SCSI Standards and Technology Update:
SAS and SCSI Express

Marty Czekalski - President, SCSI Trade Association – Seagate Technology
Greg McSorley – Vice President, SCSI Trade Association - Amphenol
SCSI Update

- SAS Overview Performance Roadmap and 12Gb/s SAS staging
- MultiLink SAS™ and Advanced Connectivity Connectivity Update
- SCSI Express
  - Express Bay
    - Backplane Receptacle Compatibility
  - SOP/PQI
- New SCSI Features Update
SCSI Logical Abstraction Layer: A foundation for Innovation

- Preserves Hardened SCSI Command Set
  - Successive Product Generations
  - Frequent Technology Perturbations
  - Multiple Vendors
  - Multiple Interconnects
  - Reduces Time to Market and Integration Costs
- Delivers Enterprise Attributes and Features
  - End-to-End Data Protection
  - Atomic Writes
  - Hinting
  - Task Management
  - Power Management
- And more on the way
SCSI Logical Abstraction Layer: A foundation for Innovation

- Operates Over Numerous Transport Layers
  - ATAPI (ATA, SATA)
  - USB
  - Memory sticks
  - Firewire
  - Infiniband
  - iSCSI
  - FC
  - Parallel SCSI
  - SAS

- And now, **SCSI Over PCIe (SOP, PQI)**

**SCSI:** The Most Widely Implemented Logical Storage Protocol
SAS – Preservation and Innovation

Enterprise Storage

Investment Protection

Platform For Innovation

- Slot serviceable
- Reliability
- Scalability
- Usability

- Diverse platform usage
- New technology adoption
- Continued industry investment

- Preserves Logical SCSI
- Enterprise Middleware
- Preserves, device types and data formats

www.scsita.org
Mature - Market Readiness

Aware Applications
- OS Support
- Device Drivers
- Management

Server
- SAS ROC
- SAS Controller
- SAS Slot
- PCI Slots

SAS ROC or Controller
- HDD
- SSD

Connectors
- Cables
- Testers
- Analyzers
- Converters

Consultants
- Enclosures
- Compatibility Test
- System Integrators
- OSVs
- Backplanes

SAS Switch
Scaling SAS Architecture

- Host Attach SAN or NAS or SAS
- SAS Switch (packaged view)
- Embedded Controllers SAS Connected
- SAS Switch
- Various JBODs
  - SATA HDDs
  - SATA HDDs
  - SAS HDDs
- SAS HDDs
SAS Performance 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


*SAS Roadmap updated Nov 2010.*
12Gb/s SAS Market Objectives

- Preserve existing SAS architecture
  - Preserve 3Gb/s SAS & 6Gb/s SAS usage models
  - Maintain backward compatibility and intermatability to current 3Gb/s & 6Gb/s SATA/SAS backplane device connectors
- Continue 6Gb/s SATA and future SATA compatibility
- Encourage improved storage system RAS attributes
- Double transfer rate
  - Improve cost/performance & power/bandwidth ratio
  - Reduced # of connections per Gb/s
- Maximize link utilization when using devices operating at less than 12Gb/s
  - Frame buffering in expanders
- Maintain and support SAS Advanced Connectivity roadmap
  - Encourage broad adoption of Mini-SAS HD required for external
- Supports MultiLink SAS™ implementations
12Gb/s SAS Industry Timelines

- T&M Equipment & Protocol Analyzers - Available
- Internal & External Connectors - Available
- Early Component/HBA Testing – Available to partners
- HDDs and SSDs – Available to partners
- Industry Plugfest – Occurred week of Aug 20th 2012
  - Next plugfest early 2013
- Standard Status
  - SPL-2 in final comment resolution
  - SAS-3 letter ballot expected Nov 2012
- System Shipments - 2H 2013
SAS Advanced Connectivity

- Drive market consistency
- Simplify cable & connector options
- 2X density improvement
- Provide converged high-density connectivity
- Provide managed connectivity standards
- Provide active copper solution to 20m
- Provide optical solution to 100m
- Support 6Gb/s SAS deployments
- External solution for 12Gb/s SAS deployments

Supply power here for active cabling

SAS-2.1 standardizes OOB for active cables

Cable provides active component for optical or copper

Passive, Active Copper, or Optical use same connector

Internal similar to External

2012 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.
SAS Advanced Connectivity Roadmap – 2011

100 meters (optical)

20 meters (active copper)

6 – 10 meters (passive copper)

SAS-1/SAS 1.1
SAS-2
SAS-2.1
SAS-3

Mini-SAS (SFF-8087/8088)

Mini-SAS Active Copper (SFF-8088)

Mini-SAS HD (SFF-8643/8644)

SFF-8470 (InfiniBand)

SFF-8088

www.scsita.org
12 Gb/s SAS Connectivity Update

12Gb/s HDD/Backplane Connectors
Redesign current 6Gb/s connector to run 12Gb/s
- 6Gb/s – SFF-8482  ➞  12Gb/s – SFF-8680

Parts Availability for Vertical Mid-plane Receptacles
- Available now

Parts Availability for Right angle Plugs for HDD’s
- Available now
MultiLink SAS™ Initiative

- STA Endorsed May 2010
- Defines New Type of SAS Drive Slot
  - Increased power/slot
  - Multiple SAS links
- No Protocol Changes
  - Works with existing infrastructure components
- 12Gb/s and 24Gb/s Link Performance
- Standards in T10 (SAS-3) & SFF (SFF8630)
MultiLink SAS™ Slot

- MultiLink SAS™
  - High performance (20+W per slot)
  - Hot swap, serviceability
  - High availability (2 fault domains possible)
  - Low implementation risk:
    - Integrates seamlessly with existing systems
      - Fully hardened protocol stack
      - Common management stack
  - Low investment
  - Flexible: Independent SSDs or wide port SSDs
  - Caching solutions isolate tier-tier traffic w/o accessing system memory
MultiLink SAS™ Connector

Port 0 (Primary)

Port 1 (Secondary)

Power

Sidebands

Port 0

Port 1

Port 2

Port 3
MultiLink SAS™ Roadmap: Backplane Slot Location

- SAS Connector (SFF-8482 & SFF-8680)
- MultiLink SAS Connector (SFF-8630)

Bandwidth (Gb/s, Full Duplex):
- 3Gb/s (1x1)
- 6Gb/s (2x1)
- 12Gb/s (2x1, 2x2)

2012 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.
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?
  - Deliver proven enterprise storage for PCIe based storage devices
  - Take advantage of lower latency PCIe to improve performance
  - Unified management and programming interface
SCSI Express Value Proposition

**Performance and Innovation**

- Increased performance through lower latency for emerging advanced technologies
- Enables new storage architectures

**Reliability**

- Proven enterprise SCSI ecosystem
- Architected for nonstop availability

**Investment Protection**

- Coexistence with SAS via Express Bay and common command set
- Leveraging robust middleware ecosystem
# SCSI Express Components

Existing industry initiatives delivering enterprise storage using PCI Express

<table>
<thead>
<tr>
<th>Technology</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCSI</td>
<td>The storage command set</td>
</tr>
<tr>
<td>SCSI Over PCIe (SOP)</td>
<td>Packages SCSI for a PQI queuing layer</td>
</tr>
<tr>
<td>PCIe Queuing Interface (PQI)</td>
<td>Flexible, high-performance queuing layer</td>
</tr>
<tr>
<td>Express Bay connector</td>
<td>Accommodates PCIe, SAS, and SATA drives</td>
</tr>
<tr>
<td>(SFF 8639)</td>
<td>Leading server I/O interconnect</td>
</tr>
<tr>
<td>PCI Express</td>
<td></td>
</tr>
</tbody>
</table>
Express Bay

- Up to 25 Watts
- 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…
SFF-8639 Multifunction Connector

Source: SFF-8639 Rev 1.2, July 2012
Connector Compatibility

1 Max two links operate
2 Four links operational
3 Two or four links operation depending on host provisioning
### SCSI Express Controllers
- Supports SOP-PQI driver functionality on the controller to the target device on the PCIe lanes
- Typically supports SAS/SATA devices

### 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 drivers planned
- Enables direct attach of SCSI Express devices to a host
Final Proposals for Approval

SOP and PQI Standards in letter ballot comment resolution process – Target completion at end of year
Key excerpts from SCSI, SOP, and PQI architecture models

SCSI initiator device: a server with a PCI Express Root Port
SCSI target device: an SSD, HDD, HBA, or RAID controller
**Simple 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.
Bridges

HBAs

- Bridges from PCI Express to another interconnect supporting SCSI
  - Maps SCSI target devices one-for-one
- Typical terms: host bus adapter (HBA), host controller, host adapter, network interface controller, converged network adapter
- 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>

2012 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.
RAID controllers

- Less complex than bridges from an SOP perspective
  - 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 drives
    - Manage with standard SCSI commands
      - REPORT LUNs reports the logical units that have been created
      - Bridge management not involved (unless it’s a hybrid HBA + RAID controller)
Queuing layer services

- SOP expects a queuing layer over PCI Express to define
  - inbound queues
    - transfer IUs from SOP initiator port to SOP target port
  - outbound queues
    - transfer IUs from SOP target port to SOP initiator port
- SOP architected to support multiple queuing layers
  - PCI Express Queuing Interface (PQI)
- Information Units (IUs)
  - Messages between a driver and a device
Information Units (IUs) 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.
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
Inbound Queues (IQs) and Outbound Queues (OQs)

- Named from the PQI device’s perspective

  - **Inbound queues (IQs)**
    - PQI host to PQI device
      - Administrator request IUs
      - SCSI request IUs (in SOP)

  - **Outbound queues (OQs)**
    - PQI device to PQI host
      - Administrator response IUs
      - SCSI response IUs (in SOP)
IQ and OQ Object Locations

- Avoids PCI Express memory reads of PIs and CIs

**IQ object locations**
- IQ element array in host memory (typical)
- IQ PI in PQI device memory space (always)
- IQ CI in host memory (typical)

**OQ object locations**
- OQ element array in host memory (typical)
- OQ PI in host memory (typical)
- OQ CI in PQI device memory space (always)
Queue Types

- Administrator queues and operational queues

- **Administrator queues**
  - Created via PQI device registers
  - Located in PQI device memory space
  - Single administrator IQ and administrator OQ
    - i.e., one administrator queue pair
  - 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
  - IUs defined by the information unit layer standard
    - e.g., SCSI over PCI Express (SOP)
SCSI Express Timeline
STA Approved timeline – updated as of May 2012

<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></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Today</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **SOP/PQI Letter Ballot**
  - SPEC Stability
  - 2H 2012

- **Express bay available**
  - 2H 2012

- **SOP proposal complete**
  - 2H 2012

- **SCSI Express Samples**
  - 1H 2013

- **SCSI Express Based drivers**
  - 1H 2013

- **Plugfest #1**
  - 2H 2013

- **Plugfest #2**
  - 1H 2014

- **SCSI Express devices/controllers available**

2012 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.
SCSI Express Summary

- Proven SCSI protocol combined with PCIe creating an industry standard path to PCIe-based storage
  - Enterprise storage for PCIe based storage devices
  - Increased performance through lower latency
  - Coexistence with SAS via Express Bay and common command set
- Unified management and programming interface

SCSI Trade Association

SCSI EXPRESS
SCSI – Looking to the Future

- SCSI Express
- 12Gb/s, 24Gb/s SAS
  - Performance and scalability
- Power Limit Control - up to 25W devices
  - Both SAS and SCSI Express
- Extended Copy Feature
- Atomic Writes
- Hinting & other NVM features
- Works in process for PCIe storage
Power Limit Control - up to 25W Devices

- Allows system to decide how much power a device is allowed to consume. Both less or more than nominal.

- For SSDs power = performance

- SAS SSDs will be able to take advantage of extra power provisioning above the nominal 9W limit in typical 2.5” slots.

  - Example: SAS drive in a SCSI Express bay (SFF-8639) could be allowed to consume up to 25W
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

Client-Server Network

Offload Write

Return Result

Physical Disk, VHD or SMB Shared Disk

Storage Array

Storage Network

Storage Array

Application To
cken

Data Movement

Copy Offload Application

Token
Extended Copy - Connecting the Tiers

Data Caching/Migration Application

Token

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 Writes

- All or nothing written capability across multiple commands
  - For single commands and across non contiguous LBA ranges

- Benefits:
  - Simplifies resilient system designs
    - Database, file system, etc.
  - Improves system performance in these applications
Hinting & Other NVM Features

- Pass “hints” to devices to make operations more efficient and increase performance
  - Targeted at SSDs and hybrid drives, but also useful for HDDs

- Direct attached devices don’t need to continually OPEN and CLOSE connections
  - Can be implemented within the existing standard
  - Reduces latency on both SSDs and HDDs

- NVM features and programming interfaces
  - Leverages ongoing work in SNIA and T10
Works in Process for PCIe Storage

- Asynchronous insertion/removal (true hot plug) not supported yet
  - Requires Downstream Port Control (DPC)
    - ECN – Scheduled for future chipsets/switches
- Work in process in PCI SIG
  - SFF-8639 electrical (channel model) specifications
Summary

- SAS and SCSI continue to innovate and provide value in the storage ecosystem
- Activity is happening in real time
- Follow activities in T10 and SFF to make sure your designs take full advantage of these improvements