Parallel SCSI Extended Addressing Proposal

X3T10/99-249R0

September 14, 1999

Charles Monia

Why increase addressability?

- **In a random I/O, transaction-intensive environment there is headroom to spare**
  - **TPC-type workloads (2K random reads, RW Ratio =2:1) use ~ 1% of the bus bandwidth per HDA**
- **Increases in HDA transactions per second have lagged behind growth in other areas.**
  - – **I/Os per second CAGR for random workloads ~ 14%.**
  - – **Compared to CAGRs for:**
    - – **Parallel SCSI Bandwidth: ~ 50%**
    - – **Areal density:~ 60%**
    - – **HDA peak data rate: CAGR ~ 25%**

LUN Bridge 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

Extended Addressing Proposal

- Define extended arbitration cycle
  - ◆ Exchanges some latent bus bandwidth for increased transaction capacity.
- Preserves some compatibility with legacy devices.
  - ◆ Restriction: Legacy devices can’t use QaS
- Assumptions:
  - ◆ Design center is LVD SCSI Wide
  - ◆ Use of bus extenders allows more physical devices to be attached
    - ◆ Fairly inexpensive
    - ◆ Device load can be distributed across several segments.

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 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

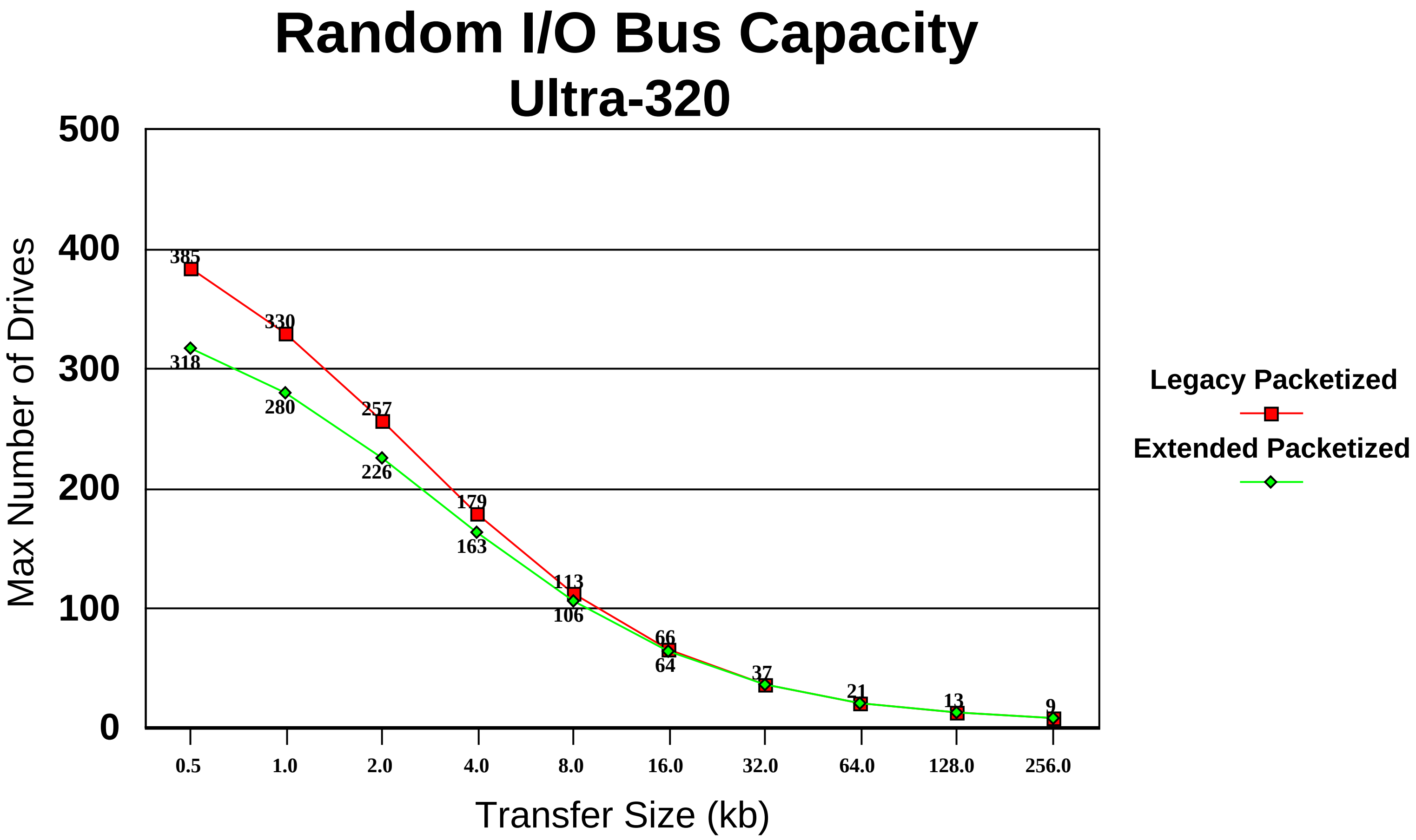
Starvation Avoidance (cont.)

- 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

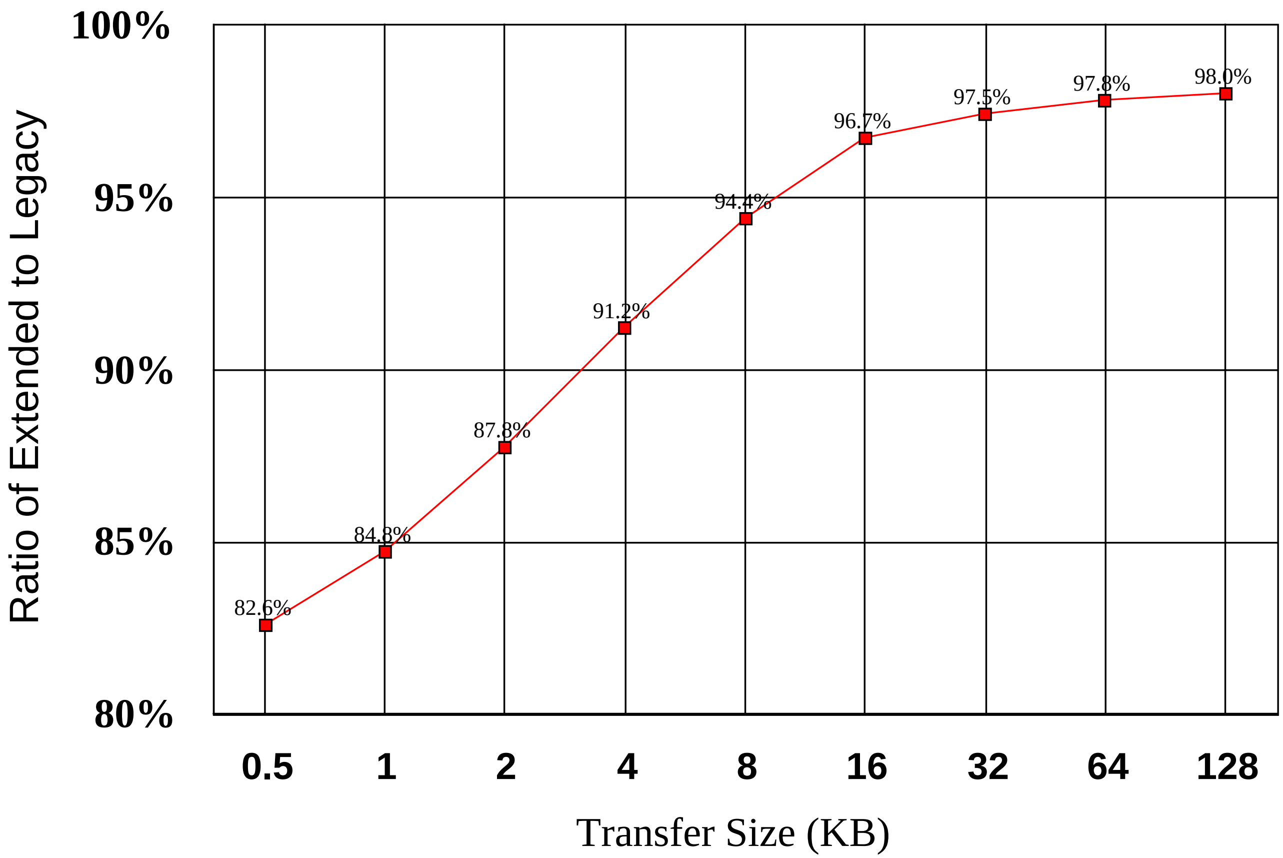
- **Scenario**
  - **Transfer Parameters**
    - **Ultra-320**
    - **Random Reads (no cache hits)**
    - **Packetized, QAS**
    - **Disconnect/Reconnect every 16KB**
  - **Drive Parameters (Year 2003 SWAG)**
    - **Drive Transfer Rate: 70MB/sec**
    - **Average seek time: 2.3ms**
    - **Average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

Estimated HDA Capacity



Estimated Effect on Bus Capacity

Change in Packetized Bus Capacity



Conclusions:

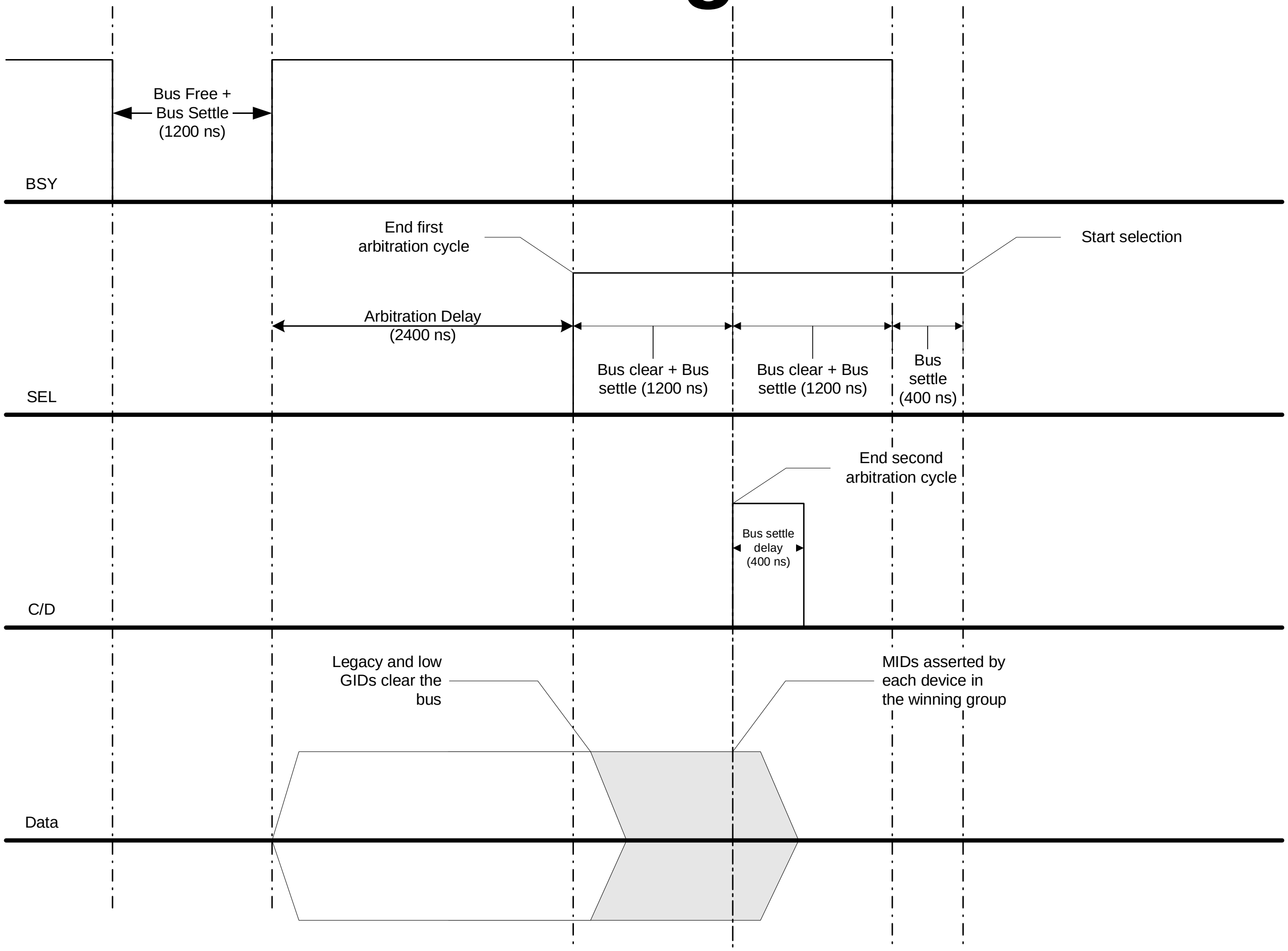
- **When to use extended addressing:**
  - **In configurations with a high HDA count**
    - **Fewer adapters required compared to legacy SCSI**
  - **When HDAs are connected to a heavily cached host or raid box**
    - **Residual drive traffic misses the HDA cache, so the hit ratio is low.**
  - **Transaction rate is HDA-limited.**
- **When to use legacy addressing**
  - **Raid boxes attached to the host via a front-side SCSI bus**
    - **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**

Work to be done

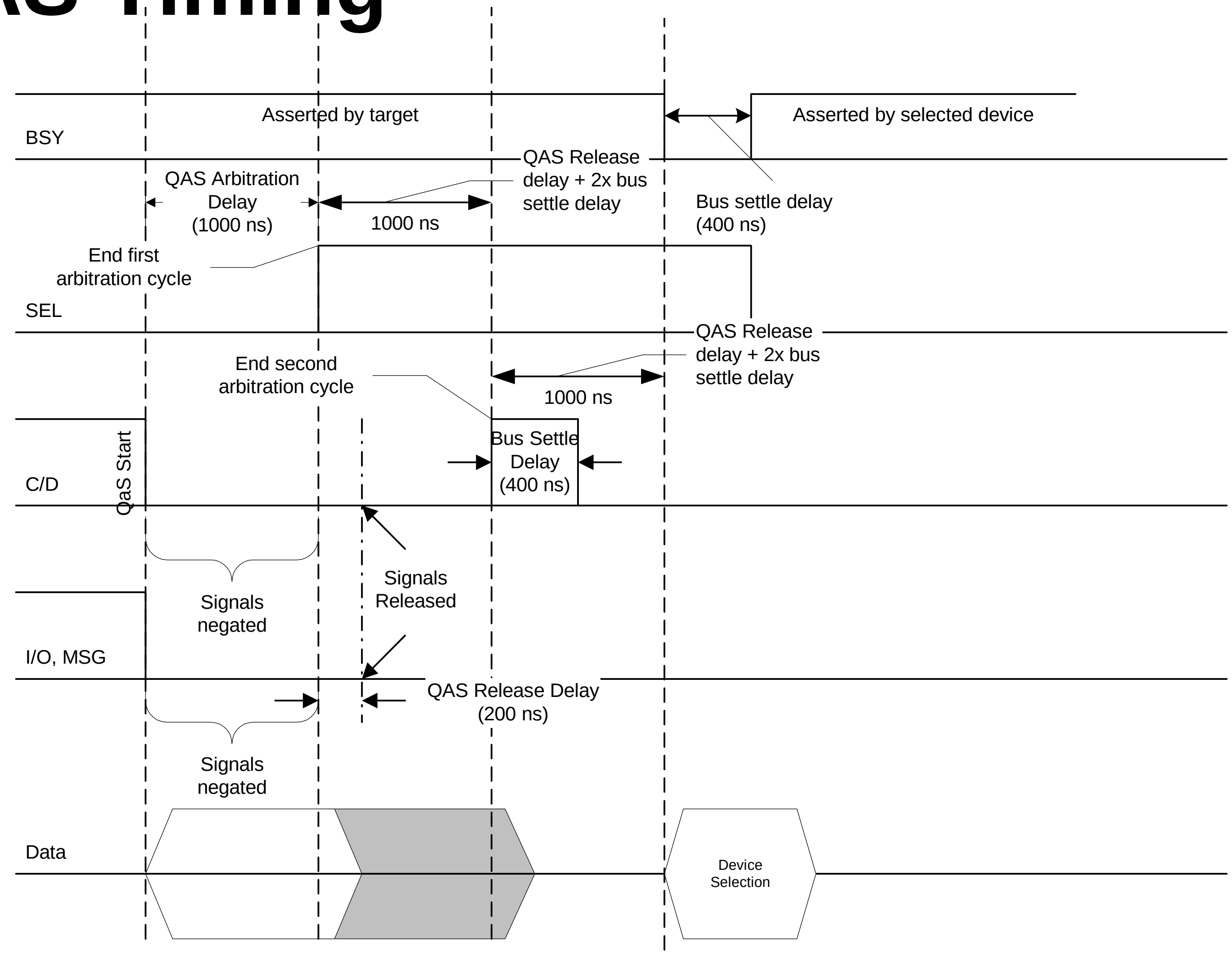
- Specify how to implement with SCA-type connector
- Assess impact on bus extenders
- Add bus configuration rules
- Analyze electrical effects on bus
    - ◆ **e.g., Wired-or effects on SELECT line.**
- Add fairness details to the proposal

Extended Addressing Timing Diagrams

Arbitration Timing



QAS Timing



Asserted by target
Asserted by selected device
BSY
QAS Release delay + 2x bus settle delay
QAS Arbitration Delay (1000 ns)
1000 ns
Bus settle delay (400 ns)
End first arbitration cycle
SEL
QAS Release delay + 2x bus settle delay
End second arbitration cycle
1000 ns
QaS Start
Bus Settle Delay (400 ns)
C/D
Signals Released
Signals negated
I/O, MSG
QAS Release Delay (200 ns)
Signals negated
Data
Device Selection