Serial Attached SCSI
General Overview

by Rob Elliott
HP Industry Standard Servers
Server Storage Advanced Technology
elliott@hp.com  http://www.hp.com
30 June 2003
Notice

- These slides are freely distributed by HP through the SCSI Trade Association (http://www.scsita.org)
- STA members are welcome to borrow any number of the slides (in whole or in part) for other presentations, provided credit is given to the SCSI Trade Association and HP
- This compilation is © 2003 Hewlett-Packard Corporation
SCSI – Small Computer System Interface
## History of parallel SCSI

<table>
<thead>
<tr>
<th>Interconnect</th>
<th>Standard</th>
<th>Year</th>
<th>Speed</th>
<th>Key features</th>
</tr>
</thead>
<tbody>
<tr>
<td>SASI</td>
<td></td>
<td>1979</td>
<td></td>
<td>Shugart Associates</td>
</tr>
<tr>
<td>SCSI-1</td>
<td>SCSI-1</td>
<td>1986</td>
<td>~2 MB/sec</td>
<td>Asynchronous, narrow</td>
</tr>
<tr>
<td>SCSI-2</td>
<td>SCSI-2</td>
<td>1989</td>
<td>10 MB/sec</td>
<td>Synchronous, wide</td>
</tr>
<tr>
<td>SCSI-3</td>
<td></td>
<td></td>
<td></td>
<td>Split command sets, transport protocols, and physical interfaces into separate standards</td>
</tr>
<tr>
<td>Fast-Wide</td>
<td>SPI/ SIP</td>
<td>1992</td>
<td>20 MB/sec</td>
<td></td>
</tr>
<tr>
<td>Ultra</td>
<td>Fast-20 annex</td>
<td>1995</td>
<td>40 MB/sec</td>
<td></td>
</tr>
<tr>
<td>Ultra 2</td>
<td>SPI-2</td>
<td>1997</td>
<td>80 MB/sec</td>
<td>LVD</td>
</tr>
<tr>
<td>Ultra 3</td>
<td>SPI-3</td>
<td>1999</td>
<td>160 MB/sec</td>
<td>DT, CRC</td>
</tr>
<tr>
<td>Ultra 320</td>
<td>SPI-4</td>
<td>2001</td>
<td>320 MB/sec</td>
<td>Paced, Packetized, QAS</td>
</tr>
</tbody>
</table>
With SCSI-3, SCSI has been broken into multiple standards (and has dropped the -3 moniker)

- Allows support for additional SCSI transport protocols

**SCSI standards structure**

- Device-type specific command sets (e.g., SBC-2, SSC-2, MMC-3)
- Primary command set (shared for all device types) (SPC-3)
- Protocols (e.g., SPI-4, FCP-2, SAS SSP)
- Interconnects (e.g., SPI-4, Fibre Channel, SAS)
There have been many serial SCSI protocols

<table>
<thead>
<tr>
<th>Interconnect</th>
<th>Standard</th>
<th>Year</th>
<th>Typical speed</th>
<th>Key features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fibre Channel</td>
<td>FCP</td>
<td>1995</td>
<td>100 MB/ sec</td>
<td>Optical</td>
</tr>
<tr>
<td>Serial Storage Architecture (SSA)</td>
<td>SSA-S2P, SSA-TL1,</td>
<td>1996</td>
<td>20 MB/ sec</td>
<td>IBM only vendor</td>
</tr>
<tr>
<td></td>
<td>SSA-PL1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Serial Storage Architecture (SSA)</td>
<td>SSA-S3P, SSA-TL2,</td>
<td>1997</td>
<td>40 MB/ sec</td>
<td>IBM only vendor</td>
</tr>
<tr>
<td></td>
<td>SSA-PL2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FireWire (IEEE 1394)</td>
<td>SBP-2</td>
<td>1998</td>
<td>50 MB/ sec</td>
<td></td>
</tr>
<tr>
<td>Fibre Channel</td>
<td>FCP-2</td>
<td>2002</td>
<td>200 MB/ sec</td>
<td></td>
</tr>
<tr>
<td>InfiniBand</td>
<td>SRP</td>
<td>2002</td>
<td>250 MB/ sec</td>
<td>4x, 12x too</td>
</tr>
<tr>
<td>Ethernet</td>
<td>iSCSI</td>
<td>2003</td>
<td>~100 MB/ sec</td>
<td>Gigabit Ethernet</td>
</tr>
</tbody>
</table>
ATA - AT Attachment
## History of parallel ATA

<table>
<thead>
<tr>
<th>Generation</th>
<th>Standard</th>
<th>Year</th>
<th>Speed</th>
<th>Key features</th>
</tr>
</thead>
<tbody>
<tr>
<td>IDE</td>
<td>Pre-standard</td>
<td>1986</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EIDE</td>
<td>ATA</td>
<td>1994</td>
<td></td>
<td>PIO modes 0-2, multiword DMA 0</td>
</tr>
<tr>
<td></td>
<td>ATA-2</td>
<td>1996</td>
<td>16 MB/sec</td>
<td>PIO modes 3-4, multiword DMA modes 1-2, LBAs</td>
</tr>
<tr>
<td></td>
<td>ATA-3</td>
<td>1997</td>
<td>16 MB/sec</td>
<td>SMART</td>
</tr>
<tr>
<td></td>
<td>ATA/ ATAPI-4</td>
<td>1998</td>
<td>33 MB/sec</td>
<td>Ultra DMA modes 0-2, CRC, overlap, queuing, 80-wire</td>
</tr>
<tr>
<td>Ultra DMA 66</td>
<td>ATA/ ATAPI-5</td>
<td>2000</td>
<td>66 MB/sec</td>
<td>Ultra DMA mode 3-4</td>
</tr>
<tr>
<td>Ultra DMA 100</td>
<td>ATA/ ATAPI-6</td>
<td>2002</td>
<td>100 MB/sec</td>
<td>Ultra DMA mode 5, 48-bit LBA</td>
</tr>
<tr>
<td>Ultra DMA 133</td>
<td>ATA/ ATAPI-7</td>
<td>2003</td>
<td>133 MB/sec</td>
<td>Ultra DMA mode 6</td>
</tr>
</tbody>
</table>
ATA standards architecture

ATA register-delivered command set (ATA/ATAPI-7 Volume 1)

Device-type specific command sets (e.g., MMC-3)

Primary command set (shared for all device types) (SPC-3)

ATAPI (Packet delivered command set) (ATA/ATAPI-7 Volume 1)

ATA/ATAPI register set (ATA/ATAPI-7 Volume 1)

Protocols (e.g., SAS STP, ATA/ATAPI-7 Volume 2 (parallel ATA), ATA/ATAPI-7 Volume 3 (Serial ATA))

Interconnects (e.g., SAS, ATA/ATAPI-7 Volume 2 (parallel ATA), ATA/ATAPI-7 Volume 3 (Serial ATA))
## History of Serial ATA

<table>
<thead>
<tr>
<th>Generation</th>
<th>Standard</th>
<th>Year</th>
<th>Speed</th>
<th>Key features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Serial ATA</td>
<td>ATA/ ATAPI-7</td>
<td>2002</td>
<td>150 MB/sec</td>
<td></td>
</tr>
<tr>
<td>Serial ATA III</td>
<td>ATA/ ATAPI-9?</td>
<td>?</td>
<td>600 MB/sec</td>
<td></td>
</tr>
</tbody>
</table>
Different classes of disk drives

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Mobile</th>
<th>Desktop</th>
<th>Enterprise</th>
</tr>
</thead>
<tbody>
<tr>
<td>rpm</td>
<td>3600, 4200, 5400 rpm</td>
<td>5400, 7200 rpm</td>
<td>10K, 15K rpm</td>
</tr>
<tr>
<td>Seek time</td>
<td>12 - 14 ms</td>
<td>8.9 - 9.5 ms</td>
<td>3.2 - 7.4 ms</td>
</tr>
<tr>
<td>Performance as file server*</td>
<td>N/A</td>
<td>79 - 136</td>
<td>146 - 366</td>
</tr>
<tr>
<td>Write cache</td>
<td>2 MB</td>
<td>2 - 8 MB</td>
<td>2 - 8 MB</td>
</tr>
<tr>
<td>Capacity</td>
<td>10 - 80 GB</td>
<td>40 - 250 GB</td>
<td>18, 36, 72, 144, 180 GB</td>
</tr>
<tr>
<td>Reliability</td>
<td>300 K hr MTBF</td>
<td>500 K hr MTBF</td>
<td>1.2 M hr MTBF</td>
</tr>
<tr>
<td>Power</td>
<td>2.5 W</td>
<td>10 W</td>
<td>15 W</td>
</tr>
<tr>
<td>Cost</td>
<td>$73 - $160</td>
<td>$75 - $240</td>
<td>$160 - $1400</td>
</tr>
<tr>
<td>Interfaces</td>
<td>ATA/ 66, ATA/ 100</td>
<td>ATA/ 100, ATA/ 133</td>
<td>Ultra 160 SCSI, Ultra 320 SCSI, FC</td>
</tr>
</tbody>
</table>

* Benchmark of many drives on [http://www.storagereview.com](http://www.storagereview.com)

Note: As of mid 2002
Bringing SCSI and ATA together...
Serial Attached SCSI (SAS)

- Serial Attached SCSI (SAS) supports both SCSI and ATA
- Three transport protocols
  - Serial SCSI Protocol (SSP)
    - Supports SAS (SCSI) disk drives, tape drives, etc.
  - Serial ATA Tunneling Protocol (STP)
    - Supports Serial ATA disk drives
  - Serial Management Protocol (SMP)
    - Supports SAS expanders
Positioning SATA, SAS, and FC

Descriptive text and diagrams explaining the positioning and usage of SATA, SAS, and FC in different environments, including desktop PC/workstation, mainstream servers/direct attached storage, and storage area network. The diagram illustrates the compatibility and design similarities between SATA and SAS, as well as the differences in flexibility and simplicity between SATA and SAS for different use cases.

Key points:
- SATA: Desktop-class disk drives (single port)
- SAS: Enterprise-class disk drives (single or dual port)
- FC: Enterprise-class disk drives (dual port)
- PC Chipset with SAS support, SAS HBA card, or SAS RAID card to drives or JBODs
- FC HBA to external storage
- Configuration and flexibility
- One host design accommodates SATA for low cost bulk storage or SAS for performance & reliability in mission critical applications
- Compatible physical layer between SATA and SATA
- Similar HDD ASIC and firmware design between SAS and FC
Direct attach = Number of drives limited to number of ports in the HBA

Expander attach = More drives than HBA ports

STP (Serial ATA Tunneling Protocol) used over SAS physical links when communicating with SATA drives through expanders

SSP (Serial SCSI Protocol) used to communicate with SAS drives

SATA (Serial ATA) used to communicate with SATA drives over SATA physical links
Serial Attached SCSI timeline

May 2002
Serial Attached SCSI introduced to T10

2002
Serial Attached SCSI Product Design

November 2002
Serial Attached SCSI standard to T10 letter ballot

2003
Serial Attached SCSI Test and Integration

May 2003
Serial Attached SCSI to ANSI public review

2004
Serial Attached SCSI Product Availability

September 2003
Serial Attached SCSI to ANSI publication

2002

2003

2004

SAS-1.1 for bug fixes

SAS-2 for 6.0 Gbps

1Q03 ASIC testing

4Q03 Emulator testing
SCSI Trade Association roadmap

- Parallel SCSI Ultra 320
- Serial Attached SCSI SAS 300
- Serial Attached SCSI SAS 600
- Serial Attached SCSI SAS 1200

SAS press releases (part 1)

- **11/26/2001 Compaq/ IBM/ LSI Logic/ Maxtor/ Seagate**
  - Industry Leaders Team to Form Serial Attached SCSI Working Group to Address Next-Generation Interconnect Needs

- **2/4/2002 STA**
  - SCSI Trade Association and Serial Attached SCSI Working Group Announce Agreement

- **5/6/2002 STA**
  - INCITS Technical Committee T10 Accepts Proposal to Begin Work on Serial Attached SCSI Standard

- **6/12/2002 Dell/ HP/ Intel**
  - Industry Leaders Collaborate on New Choices for Next-Generation Serial Architectures for Server Storage

- **1/9/2003 STA**
  - Serial Attached SCSI Letter Ballot Released by T10

- **1/20/2003 STA**
  - STA and SATA II Working Group Agree to Collaborate on Specification Alignment

- **1/20/2003 HP/ Seagate**
  - Seagate Demonstrates Industry’s First Serial Attached SCSI Storage Solutions with HP

- **1/20/2003 LSI Logic**
  - LSI Logic and Leading Hard Disk Drive Manufacturers Initiate Serial Attached SCSI Product Planning
SAS press releases (part 2)

- 3/18/2003 Adaptec
  - Adaptec First Host Bus Adapter Provider to Transmit Serial Attached SCSI Protocol Packet

- 3/25/2003 Hitachi/LSI Logic
  - Hitachi and LSI Logic to Accelerate Delivery of Serial Attached SCSI

- 4/15/2003 Emulex/Intel
  - Emulex and Intel to Develop first Storage Processors for Serial ATA, SAS and Fibre Channel within a Single Architecture

- 4/21/2003 Adaptec/Fujitsu/Hitachi/Maxtor/Seagate
  - Adaptec Teams with Fujitsu, Hitachi, Maxtor and Seagate to Bring First Serial Attached SCSI Solutions to Market

- 4/29/2003 LSI Logic/Seagate/Tabernus
  - Tabernus, LSI Logic and Seagate Technology Work Together for Serial Attached SCSI

- 5/9/2003 LSI Logic
  - LSI Logic First To Demonstrate Serial Attached SCSI Initiator-to-Target Functionality with Full Read/Write Validation

- 6/11/2003 Adaptec/Maxtor/Tabernus
  - Tabernus, Adaptec, and Maxtor Work to Ensure Compatibility of Serial Attached SCSI Solutions, Speed Time to Market
SAS press releases (part 3)

- **6/16/2003 Adaptec/HP/Seagate**
  - Adaptec, HP and Seagate Technology Team to Debut Compatibility of Serial Attached SCSI and Serial ATA at CeBIT America
- **6/24/2003 LSI Logic/Maxtor**
  - LSI Logic and Maxtor Extend Testing of Serial Attached SCSI Prototypes
- **6/26/2003 LSI/Seagate**
  - LSI Logic Teams With Seagate to Drive Serial Attached SCSI Interoperability
- **6/30/2003 Maxtor**
  - Maxtor Demonstrates Early Success with Serial Attached SCSI Interface

**to be continued…**
Serial Attached SCSI standard
SAS standard layering

- **SCSI application layer**
- **ATA application layer**
- **Management application layer**
  - **SSP transport layer**
  - **STP transport layer**
  - **SMP transport layer**
  - **SAS port layer**
    - **SSP link layer**
    - **STP link layer**
    - **SMP link layer**
    - **SAS link layer**
    - **SAS phy layer**
    - **SAS physical layer**

**Features and Functions**

- **SCSI:** mode pages, log pages, spinup
- **Mgmt:** SMP functions
- **Frame definitions**
- **Wide port handling**
- **Frame transmission**
- **Connection management Primitives**
- **OOB, speed negotiation**
- **8b10b encoding**
- **Electrical specs**
- **Cables and connectors**

6/30/2003 SAS general overview page 23
Key features of SAS by layer

- **Physical layer** - connectors
- **Architecture** - Expanders and topologies
- **Phy layer** - OOB and speed negotiation
- **Link layer** - connections
- **Transport layer** - frames
- **Wrap Up**
Physical layer
• SATA 1.0 was defined for internal use only (e.g. inside a PC)
• 1 m internal cable
• No external connectors/ cables
Appearance of Serial ATA Connectors
(Drawing courtesy of Molex)

Device plug connector

Serial ATA signal connector (pin S1)

Device connector sizes and locations

<table>
<thead>
<tr>
<th>Type</th>
<th>Dimensions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Serial</td>
<td>2.5&quot;</td>
</tr>
<tr>
<td></td>
<td>power</td>
</tr>
<tr>
<td></td>
<td>signal</td>
</tr>
<tr>
<td>Serial</td>
<td>3.5&quot;</td>
</tr>
<tr>
<td></td>
<td>power</td>
</tr>
<tr>
<td></td>
<td>signal</td>
</tr>
<tr>
<td>Legacy Power</td>
<td>(vendor specific)</td>
</tr>
<tr>
<td>Parallel</td>
<td>3.5&quot;</td>
</tr>
<tr>
<td>parallel ATA signals</td>
<td>4-pin power</td>
</tr>
</tbody>
</table>

(5.25" form factor also defined for devices like tape drives and DVDs)

in comparison...
Similarly, SAS defines internal environments.

Backplanes support two physical links.

Cables mainly support one phy (for debug, two physical link cables might be useful).

HP proposing 4x internal connector for SAS-1.1
SAS plug and backplane receptacle connectors

SAS primary physical link

SAS secondary physical link (on backside)

Power

SATA physical link

Serial attached SCSI

SAS backplane receptacle connector

Power

SATA/ SAS primary physical link

SAS secondary physical link

Note: SATA backplane connectors Will NOT accept SAS drives
• SAS defines an external (box-to-box) environment using InfiniBand 4x connectors and cables

External cabled environment

(SAS external cable connects the Tx signal pins to the Rx signal pins on each physical link)
SAS external receptacle connector

- (picture courtesy FCI/ Berg)
Architecture
Physical links and phys

- A phy contains one transceiver
- A physical link attaches two phys together
Differential signaling

Positive signal (single-ended) (non-inverted)

1500 mV
1200 mV
900 mV
0 V

Common mode voltage (the level is not very important)

Negative signal (single-ended) (inverted)

1500 mV
1200 mV
900 mV
0 V

Common mode voltage

Differential signal

600 mV
0 V

Differential signal is immune to noise common to both single-ended signals

positive - negative
1500-900 = 600 mV
or
900-1500 = -600 mV

101011

1 0 1 0 1 1
Physical link rate

- Each direction runs 1.5 Gbps or 3.0 Gbps (150 MB/sec or 300 MB/sec)
  - Both directions use the same physical link rate
- Dual simplex (full duplex) operation – 600 MB/sec total bandwidth
- Example: peak bandwidth needs of an HBA with 8 phys
  - 2400 MB/sec half duplex, 4800 MB/sec full duplex
SAS devices and ports

- SAS devices contain ports
- Ports contain phys
- Ports are virtual constructs
  - Groups of phy with the same SAS address, attached to another group of phys with the same SAS address

Each horizontal line represents a differential signal pair.
Each SAS port and expander device has a worldwide unique 64-bit SAS address

Same namespace as the Fibre Channel Port_Name

<table>
<thead>
<tr>
<th>Byte\Bit</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NAA (5h)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>IEEE Company ID (24 bits)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>Vendor-Specific Identifier (40 bits)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
End devices

- End device is a SAS device that is not an expander device
- Sample end devices
  - HBA - 8 phys
    - One SAS address for all 8 phys
      - Potentially all one (very) wide port
    - One SAS address for 4 phys, another SAS address for 4 phys
      - Guarantees at least two ports
      - Good match for 4-wide connectors
  - Eight SAS addresses
  - Disk drive - 2 phys
    - Separate SAS address for each phy
    - Guarantees two ports
    - Never a wide port
Expander devices

- Expander device contains expander ports
- May contain SAS devices too (e.g. for enclosure management)
- Each expander device has a SAS address
- Each expander phy has a phy identifier unique within that expander device
Expander device types - edge vs. fanout

- **Edge expander device**
  - Always part of an “edge expander device set”
  - May perform subtractive routing
- **Fanout expander device**
  - Never does subtractive routing
  - Usually supports larger tables for table routing
- **Topologies described later**
• A simple SAS domain contains SAS devices and expander devices.

• An ATA domain contains a SATA host and a SATA device.

Note: When expander devices are present, SAS target ports may be located in SAS devices contained in expander devices.
SAS domain bridged to ATA domains

SAS domain

Service delivery subsystem with expander devices

SSP

initiator port(s)

STP

initiator port(s)

SMP

initiator port(s)

SSP target port(s)

STP target port(s)

SMP target port(s)

STP/SATA bridge(s)

ATA domain

ATa device port

ATA domain

(informative)
SAS devices in multiple SAS domains

SAS initiator device
- Initiator port

SAS target device
- Target port

Expander devices

SAS domain
Edge expander device set

- Set of edge expander devices
- 128 SAS addresses per set
- Typically bounded by a subtractive port (to a fanout expander device, or to another edge expander device set)
- Edge expander devices uses table routing and direct routing “downstream” and subtractive routing “upstream”
- Wide links between expanders are allowed
- No loops
Expander topologies

- Maximum of one fanout expander device in a SAS domain
- If no fanout expander, maximum of two edge expander device sets (attached via subtractive decode ports)
- End devices may be attached at any level
  - Directly to fanout expander device
  - Any level edge expander device
- Wide links possible between any two devices
- No loops
- No multiple paths
One fanout expander device

Any of the physical links could be wide

Maximum of 128 SAS addresses per edge expander device set

128 edge expander device sets
Two edge expander device sets

The root edge expander device in each edge expander device set uses the subtractive routing method on the expander phy attached to its peer.

Any of the physical links could be wide.

Upstream phys use the subtractive routing method; downstream phys use table routing method or direct routing methods.

Two edge expander device sets
Expander routing attributes and methods

- Each expander phy has an expander routing attribute
- The attribute determines the routing methods the expander uses with each phy

<table>
<thead>
<tr>
<th>Routing attribute</th>
<th>When attached to</th>
<th>Routing method used</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct</td>
<td>End device</td>
<td>Direct</td>
</tr>
<tr>
<td>Table</td>
<td>Expander device</td>
<td>Table + Direct</td>
</tr>
<tr>
<td></td>
<td>End device</td>
<td>Direct</td>
</tr>
<tr>
<td>Subtractive</td>
<td>Expander device</td>
<td>Subtractive (+ Direct)</td>
</tr>
<tr>
<td></td>
<td>End device</td>
<td>Direct</td>
</tr>
</tbody>
</table>

- Direct = route requests to the attached SAS port through this phy
- Table = route requests that match in routing table through this phy
- Subtractive = route unresolved requests through this phy
**Expander route table contents**

- **Expander route table**
  - | 0 | 1 | 2 | ... | N |
  - |---|---|---|-----|----|
  - | 0 |   |   |     |    |
  - | 1 |   |   |     |    |
  - | 2 |   |   |     |    |
  - | ... | |   |     |    |
  - | M | |   |     |    |

- **Phy identifier**
  - A phy identifier for each expander phy of the expander device.
  - \( N \leq \text{number of phys - 1} \)

- **Expander route entry**
  - Includes:
    - Routed SAS address
    - Enable/disable bit

- **Expander route index**
  - An expander route index for each expander route entry.
  - \( M = \text{expander route indexes - 1} \)
Connections

• Connection = temporary association between an initiator phy and a target phy
  - Source phy transmits an OPEN address frame
    • Contains a destination SAS address
  - Expanders route it to a matching destination phy
  - Destination phy replies with an OPEN_ACCEPT primitive
  - Connection is established
  - Both sides exchange CLOSE primitives to close
• Connections are addressed to ports but established phy-to-phy
• N-wide ports may establish N connections at a time (to up to N other ports)
• Wide ports may establish multiple connections to other wide ports simultaneously
The green request is blocked until the red connection finishes.
Connection rules

• Connections are addressed to SAS ports but are established from phy to phy

• Wide ports may establish multiple connections at a time (to up to one per phy) to different destinations

• Wide ports may establish multiple connections to other wide ports simultaneously (wide initiator port to wide target port)
  - SAS disk drives will offer two narrow ports
  - Only HBAs and RAID controllers will offer wide ports
Rate matching (connection rate)

- Let's 1.5 Gbps and 3.0 Gbps SAS ports communicate
  - Especially 3.0 Gbps initiators and 1.5 Gbps SATA drives
- On faster physical links, insert ALIGN primitives to slow down the effective data rate to match the slowest physical link in the connection
Phy layer
8b10b coding converts 8-bit bytes into 10-bit data characters for transmission on the wire

- Clock recovery
- DC balance
- Special characters
- Error detection

Mapping done with two simple logic blocks
- 5b6b and 3b4b
- Full table in SAS and SATA

Invented by IBM in 1983

Used by Fibre Channel, Gigabit Ethernet, 1394b, and many other standards
Dwords

- **Byte** = 8 bits (xxh)
- **Character** = 10 bits (Dxx.y or Kxx.y)
- **Dword** = 4 characters (or 4 bytes, depending on context)
  - 40 bits on the wire
  - Usually represents 32 bits of data
  - A dword may flip disparity or leave it the same
  - A dword is either a **data dword**, a **primitive**, or an **invalid dword**

<table>
<thead>
<tr>
<th>first</th>
<th>second</th>
<th>third</th>
<th>fourth</th>
</tr>
</thead>
<tbody>
<tr>
<td>char</td>
<td>char</td>
<td>char</td>
<td>char</td>
</tr>
</tbody>
</table>
Primitives

- 1 control character and 3 data characters
  - First character is K28.5 (for SAS primitives), K28.3 (for SATA primitives), or K28.6 (special SATA error primitive)
    - K28.6 primitive serves as an invalid dword for SATA
  - Last three characters are data characters
  - Endianness does not matter
    - both SAS and SATA primitives always have the control character first on the wire

<table>
<thead>
<tr>
<th>first</th>
<th>second</th>
<th>third</th>
<th>fourth</th>
</tr>
</thead>
<tbody>
<tr>
<td>K28.5</td>
<td>Dxx.y</td>
<td>Dxx.y</td>
<td>Dxx.y</td>
</tr>
<tr>
<td>K28.3</td>
<td>Dxx.y</td>
<td>Dxx.y</td>
<td>Dxx.y</td>
</tr>
<tr>
<td>K28.6</td>
<td>Dxx.y</td>
<td>Dxx.y</td>
<td>Dxx.y</td>
</tr>
</tbody>
</table>
OOB signals

- OOB signal is a pattern of idle times and burst times
  - **Idle time** (and negation time)
    - Differential 0 V (Positive signal = negative signal)
    - No transitions (DC idle)
  - **Burst time**
    - Transmitted as a burst of ALIGN (0) primitives
    - Received as presence of edges (whether they are valid ALIGNs is irrelevant)
- Designed to be detectable by analog squelch detection logic
- Length of idle time distinguishes between OOB signals
  - COMINIT, COMWAKE, and COMSAS
**SAS OOB sequence**

- Transmit and receive COMINIT
- Transmit COMSAS
  - If COMSAS received, physical link is SAS to SAS
  - If COMSAS is not received, physical link is SAS to SATA
- There are a few hot-plug situations where COMINIT leads directly to COMSAS; this is allowed
SAS to SATA OOB sequence

- If COMSAS is not received, physical link is SAS to SATA
- SAS OOB sequence morphs into the SATA OOB sequence
- Since the SAS device drives COMWAKE, it looks like a SATA host
- Only supports SATA device not SATA host
SAS speed negotiation sequence

- Slow-to-fast
- Both phys run same set of speed negotiation windows
- 1.5, then 3.0, then 6.0 (if needed), etc. until they find:
  - a supported rate; then
  - a (faster) non-supported rate
- Last window returns to the highest supported rate detected
Link layer
Clock skew management

- Slow down the transmitter by including dwords that can be deleted
  - ALIGN (and NOTIFY for SAS) primitives stuck in
  - Throttles the transmitter to be slower than the slowest receiver
    - SATA (and SAS STP) requires 2 per 256 dwords (0.7812%)
      - Must always appear in pairs
    - SAS requires 1 per 2048 dwords (0.0488%)
  - Receiver throws away incoming ALIGNs (and NOTIFYs for SAS)

![Diagram showing transmitter and receiver with dwords]

| ... | dword e | dword f | ALIGN | dword g | dword h | dword i |
Connections

- Established between a SAS initiator phy and a SAS target phy
- Three types of connections
  - SSP (Serial SCSI protocol)
  - SMP (Serial management protocol)
  - STP (Serial ATA Tunneling protocol)
- Basic sequence
  - 1. Transmit OPEN address frame
  - 2. Receive AIP (arbitration in progress) primitives if expanders are involved
  - 3. Receive final result
    - OPEN_ACCEPT primitive – connection established
    - OPEN_REJECT primitive – rejected, go back to idle
- SATA has no concept of connections
**Connection – basic sequence**

- **OPEN**
  - address
  - frame

- **SOAF**
  - 8 data dwords

- **EO AF**
  - idle dwords

- **idle dwords**

- **AIP**

- **idle dwords**

- **AIP**

- **idle dwords**

- **AIP**

- **idle dwords**

- **AIP**

- **idle dwords**

- **open_address_frame**

- **OPEN_ACCEPT**

- **or OPEN_REJECT**
Arbitration fairness and deadlocks

• SAS implements a Least Recently Used (LRU) arbitration fairness scheme
  - Each OPEN address frame includes the age of a request
  - In case of ties, SAS addresses are used to pick the highest priority

• SAS implements deadlock detection and recovery using Partial Pathway timers
  - When expanders detect a hang, one request is backed off to break the likely deadlock
  - SAS addresses are eventually used to decide which request is the lowest priority and gets backed off
Closing a connection

- Exchange CLOSE primitive sequences end-to-end to close a connection
- SSP and SMP require traffic be stopped first
  - CLOSE ignored by SAS phy state machines if not at the correct time
- STP connection should be idle before closing
  - Not enforced by SL state machine
- Expanders always honor CLOSE regardless of connection state

Example 1: CLOSEs sent one at a time

Example 2: CLOSEs sent simultaneously

CLOSE primitive sequence

CLOSE primitive sequence

CLOSE primitive sequence

CLOSE primitive sequence
SSP overview

- Inside an SSP connection...
- Phys exchange **SSP frames**
  - SSP frame = SOF primitive, data dwords, EOF primitive
  - Each frame results in an ACK or NAK primitive
- Credit-based flow control
  - Permission to send a frame must be granted with RRDY primitives
- Full duplex
  - SSP frames can be sent in both directions simultaneously
  - Independent credit for each direction
- ACK, NAK, and RRDY primitives may be interjected among frame dwords
  - (so can ALIGNs and NOTIFYs)
SSP interlocked frames example

- Showing frames only in one direction

Send frame

Interlocked frame

SOF

idle dwords

Wait for ACK or NAK

idle dwords

...
- Showing frames in one direction

Non-interlocked frame
EOF, SOF, non-interlocked frame
EOF, SOF, non-interlocked frame
EOF, SOF, non-interlocked frame

4
3
2
1

idle dwords
ACK or NAK, idle dwords
ACK or NAK, idle dwords
ACK or NAK, idle dwords
ACK or NAK, idle dwords
SMP overview

• Only an initiator can open an SMP connection
  - Target not allowed to open an initiator

• Inside an SMP connection...

• Two SMP frames are transferred
  - SMP frame = SOF primitive, data dwords, EOF primitive
  - 1. Initiator transmits one SMP_REQUEST frame to target
  - 2. Target transmits one SMP_RESPONSE frame to initiator
SMP example

SMP initiator phy

Send request frame
SOF
SMP_REQUEST frame
EOF
idle dwords
...
CLO SE

OPEN_ACCEPT

SMP target phy

Send response frame
SOF
SMP_RESPONSE frame
EOF
idle dwords (optional)
CLO SE
STP and SATA overview

• In SATA, SATA host and SATA device just communicate directly – no connections

• In SAS, once an STP connection is open, STP initiator and STP target communicate as if they were SATA host and SATA device directly attached on a physical link
  - Extra latency introduced as dwords flow through expanders

• Half duplex
  - SATA never transmit frames in both directions at one time
  - Usually the frame goes one way and R_IP primitives go the other way

• Note: SAS prefixes the SATA primitive names with “SATA_”
  - SATA_SOF (used in STP) is different than SOF (used in SSP)
  - In this presentation, prefix not always used
**SATA basic frame transmission**

- **Idle**
- **Request permission**
- **Send frame**
- **Wait for termination**
- **Idle**

**SATA_SYNC**

- **SATA_X_RDY**
- **SATA_SOF**
- **Data dwords**
- **SATA_EOF**
- **SATA_WTRM**

**SATA_R_RDY**

- **SATA_R_OK** or **SATA_R_ERR**

**SATA_R_IP**

- **SATA_R_OK** or **SATA_R_ERR**

**SATA_SYNC**

- **Receipt in progress**
- **Acknowledge frame receipt**

**SATA_SYNC**

- **Idle**
Transport layer
### SSP frames

- **One common SSP frame format**
  - **Frame header**: 24 bytes
  - **Information Unit**: 0 to 1024 bytes
  - **Fill bytes**: 0 to 2 bytes
  - **CRC**: 4 bytes
- Frame format is based on Fibre Channel

<table>
<thead>
<tr>
<th>Byte</th>
<th>Field(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Frame Type</td>
</tr>
<tr>
<td>1 to 3</td>
<td>Hashed Destination SAS address</td>
</tr>
<tr>
<td>4</td>
<td>Reserved</td>
</tr>
<tr>
<td>5 to 7</td>
<td>Hashed Source SAS address</td>
</tr>
<tr>
<td>8 to 9</td>
<td>Reserved</td>
</tr>
<tr>
<td>10</td>
<td>Reserved</td>
</tr>
<tr>
<td>11</td>
<td>Reserved</td>
</tr>
<tr>
<td>12 to 15</td>
<td>Reserved</td>
</tr>
<tr>
<td>16 to 17</td>
<td>Tag</td>
</tr>
<tr>
<td>18 to 19</td>
<td>Target Port Transfer Tag</td>
</tr>
<tr>
<td>20 to 23</td>
<td>Data Offset</td>
</tr>
<tr>
<td>24 to m</td>
<td>Information Unit</td>
</tr>
<tr>
<td>m to (n-3)</td>
<td>Fill bytes, if needed</td>
</tr>
<tr>
<td>(n-3) to n</td>
<td>CRC</td>
</tr>
</tbody>
</table>
**SSP frame types**

- Information Units are nearly identical to Fibre Channel Protocol (FCP)
  - Separate TASK frame is new to SAS

<table>
<thead>
<tr>
<th>Command</th>
<th>Information Unit field size</th>
<th>Direction</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>COMMAND</td>
<td>28 to 284</td>
<td>I to T</td>
<td>Send a command</td>
</tr>
<tr>
<td>TASK</td>
<td>28</td>
<td>I to T</td>
<td>Send a task management function</td>
</tr>
<tr>
<td>XFER_RDY</td>
<td>12</td>
<td>T to I</td>
<td>Request write data</td>
</tr>
<tr>
<td>DATA</td>
<td>1 to 1024</td>
<td>I to T or T to I</td>
<td>Write data (I to T) or read data (T to I)</td>
</tr>
<tr>
<td>RESPONSE SE</td>
<td>24 to 1024</td>
<td>T to I</td>
<td>Send SCSI status (for commands) or task management response (for task management functions)</td>
</tr>
</tbody>
</table>
SSP task management frame sequence

SSP initiator port

Send Task Management Request ()

Received Task Management Function Executed ()

time

SSP target port

Task Management Request Received ()

Task Management Function Executed ()

time

TASK frame

RESPONSE frame
SSP non-data command sequence

SSP initiator port

Send SCSI Command ()

Command Complete Received ()

time

SSP target port

COMMAND frame

SCSI Command Received ()

RESPONSE frame

Send Command Complete ()

time
SSP write command sequence

SSP initiator port

Send SCSI Command ()

SSP initiator port replies to XFER_RDY with one or more DATA IUs

Command Complete Received ()

data

time

SSP target port

SCSI Command Received ()

Receive Data Out ()

Send Command Complete ()

SSP target port sends XFER_RDY frames and receives DATA frames until all write data has been transferred

Data Out Received ()

data

time
SSP read command sequence

SSP initiator port

Send SCSI Command ()

Command Complete Received ()

time

COMMAND frame

DATA frame

RESPONSE frame

time

SSP target port

SCSI Command Received ()

Send Data In ()

Data In Delivered ()

SSP target port sends DATA frames until all read data has been transferred

Send Command Complete ()
### SSP bidirectional command sequence

<table>
<thead>
<tr>
<th>SSP initiator port</th>
<th>SSP target port</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="commandframe.png" alt="COMMAND frame" /></td>
<td><img src="commandframe.png" alt="COMMAND frame" /></td>
</tr>
<tr>
<td>Send SCSI Command ()</td>
<td>SCS Command Received ()</td>
</tr>
<tr>
<td><img src="xferframed.png" alt="XFER_RDY frame" /></td>
<td>Receive Data Out ()</td>
</tr>
<tr>
<td>SSP initiator port replies to XFER_RDY with one or more DATA IUs</td>
<td>SSP target port sends XFER_RDY frames until all write data has been transferred</td>
</tr>
<tr>
<td><img src="dataframe.png" alt="DATA frame" /></td>
<td>Data Out Received ()</td>
</tr>
<tr>
<td>Send Data In ()</td>
<td>SSP target port chooses how to interleave write data and read data</td>
</tr>
<tr>
<td>Data In Delivered ()</td>
<td></td>
</tr>
<tr>
<td><img src="responseframe.png" alt="RESPONSE frame" /></td>
<td>Send Command Complete ()</td>
</tr>
<tr>
<td>Command Complete Received ()</td>
<td>time</td>
</tr>
<tr>
<td>time</td>
<td></td>
</tr>
</tbody>
</table>
Wrap Up
Additional tutorials

- General overview
- More detailed 7-part tutorial:
  - #1: General introduction and architectural overview
  - #2: Physical and phy layers
  - #3: Link layer
    - Part 1) Primitives, address frames, connections
    - Part 2) Arbitration fairness, deadlocks and livelocks, rate matching, SSP, STP, and SMP frame transmission
  - #4: Upper layers
    - Part 1) SCSI application and transport layers
    - Part 2) ATA application and transport layers
    - Part 3) Management application and transport layers, plus port layer
Key SCSI standards

- Working drafts of SCSI standards are available on http://www.t10.org
- Published through http://www.incits.org
  - Serial Attached SCSI
  - SCSI Architecture Model - 3
  - SCSI Primary Commands - 3
  - SCSI Block Commands - 2
  - SCSI Stream Commands - 2
  - SCSI Enclosure Services - 2
- SAS connector specifications are available on http://www.sffcommittee.org
  - SFF 8482 (internal)
  - SFF 8470 (external)
  - SFF 8223, 8224, 8225 (form factors)
Key ATA standards

- Working drafts of ATA standards are available on http://www.t13.org
  - Serial ATA 1.0a (private WG specification)
  - ATA/ATAPI-7 Volume 1 (architecture and commands)
  - ATA/ATAPI-7 Volume 3 (Serial ATA standard)
- Serial ATA II specifications are available on http://www.t10.org and http://www.serialata.org
  - Serial ATA II: Extensions to Serial ATA 1.0
  - Serial ATA II: Port Multiplier
  - Serial ATA II: Port Selector
  - Serial ATA II: Cables and Connectors Volume 1
For more information

- International Committee for Information Technology Standards
  - [http://www.incits.org](http://www.incits.org)
- T10 (SCSI standards)
  - [http://www.t10.org](http://www.t10.org)
  - Latest SAS working draft
  - T10 reflector for developers
- T13 (ATA standards)
  - [http://www.t13.org](http://www.t13.org)
  - T13 reflector for developers
- T11 (Fibre Channel standards)
  - [http://www.t11.org](http://www.t11.org)
- SFF (connectors)
  - [http://www.sffcommittee.org](http://www.sffcommittee.org)
- SCSI Trade Association
  - [http://www.scsita.org](http://www.scsita.org)
- Serial ATA Working Group
  - [http://www.serialata.org](http://www.serialata.org)
- SNIA (Storage Networking Industry Association)
  - [http://www.snia.org](http://www.snia.org)
- Industry news
  - [http://www.infostor.com](http://www.infostor.com)
  - [http://www.byteandswitch.com](http://www.byteandswitch.com)
  - [http://www.wwpi.com](http://www.wwpi.com)
  - [http://searchstorage.com](http://searchstorage.com)
- Training
  - [http://www.knowledgetek.com](http://www.knowledgetek.com)