DMA Architecture for InfiniBand™

Rob Haydt
Program Manager
Base OS Development
Microsoft Corporation

Basic IB features

- Point to point links
- Switched fabric
- Multi-host capable
- Connection based with support for unconnected service
- Message oriented
- Supports remote direct memory access (RDMA)

Why adapt SBP-2 for IB?

- Designed around remote memory access model
- Login protocol supports multi-host
- Reconnect supports persistent connections
- Configuration ROM provides a flexible, understandable basis for enumerating resources and options
- Consistent model for scatter/gather list (page tables)
- No need to “re-invent the wheel”

IB DMA Architecture Issues

- Need a connection to do RDMA
- 64 Bit buffer addresses are qualified by a 32 bit Remote Key (Rkey)
- IB has no direct equivalent to the 1394 notification mechanism used to implement SBP-2 *status_FIFO*
- I/O controllers aren't required to be RDMA targets
- I/O controllers should implement message level End-to-End flow control

IB DMA Architecture Issues (cont.)

- Storage Issues
  - Additional SCSI encapsulations (>12 byte CDB's)
  - Need to distinguish between native storage controllers and FC bridges (and other back-end shared media)
  - Possible performance enhancements

Suggested Approach

- Initiator connects with target
- Use messages for Initiator access to Target registers
- Use messages for Target writes to *status_FIFO* locations
- Add Rkey to ORB pointer registers
- Define additional Config. ROM entries for additional functionality
- SCSI encapsulation enhancements

ORB Processing

- Target reads ORB's from initiator's memory
- Initiator sets the ORB pointer by sending a message containing the register address, the pointer and the Rkey for ORB
- Initiator writes the doorbell by sending a message that contains the register address
- All pointers in an ORB are accessed using the Rkey from ORB pointer
  - *next_ORB, password*
  - *data_descriptor, login_response*
  - *status_FIFO*

Connection & Login

- Initiator establishes connection
  - connect to target
  - read Configuration Rom
  - Multiple logins allowed on one connection
- Initiator login
  - allocate Login ORB, get Rkey
  - Write Management ORB pointer
- Target
  - Read Login ORB
  - Write Response
  - Send status_FIFO message

Nits

- Can 1394 fields be redefined for IB?
- IB supports a 64 bit memory address
  - Not really a problem if the address is only physical (48 PA bits will hold us for a while)
  - Would it be better just to expand all pointers to 96 (or even 128) bits?
  - Should page table entries be expanded to 96 (or even 128) bits?

SCSI additions

- >12 byte CDB's
- SCSI-3

Performance Enhancements

- Support a separate connection for data transfers
  - Isolate control traffic to a single channel
  - Allow a dedicated data channel for each LUN
- Prefetch beyond ORB end
  - Better network utilization
  - Contiguous ORB's and Page Tables could be read in a single transfer