SCSI and FC standards update

Frederick Knight
NetApp Inc
INCITS standards process

1. Proposals (at each TC - T10, T11, T13)
2. Working Draft (many revs)
3. TC Letter Ballot (+ comment resolution)
4. INCITS Public Review (+ comment resolution)
5. ANSI Standard published
6. ISO process
   - NWIP / CD / review + resolve / FCD / publish
T10 Document Status

- SCSI Primary Commands (SPC-4) (July 2012)
- SCSI / ATA Translation (SAT-3) (Oct 2012)
- SCSI Block Commands (SBC-3) (Nov 2012)
- SCSI Express (SOP / PQI) (Sept 2012)
- Serial Attach SCSI (SAS-3) (Mar 2013)

Currently resolving > 18,000 comments
What’s Driving Change

- HDD ➔ SSD
  - Bandwidth
    - 145 – 200 Mb/s ➔ 3.2 – 14.4 GB/s (*)
  - IOPS
    - 150 – 400 IOPS ➔ 785 – 3.5M IOPS (*)
  - Latency
    - 5 – 13 ms ➔ 5 – 50 usec

- Flash media life cycle issues

(*) = PCIe 3.0 x 16
New SCSI (T10) Features

- Atomic commands
  - All or none
  - Easy for flash to implement
    - Flash = not overwrite in place
  - Interactions with traditional R / W commands
- Scatter / Gather commands
  - Multiple LBA and length pairs
    - Write data buffer includes the pairs and the data
    - Read requires bi-di command (pairs out; data in)
  - Databases, Filesystems, etc
New SCSI (T10) Features

- Example:

- Write LBAs 22 to 32 (error on LBA 26) - what happens?
  - Indeterminate data (old vs. new) in all LBAs 22 to 32
  - For Atomic operation, all of 22 to 32 are written
    - If error on LBA 26 occurs, then NONE of LBA 22 to 32 are written
New SCSI (T10) Features

- **SCATTER WRITE command**
- **CDB**
  - # of range entries
  - # of LBAs to write
- **DATA-OUT**
  - LBA range lists; data for those LBAs
New SCSI (T10) Features

- **GATHERED READ command**
- **Bi-Di command**
- **CDB**
  - # of range entries
  - # of LBAs to read
- **DATA-OUT**
  - LBA range lists
- **DATA-IN**
  - data for those LBAs
New SCSI (T10) Features

- Simplified feature Discovery
  - Defined feature sets with single command discovery
  - Base feature set
  - Drive management feature set
  - Basic provisioning feature set
  - Others?
  - Obsolete old (unused?) features
New SCSI (T10) Features

- Base Feature set
  - INQUIRY, MODE SENSE / SELECT (10), REPORT LUNS, REPORT SUPPORTED OPCODES / TMF, FORMAT UNIT, REQUEST SENSE, START STOP UNIT, READ / WRITE / VERIFY / WRITE SAME (16), READ CAPACITY (16)
  - Eliminate duplicates (no READ / WRITE 6/10/12)
New SCSI (T10) Features

- Drive Management Feature set
  - LOG SENSE / SELECT (10), FORMAT UNIT, READ / WRITE BUFFER, SEND DIAGNOSTIC, REASSIGN BLOCKS, READ DEFECT DATA (12), SANITIZE, WRITE LONG (16) (w/WR_UNCOR bit)
  - Mainly for test purposes
New SCSI (T10) Features

- Basic Provisioning Feature set
  - GET LBA STATUS, READ CAPACITY (16), UNMAP, WRITE SAME (16) (w/UNMAP bit)
- Additional advanced features?
  - ALUA (HA features)
  - Scatter / Gather
  - Atomic
  - DIF (PI)
  - Copy Offload (EXTENDED COPY)
New SCSI (T10) Features

- Security erasure (SANITIZE command)
  - Overwrite
  - Crypto scramble
  - Block erase
  - WHOLE logical unit only (for now)
- New VERIFY SAME option (in VERIFY command)
- TLC (Time Limited Commands)
  - Command completes or terminates
- Additional status info (aka: GOOD with Sense)
New SCSI (T10) Features

- Conglomerate Logical units (Virtual Volumes)
  - Large scale configurations
    - Billions and Billions of logical units
    - Inventory discovery (REPORT LUNS)
    - Unit Attention Coalescing
  - Administrative & Subsidiary Logical units
- LU movement
  - Binding / unbinding of data to LUN
New SCSI (T10) Features

Conglomerates

Data Set

LUN

Administrative

Relationship

LUN

Subsidiary

LUN

Subsidiary

LUN

Subsidiary

Bind
New SCSI (T10) Features

- Device crash dump
  - In-band extract of debug info from a drive
  - Drive generated or initiator generated
  - Uses READ BUFFER command
  - Returns VS buffer format (with standard header)
  - Vendor ID indicates company that can decode the VS information in the buffer
- Common SCSI & ATA methods
  - Direct attach and Backend devices
- Banding (SMR)
SMR (Shingled Magnetic Recording)
SMR (Shingled Magnetic Recording)
New SCSI (T10) Features

- Banding (SMR)
  - Hints from storage to host
  - Legacy use patterns = slow
  - Intelligent use patterns = fast
  - Demands from storage to host

- Hints
  - From host to storage
SCSI Transport updates

- SAS 12GB now complete (SAS-3)
  - Performance enhancements (long lived connections)
  - Work started on 24GB (SAS-4)

- SCSI Express
  - SCSI over PCIe (SoP / PQI…) (in LB resolution)
  - Direct connect storage (flash)
iSCSI Transport updates

- iSCSI being updated for new features
  - iSER and MIB updated
  - ipSec update underway
  - Consolidated RFC (all of iSCSI in one place)
  - New features RFC (SAM-2 to SAM-5 updates)
    - Negotiated (iSCSIProtocolLevel)
    - Priority field
    - additional response information
  - Fully backward compatible
The iSCSI Command PDU

<table>
<thead>
<tr>
<th>Byte/</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>/</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0 1 2 3 4 5 6 7</td>
<td>0 1 2 3 4 5 6 7</td>
<td>0 1 2 3 4 5 6 7</td>
<td>0 1 2 3 4 5 6 7</td>
</tr>
</tbody>
</table>
| +---------------+---------------+---------------+---------------+
| 0 |.|I| 0x01 |F|R|W|. .|ATTR | PRI | Reserved |
| +---------------+---------------+---------------+---------------+
| 4 |TotalAHSLength | DataSegmentLength |
| +-----------------------------------------------+|
| 8 | Logical Unit Number (LUN) |
| +-----------------------------------------------+|

- PRI field (previously reserved space)
### The iSCSI Response PDU

<table>
<thead>
<tr>
<th>Byte/</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>/</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 1 2 3 4 5 6 7</td>
<td>0 1 2 3 4 5 6 7</td>
<td>0 1 2 3 4 5 6 7</td>
<td>0 1 2 3 4 5 6 7</td>
<td></td>
</tr>
</tbody>
</table>

+---------------+---------------+---------------+---------------+
| 0 | 0x21 | 1 | Response | Status |
+---------------+---------------+---------------+---------------+

+---------------+---------------+---------------+---------------+
| 4 | TotalAHSLength | DataSegmentLength |
+---------------+---------------+---------------+---------------+

+---------------+---------------+---------------+---------------+
| 8 | Status Qualifier | Reserved |
+---------------+---------------+---------------+---------------+

+---------------+---------------+---------------+---------------+
| 12 | Reserved |
+---------------+---------------+---------------+---------------+

#### Status Qualifier field (previously reserved space)

#### Explicit – Sense data on non-CHECK CONDITION (aka: GOOD with sense)
iSCSI Transport updates

- Additional Task Management Functions
  - QUERY TASK
  - QUERY TASK SET
  - I_T NEXUS RESET
  - QUERY ASYNCHRONOUS EVENT

- New TMF response
  - FUNCTION SUCCEEDED
Fibre Channel (T11) updates

- Peer Zoning (aka Target Driven Zoning – TDZ)
  - Targets inform SAN about zoning

- FC Speed increasing
  - 32GFC just about done
  - 128GFCp underway (4x32GFC)
Fibre Channel (T11) updates

- 128GFC Port
- 128GFC FEC
- 32GFC FEC
- 32GFC FEC
- 32GFC FEC
- 32GFC FEC

Active

- I28GFC QSFP28

Inactive

- ASIC
- 32GFC FEC
- 32GFC SFP+
Fibre Channel (T11) updates

All 4 32GFC FEC Inactive
Fibre Channel (T11) updates

- FC-BB-6 - FCoE VN2VN (point to point)
  - Ethernet switch only
  - Ethernet FC fabrics (no FC switching elements)
Fibre Channel (T11) updates

- FC-HBA-2 completed INCITS public review
- FC-SW6
  - Distributed switch
  - Increased domain count
  - Increased port count
- FC-LS-3 and FC-FS-4
  - New TLV based management operations
Fibre Channel (T11) updates

- New Efforts
  - New EEFC (Energy Efficient FC)
    - Copper done (IEEE method) / Optical ongoing
  - Energy Star for switches