Why increase addressability?

- Bus bandwidth is increasing at CAGR of 50% while connectivity is at a standstill.
- Increases in sustainable HDA transactions per second have lagged behind growth in other areas of HDA performance.
  - I/Os per second CAGR for random workloads ~ 14%.
  - Compared to CAGRs for:
    - Areal density: ~ 60%
    - HDA peak data rate: CAGR ~ 25%
- Because of improvements in the protocol and electrical layer, there is more bus headroom for processing transaction-intensive workloads
  - Example: TPC-type workloads (2K random reads, RW Ratio =2:1) use ~ 1% of the bus bandwidth per HDA.
- ∴ For HDA-limited workloads, more devices per bus = higher throughput
Extended Addressing Proposal

- **Goals:**
  - Increase connectivity of Wide SCSI LVD by a factor of 4 (up to 64 devices)
  - Increase I/Os per second by exchanging latent bus bandwidth for increased device count.
  - Preserve compatibility with legacy SCSI
    - No change to the SCSI LVD Wide electrical layer
      - Changes are in the Arbitration and Selection protocols
    - Extended devices are fully compatible with legacy arbitration and selection protocols
      - Device that supports extended addressing can operate in legacy SCSI mode.
    - Legacy devices can operate on extended busses
      - Restriction: Legacy devices can't use QaS on an extended bus
  - No change to the SCSI LVD Wide electrical layer
  - Changes are in the Arbitration and Selection protocols
  - Extended devices are fully compatible with legacy arbitration and selection protocols
  - Device that supports extended addressing can operate in legacy SCSI mode.
  - Legacy devices can operate on extended busses
    - Restriction: Legacy devices can't use QaS on an extended bus

- **Assumptions:**
  - Design center is LVD SCSI Wide
  - Use of bus expanders allows more physical devices to be attached
  - Fairly inexpensive
  - Device load can be distributed across several segments.

Protocol Overview

- **Extended device address**
  - 16-bit format, two bits per device
    - Extended Group ID (GID) in bits 7--0
    - Group IDs 15--8 reserved for legacy devices
      - Legacy device addresses have no MID component.
    - Group member ID (MID) in bits 15--8
  - Addressability is 64 extended devices.
  - GID/MID combination is unique for each device.
  - Device automatically operates in extended mode if extended address is assigned
### Extended Address Format - Group I/Ds

**Table 1 -- Group I/D Arbitration Priority**

<table>
<thead>
<tr>
<th>Group ID</th>
<th>DB 15</th>
<th>DB 8</th>
<th>DB 7</th>
<th>DB 0</th>
<th>Priority</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>8</td>
</tr>
</tbody>
</table>

**Table 2 -- Member I/D Arbitration Priority**

<table>
<thead>
<tr>
<th>Member ID</th>
<th>DB 15</th>
<th>DB 8</th>
<th>DB 7</th>
<th>Unused</th>
<th>DB 0</th>
<th>Priority</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>3</td>
</tr>
<tr>
<td>12</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>4</td>
</tr>
<tr>
<td>11</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>5</td>
</tr>
<tr>
<td>10</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>6</td>
</tr>
<tr>
<td>9</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>8</td>
</tr>
</tbody>
</table>
Extended Arbitration

- Two round elimination
  - First round -- Group and legacy device arbitration
    - Identical to legacy arbitration cycle
    - Devices in the highest priority group advance to next round
    - Legacy devices that lose drop out
    - Legacy device that wins bypasses second round, proceeds directly to selection phase
  - Second round -- Group member arbitration
    - Device with highest priority MID wins
  - Estimated additional arbitration overhead for the second cycle
    - Added Arbitration time: +1.2 us
      - % Increased Arb overhead = (3600+1200)/3600 = 33%
    - QaS: +1 us
      - % Increased QaS overhead = (2000 + 1000)/2000 = 50%

Extended Selection

- No change in timing
- Approach:
  - Snoop arbitration phase to build selection mask
    - Snooping is already used for fairness
  - Selection Mask = ID of ARB Winner | Device ID
- Discriminating between legacy and extended selection
  - Three or four bits asserted during extended selection
  - Only two data bits asserted during legacy selection
Starvation Avoidance

◆ Each extended device implements two “fairness” registers
  ♦ Group
  ♦ Group member
◆ Mask registers with one bit set for each arbitrating group or group member ID whose priority is less than the device.
◆ On each arbitration cycle
  ♦ Each device updates its group fairness register
  ♦ Each device updates its group member fairness register from the winning group MIDs

Starvation Avoidance (cont.)

◆ A device may arbitrate when both its Group and Group Member fairness registers are 0.
◆ Legacy device fairness
  ♦ Group I/Ds in the range 8 – 15 are reserved for legacy devices.
  ♦ Legacy devices update their fairness registers with the group I/Ds of lower priority contending devices.
  ♦ Extended devices will defer to legacy devices.
  ♦ Legacy devices will defer to lower priority legacy devices.
Performance

◆ Scenario
  ◆ Transfer Parameters
    - Ultra-320
    - Random Reads (no cache hits)
    - Packetized, QAS
    - Disconnect/Reconnect every 16KB
  ◆ Drive Parameters (Year 2003 SWAG)
    - Drive Transfer Rate: 70MB/sec
    - Average seek time: 2.3ms
    - Average rotational delay: 1.35ms

Estimated HDA Capacity

Random I/O Bus Capacity

Ultra-320

<table>
<thead>
<tr>
<th>Transfer Size (kb)</th>
<th>Max Number of Drives</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td></td>
</tr>
<tr>
<td>2.0</td>
<td></td>
</tr>
<tr>
<td>4.0</td>
<td></td>
</tr>
<tr>
<td>8.0</td>
<td></td>
</tr>
<tr>
<td>16.0</td>
<td></td>
</tr>
<tr>
<td>32.0</td>
<td></td>
</tr>
<tr>
<td>64.0</td>
<td></td>
</tr>
<tr>
<td>128.0</td>
<td></td>
</tr>
<tr>
<td>256.0</td>
<td></td>
</tr>
</tbody>
</table>

Legacy Packetized
Extended Packetized
Estimated Effect on Bus Capacity

Change in Packetized Bus Capacity

Conclusions:

◆ When to use extended addressing:
  ◆ In configurations requiring a high device count
  ◆ When HDAs are connected to a heavily cached host or raid box
    – Residual drive traffic tends to miss the HDA cache, so the hit ratio is low.
  ◆ Transaction rate is HDA-limited.

◆ When to use legacy addressing
  ◆ SCSI bus connected to external RAID box
    – There is a large percentage of cache hits
    – Device count on the bus is less important than response time
  ◆ Device count is low
    – e.g., Desktop, entry-level servers
Next steps

- Specify how to implement with SCA-type connector
- Define bus configuration rules
- Explore bus expander issues
- Analyze electrical effects on bus
  - e.g., Wired-or effects on SELECT line.
- Add fairness details to the proposal
- Develop SES/Workbench model to simulate extended addressing

Backup Material
Bus Expander Considerations

- Allocating a group address to a single bus segment preserves arbitration properties.
- SELECT assertion at the completion of the first arbitration cycle originates from one side of the expander.
- Are there other issues?

A topology example

[Diagram of a topology example showing a backbone SCSI bus with branches leading to different groups of SCSI buses.]
Extended Addressing Timing
Diagrams

Arbitration Timing

- Bus Free + Bus Settle (1200 ns)
- Arbitration Delay (2400 ns)
- End first arbitration cycle
- End second arbitration cycle

- Legacy and low GIDs clear the bus
- Bus settle delay (400 ns)
- Bus clear + Bus settle (1200 ns)
- Start selection

- MIDs asserted by each device in the winning group
QAS Timing

- **Signals negated**
- **C/D Asserted by target**
- **BSY Asserted by selected device**
- **QAS Release delay + 2x bus settle delay**
- **QAS Release delay + 2x bus settle delay**
- **Bus settle delay (400 ns)**
- **QAS Release Delay (200 ns)**
- **End first arbitration cycle**
- **End second arbitration cycle**
- **Data Signals negated**
- **Sel Signals negated**

LUN Bridge as a Connectivity Solution

- **Cost and Complexity**
  - Bridge must emulate multi-lun target and initiator
- **Performance**
  - Device access requires at least two full arb cycles plus internal bridge delays
- **Other Issues**
  - How to handle multi-host configurations
    - Tagged queuing
    - Reserve/release, Persistent reserve, etc
  - How to handle select/reselect collisions