defined by PQI

- **Operational queues**
  - Created via PQI administrator functions
    - Delivered over the administrator queues
  - Any number of operational IQs and operational OQs
    - Not in pairs
    - Can be specific to different cores
  - IUs defined by the information unit layer standard
    - e.g., SCSI over PCI Express (SOP)
Information Unit (IU) Types

- **SCSI Request/Response IUs**
  Commands, Task Management, Success, Command Response, Task Management Response, etc.

- **General Management Request/Response IUs**

- **Bridge Management Request/Response IUs**

- **Administrator Request/Response IUs**

- **Other**
  Null IU, etc.
IUs and Queues

IU smaller, equal, or larger than queue element

Circular
Avoiding Downstream PCI Express Memory Reads

IQ element array
IQ CI
PQI host

OQ element array
OQ PI
Interrupt receiver

MemWr

IQ PI

MemRd
MemWr

PQI device

MemWr
MemWr
MemWr
MemWr

OQ CI

MemWr
KEY PQI Features

- **NUMA Support**
  - Directed MSI-X
  - Flexible queue organization

- **Interrupt Coalescing**
  - Single interrupt for multiple queue entries
  - Tuning via; count, min and max times

- **Queue Element Spanning**

- **Scatter Gather Lists (SGL)**
  - Describes a data buffer and how it is distributed across non contiguous chunks of memory
  - Can be embedded in IUs or a separate list
  - Widely supported method across multiple OSs
Drivers

- **Open Source Linux Driver**
  - [https://github.com/HPSmartStorage/scsi-over-pcie](https://github.com/HPSmartStorage/scsi-over-pcie)
  - Block driver in development
    - Bypasses native Linux SCSI stack for maximum performance
    - Performance enhancements have not started
  - Low level SCSI driver is dormant
    - SCSI path enhancements proposed (2013 Kernel Summit)

- **Windows**
  - Proprietary Storport drivers available today
SCSI Express Roadmap

<table>
<thead>
<tr>
<th>CY2012</th>
<th>CY2013</th>
<th>CY2014</th>
</tr>
</thead>
<tbody>
<tr>
<td>1H'12</td>
<td>2H'12</td>
<td>1H’13</td>
</tr>
<tr>
<td>SOP proposal complete 2H 2012</td>
<td>SOP/PQI Letter Ballot SPEC Stability 2H 2012</td>
<td>SCSI Express Samples 1H 2013</td>
</tr>
</tbody>
</table>

- SOP proposal complete 2H 2012
- Express Bay available 2H 2012
- SOP/PQI Letter Ballot SPEC Stability 2H 2012
- SCSI Express Samples 1H 2013
- SCSI Express Based drivers 1H 2013
- Plugfest #1 1H 2014
- Plugfest #2 2H 2014
- SCSI Express devices/controllers available 2H 2014
New SCSI Features for FLASH and Performance

- Extended Copy Feature
- Atomic Writes
- SCSI-SF
- Power Limit Control
Extended Copy Using Tokens

- New feature allowing direct movement of data between storage devices on the same fabric
  - Leverages the ability of SCSI devices to act as both an initiator and a target
- Greatly improves performance
- Greatly reduces overhead
  - Eliminates multiple passes of data over PCIe
  - Eliminates use of system memory as a buffer
How Does Offloaded Data Transfer Work?

Server1 or Hyper-V VM1

Offload Read

Return Token

Physical Disk, VHD or SMB Shared Disk

Copy Offload Application

Token

Server2 or Hyper-V VM2

Offload Write

Return Result

Physical Disk, VHD or SMB Shared Disk

Copy Offload Application

Token

Client-Server Network

Data Movement

Storage Network

Storage Array

Storage Array

How Does Offloaded Data Transfer Work?

Server1 or Hyper-V VM1

Offload Read

Return Token

Physical Disk, VHD or SMB Shared Disk

Copy Offload Application

Token

Server2 or Hyper-V VM2

Offload Write

Return Result

Physical Disk, VHD or SMB Shared Disk

Copy Offload Application

Token

Client-Server Network

Data Movement

Storage Network

Storage Array

Storage Array
Extended Copy - Connecting the Tiers

Data Caching/Migration Application

Offload Write
Return Result

Offload Read
Return Token

DATA Movement

PCIe Fabric

SCSI Express SSD
SOP/PQI RAID
SOP/PQI Bridge (HBA)

Storage Network
Atomic Operations

- **Atomic Write** – all or nothing is written
  - For single commands and across non contiguous LBA ranges (Scatter)

- **Atomic Read** - data read is consistent at a point in time
  - No partial updates in process
  - Multiple extents (Gather)

- **Benefits:**
  - Simplifies resilient system designs
    - Database, file system, etc.
  - Improves system performance in these applications
Atomic Writes

- Single extent Atomic Writes
  - General agreement on proposal

- Multiple extent (Scatter) Atomic Writes – two versions under discussion
  - Each extent is individually atomic, but no requirement that all extents be completed
    - Reduces overhead
    - All extents must be completed or none are competed
      - Additional programming efficiencies

- Discovery of capabilities for system/application use
SCSI-SF (Simplified Features)

- SCSI contains a rich feature set with multiple methods and options
- SCSI-SF is targeted as a common subset of features for increased efficiency of implementations, qualifications and maintenance
- Proposals under discussion in T10 (Dell, HP, Intel, and Microsoft)
  - Base Feature Set
  - Provisioning Feature Set
  - Maintenance Feature Set
Power Limit Control

SSD Performance Scales with Power

\[ P \approx P_{\text{base}} + P_{\text{I/O}} \times \text{IOPs} \]

Power Consumption

- 9W
- 25W

IOPs
Power Limit Control

What power levels are supported?

e.g. 9W, 15W, 25W

Set Power Level
Express Bay Components

- 25W Power
- Cooling
- Multifunction Connector

Accessibility / Serviceability

Traditional Drive Form Factor
Express Bay

- **Express Bay**
  - Up to 25 Watts
    - For both SAS and PCIe
  - SFF-8639 connector
  - PCI-SIG electrical specification

- **Objectives**
  - Preserve the enterprise storage experience for PCI Express storage
  - Meet SSD performance demands
  - Serviceable, hot-pluggable Express Bay opens up new possibilities…
SAS Connector Compatibility

1 Max two links operate
2 Four links operational
3 Two or four links operation depending on host provisioning
Dual Port With PCIe for High Availability

Server Farm

Socket1/Root Complex1

Network1

Socket2/Root Complex2

Network2

Storage Enclosure

PCIe SSD
PCIe SSD
PCIe SSD
PCIe SSD
PCIe SSD
PCIe SSD
PCIe SSD
PCIe SSD

Dual Port With PCIe for High Availability
Wide Port SAS for Increased Throughput

2.4 GB/s full-duplex per SSD
Express Bay Summary

- Preserve the enterprise storage experience for PCI Express storage
- Meet SSD performance demands with PCIe, SAS, or SATA
- Serviceable, accessible bay offers configurability
Beyond 12Gb/s - SAS Continues to Evolve

- **SAS Roadmap Moving Forward**
  - 24 Gb/s SAS specification in development
  - Preserve 6Gb/s SAS, 12Gb/s SAS, and 6 Gb/s SATA usage models and compatibility
  - 12Gb/sec SAS controllers now shipping at >1 million IOPs

- **Proven Reliability, High Availability, and Serviceability**

- **SAS Ecosystem in Place**
  - Test & measurement equipment
  - Internal & external connectors and cabling
  - HDDs and SSDs shipping today
SAS Continues to Evolve
- Performance Gains without Protocol Changes

Performance (4K Sequential IOPS)

2009

2010

2011

2012

2013

Expected Improvements w/12Gb/s SAS
- Protocol execution
- Application hints
- OS improvements
- Controller caching

6G

900+K

450+K

300+K

150+K

80+K

3G

12G

Note: 12Gb/s SAS shipping at >1M IOPS (small block, seq.) in 2011!

6-10X performance gains in ~3 years

Feature Functionality

SCSI Standards and Technology Update
© 2013 Storage Networking Industry Association. All Rights Reserved.
SAS Roadmap

First Plugfest (leading edge)

First End-User Products (approximately 12–18 months later)


3Gb/s SAS

6Gb/s SAS

12Gb/s SAS

24Gb/s SAS
12Gb/s SAS External Interconnect

- Drive market consistency
- Simplified cable & connector options
- 2X density improvement
- Passive copper to 7m
- Active copper solution to 20m
- Active Optical (AOC) solution to 100m
- Managed connectivity standards

Cable provides active component for optical or copper

Supply Power for active cabling

SAS-2.1 standardizes OOB for active cables

Internal similar to External

Passive, Active Copper, or Optical use same connector
Managed Cable System

- New to SAS
- OoB (Out of Band) method of controlling the interface
- Every pluggable device has an EEPROM or microprocessor that communicates with the system via a low speed, two wire interface.
- Allows each port to support short passive copper cables to 100m active optical cables
STA 24Gb/s SAS MRD

- Preserve existing SAS architecture
- Continue 6Gb/s SATA compatibility
- Maintain and Support SAS backward compatibility
  - Must be backward compatible 2 generations: 6Gb/s SAS and 12Gb/s SAS
- Maximize link utilization when using devices operating at less than 24Gb/s
- Encourage improved storage system RAS attributes
- Double the transfer rate
Basic Link Budgets

SAS 3.0 Total Insertion Loss = 24dB

Higher Frequency = More Insertion Loss

SAS 4.0 Total Insertion Loss = 28dB

Basic Link Budgets
12Gb/s SAS Drive Connector

**Differential Insertion Loss**

**Common Mode Return Loss**

**Differential Return Loss**

**Differential NEXT**
SAS / Express Bay Connector

- SFF-8639 Multifunction 12 Gb/s SAS 6X Unshielded Connector

![Diagram of 68pin PCIe & SAS Combo Connector]

Vertical Receptacle mating R/A Plug
SI 24Gb/s SAS Performance

Differential Transmission

Common Mode Return Loss

Differential Reflection

Differential Near End Crosstalk
Encoding vs. Line Rate

<table>
<thead>
<tr>
<th>Encoding</th>
<th>Line Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>8b10b</td>
<td>24.0 Gbps</td>
</tr>
<tr>
<td>64b66b</td>
<td>19.8 Gbps</td>
</tr>
<tr>
<td>128b130b</td>
<td>19.5 Gbps</td>
</tr>
<tr>
<td>256b257b</td>
<td>19.28 Gbps</td>
</tr>
<tr>
<td>512b513b</td>
<td>19.24 Gbps</td>
</tr>
<tr>
<td>1024b1025b</td>
<td>19.22 Gbps</td>
</tr>
</tbody>
</table>

- Longer encoding lengths offer similar bandwidth efficiencies and yield minimal reduction in line rate
- Longer encoding lengths increase buffering requirements and increase protocol handshake latency
- SAS-4 line rate range should be 19.5 Gbps to 24Gbps
## Forward Error Correction Performance

<table>
<thead>
<tr>
<th>Encoding</th>
<th>FEC Bits</th>
<th>Overall coding length (bits)</th>
<th>Line Rate¹</th>
<th>SI Gain² @ BER of 1e-15³</th>
<th>FEC Latency Adder⁴</th>
</tr>
</thead>
<tbody>
<tr>
<td>8b10b</td>
<td>0</td>
<td>8b10b</td>
<td>24.0 Gbps</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>64b66b</td>
<td>0</td>
<td>64b66b</td>
<td>19.8 Gbps</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>128b130b</td>
<td>0</td>
<td>128b130b</td>
<td>19.5 Gbps</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>64b66b</td>
<td>14(5)</td>
<td>64b80b</td>
<td>24.0 Gbps</td>
<td>5.8 dB</td>
<td>~2.7 ns</td>
</tr>
<tr>
<td>128b130b</td>
<td>16(5)</td>
<td>128b146b</td>
<td>21.9 Gbps</td>
<td>5.8 dB</td>
<td>~5.3 ns</td>
</tr>
<tr>
<td>256b257b</td>
<td>18(5)</td>
<td>256b275b</td>
<td>20.63 Gbps</td>
<td>5.6 dB</td>
<td>~10.6 ns</td>
</tr>
<tr>
<td>512b513b</td>
<td>20(5)</td>
<td>512b533b</td>
<td>19.99 Gbps</td>
<td>5.6 dB</td>
<td>~21.2 ns</td>
</tr>
<tr>
<td>1024b1025b</td>
<td>88(6)</td>
<td>1024b1113b</td>
<td>20.87 Gbps</td>
<td>7.4 dB</td>
<td>~53.2 ns</td>
</tr>
</tbody>
</table>

¹Raw data throughput of 19.2Gb/s.
²SI gain is addition IL that the system can tolerate (~2x the FEC gain at the slicer)
³Assumes 1e-15 as a target BER.
⁴Additional latency imposed by use of FEC
⁵Differential encoding and BCH algorithm for FEC.
⁶Reed-Solomon algorithm with T=4 for FEC.
## Channel Lengths for a Range of Dielectric Materials

<table>
<thead>
<tr>
<th>Material</th>
<th>8b10b 24 Gbps</th>
<th>64b66b 19.8 Gbps</th>
<th>64b66B + FEC 24 Gbps</th>
<th>128b130b + FEC 21.9 Gbps</th>
<th>256b257b + FEC 20.63 Gbps</th>
<th>512b513b + FEC 19.99 Gbps</th>
<th>1024b1025b + FEC 20.87 Gbps</th>
</tr>
</thead>
<tbody>
<tr>
<td>Megtron6</td>
<td>18.2”</td>
<td>22.1”</td>
<td>25.8”</td>
<td>28.3”</td>
<td>29.7”</td>
<td>30.6”</td>
<td>32.3”</td>
</tr>
<tr>
<td>Nelco 4000-13SI</td>
<td>16.4”</td>
<td>19.9”</td>
<td>23.2”</td>
<td>25.5”</td>
<td>26.8”</td>
<td>27.6”</td>
<td>29.2”</td>
</tr>
<tr>
<td>Nelco 4000-13</td>
<td>15.6”</td>
<td>18.9”</td>
<td>22.0”</td>
<td>24.1”</td>
<td>25.3”</td>
<td>26.1”</td>
<td>27.6”</td>
</tr>
<tr>
<td>Nelco 4000-6</td>
<td>8.3”</td>
<td>10.1”</td>
<td>11.8”</td>
<td>12.9”</td>
<td>13.6”</td>
<td>14.0”</td>
<td>14.8”</td>
</tr>
<tr>
<td>FR4 (Nelco 4000)</td>
<td>4.2”</td>
<td>5.1”</td>
<td>5.9”</td>
<td>6.5”</td>
<td>6.8”</td>
<td>7.0”</td>
<td>7.4”</td>
</tr>
<tr>
<td>SAS3 Cable</td>
<td>103” (2.6m)</td>
<td>125” (3.2m)</td>
<td>146” (3.7m)</td>
<td>160” (4.1m)</td>
<td>168” (4.3m)</td>
<td>174” (4.4m)</td>
<td>183” (4.7m)</td>
</tr>
<tr>
<td>SAS4 Cable¹</td>
<td>136” (3.5m)</td>
<td>164” (4.2m)</td>
<td>192” (4.9m)</td>
<td>211” (5.4m)</td>
<td>222” (5.6m)</td>
<td>229” (5.8m)</td>
<td>241” (6.1m)</td>
</tr>
<tr>
<td>Latency</td>
<td>~2.7 ns</td>
<td>~5.3 ns</td>
<td>~10.6 ns</td>
<td>~21.2 ns</td>
<td>~53.2 ns</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Conclusions

- 24Gb/s is definitely feasible
- Will need more efficient encoding
- T10 to investigate forward error correction implications
- Better board materials can help
The SNIA Education Committee thanks the following individuals for their contributions to this Tutorial.

**Authorship History**

Marty Czekalski

**Updates:**

**Additional Contributors**

Joe Foster
Rob Elliot
Mike James
Dave Allen
Greg McSorley
Tim Symons
STA members

Please send any questions or comments regarding this SNIA Tutorial to tracktutorials@snia.org