Parallel SCSI Extended Addressing Proposal
T10/99-249R1
September 14, 1999

Charles Monia

adaptec

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.**
- **$\therefore$ 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**
- **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.**

adaptec

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

adaptec

Extended Address Format - Group I/Ds

Table 1 -- Group I/D Arbitration Priority

Group I/D	DB 15	Legacy devices only							DB 8	DB 7	Legacy or extended devices						DB 0	Priority
7	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	1
6	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-	2
5	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	3
4	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	4
3	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-	5
2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-	6
1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	7
0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	8
15	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	9
14	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	10
13	--	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	11
12	--	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	12
11	--	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	13
10	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	14
9	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	15
8	--	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	16

Extended Address Format -- Member I/Ds

Table 2 – Member I/D Arbitration Priority

Member I/D	DB 15	Extended devices only							DB 8	DB 7	Unused						DB 0	Priority
15	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1
14	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2
13	--	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	3
12	--	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	4
11	--	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	5
10	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	6
9	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	7
8	--	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	8

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 loose 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

adaptec

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.

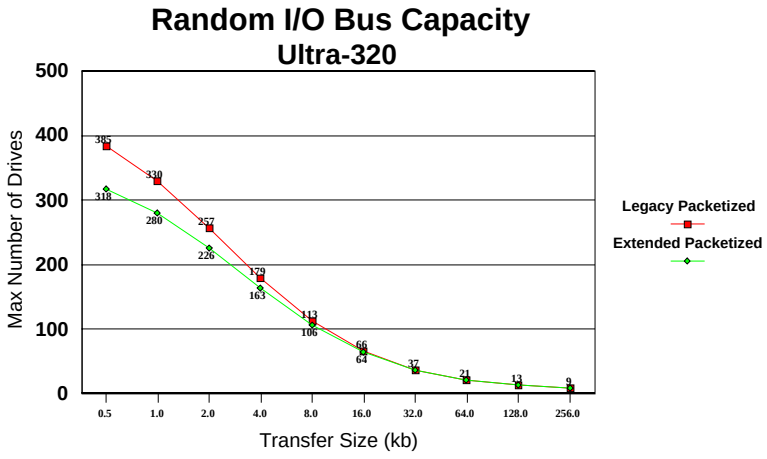
adaptec

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**

adaptec

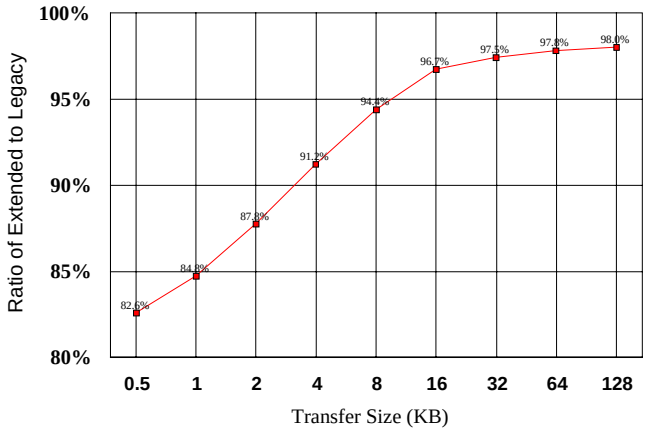
Estimated HDA Capacity



adaptec

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

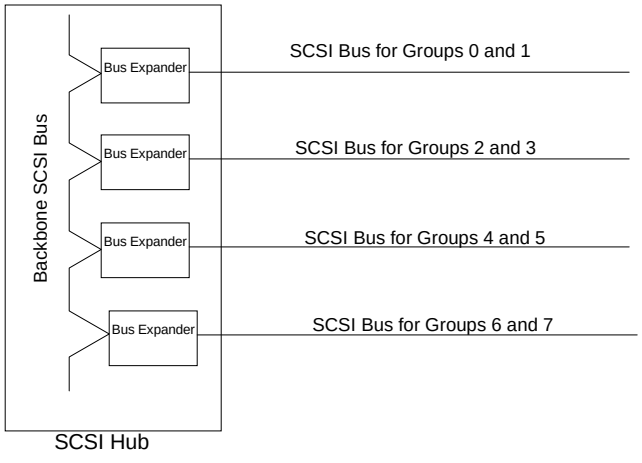
Backup Material

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?

adaptec

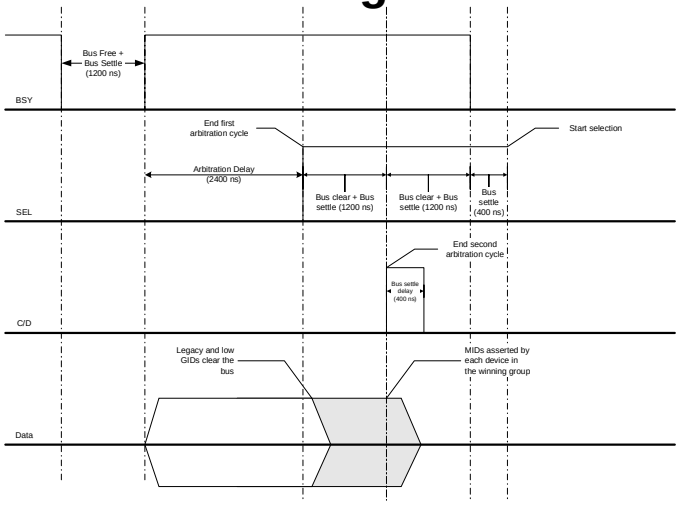
A topology example



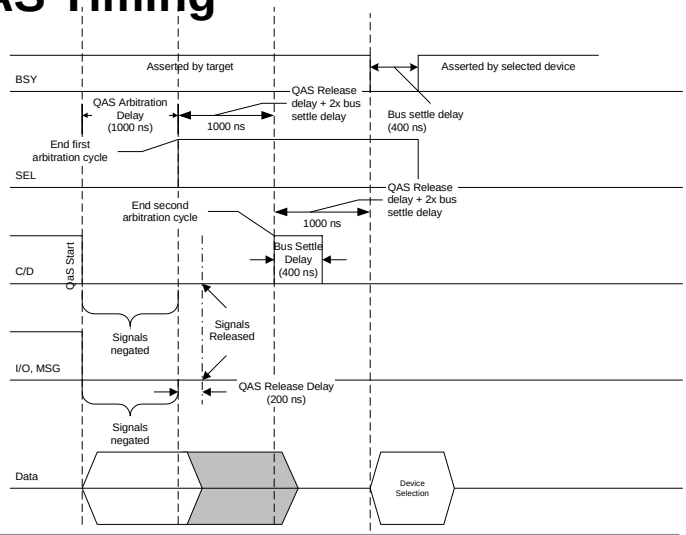
adaptec

Extended Addressing Timing Diagrams

Arbitration Timing



QAS Timing



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**