

User-Level Protocols Papers 2a, 2b, 2c

Gregory Marsh
CSE 788.08P: Network-Based
Computing
January 17, 2007

1

Summary of Papers

- 2a: [User-Level Network Interface Protocols](#), 1998.
 - 6 general design issues applied to Myrinet.
- 2b: [EMP: Zero-copy OS-bypass NIC-driven Gigabit Ethernet Message Passing](#), 2001.
 - Design for Gigabit Ethernet implementation.
- 2c: [InfiniBand Host Channel Adapter Verb Implementer's Guide](#), 2003.
 - Specifications for InfiniBand software functions.

2

2a: [User-Level Network Interface Protocols](#), IEEE Computer, Nov. 1998

- Presents 6 design issues for communication architecture applied to late-90's Myrinet.
- Design Goal: improve performance by reducing use of host's OS, CPU, & other resources.
- How: Leverage advanced network interface features. Term: "hardware off-loading."

3

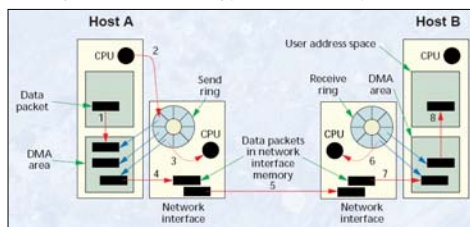
6 Design Issues for Network Communication Architecture

1. Data Transfer
2. Address Translation
3. Protection
4. Control Transfer
5. Reliability
6. Multicast

4

Design Issue 1: Data Transfer

- Path from sender's host memory to receiver's host memory. Schematic of hypothetical "Simple Case:"



5

Design Issue 1: Data Transfer

- More transfer steps & operations means slower performance.
- 3 essential steps:
 1. Host-to-Interface (sending side)
 2. Interface-to-Interface
 3. Interface-to-Host (receiving side)

6

Host-to-Interface Transfer Alternatives

(same for Interface-to-Host on receiving side)

1. Direct Memory Access (DMA)
2. Programmed I/O (PIO)

7

Host-to-Interface Transfer Alternative: Direct Memory Access (DMA)

- Transfers data between host & interface memory without using OS & CPU.
- Fast: entire packet in large burst.
- Problem: Asynchronous with OS. OS may swap target memory of running DMA operation.
- Solution: memory pinning to prevent swap during DMA. Cost: Requires OS system call.
- (More about DMA during Address Translation.)

8

Host-to-Interface Transfer Alternative: Programmed I/O (PIO)

- OS transfers data between host memory & network interface memory.
- Cost: Many I/O bus transactions unless using "write-combining buffers."
- Benefit: Does not need memory pinning.

9

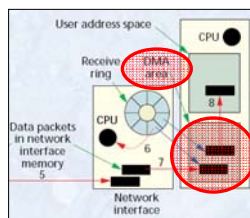
Design Issue 2: Address Translation

- DMA needs physical memory address to copy to/from host memory.
- However, OS forces virtual-to-physical memory mapping. DMA requires a translation.
- Solution Alternatives
 1. PIO. No pinning, no translation. Bus I/O costs.
 2. DMA Area (Cache Coherent DMA + memory copy)
 3. Pinning by Application (Cache Coherent DMA)

10

Address Translation Alternative 2: DMA Area

- Intermediary storage between host memory and network interface.



11

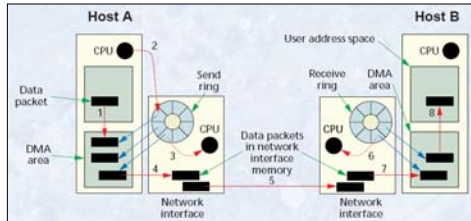
Address Translation Alternative 2: DMA Area

- Reserved and pinned section of host memory. Can't be swapped.
- Network interface does not need address translation. User application copies host memory data to/from DMA area.
- Benefit: interface can access DMA area without OS involvement.
- Cost: extra memory copies in data transfer path. Now has 5 steps: 3 essential + 2 area-to-interface.

12

Address Translation Alternative 2: DMA Area

- Review of Data Transfer steps:



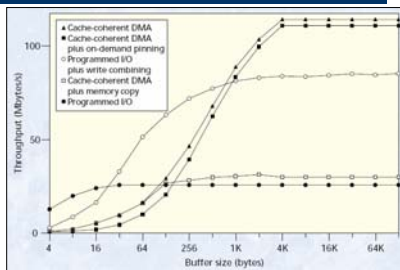
13

Address Translation Alternative 3: Pinning by Application (Cache Coherent DMA)

- Direct transfer between host memory & network interface memory. No DMA area.
- An OS kernel module informs interface of address translation.
- User application pins relevant host memory pages. Drawback: Interface cannot confirm page validity.
- (On-Demand Pinning: kernel and interface interact. Valid pages confirmed.)

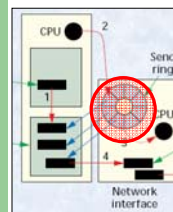
14

Performance of Data Transfer and Address Translation Alternatives



15

Design Issue 3: Protection



- Dealing with multiple users wanting access to network interface.
- Alternatives
 - One user at a time (undesirable).
 - Virtualize interface memory with user address spaces. (Term: Descriptor Management. More later.)

16

Design Issue 4: Control Transfer

- How host applications detect received packets.
- Alternatives
 - I/O Interrupts on CPU. Expensive in terms of time.
 - Polling
 - Flag in network interface memory. Cost: I/O bus traffic.
 - Flag in host cache memory. Faster polling performance.

17

Design Issue 5: Reliability

- Paper focuses on Myrinet which has low error rate. Table 1 shows 11 Myrinet real schemes. Most assume network reliability.
- Some schemes implement flow control.
 - Recovering from buffer overflow: ACK/NACK to cause dropped packet retransmission.
 - Preventing buffer overflow: credit schemes to limit transmission.

18

Design Issue 6: Multicast

- Few of 11 Myrinet schemes in Table 1 support multicast.
- Alternatives at time of publication (1998)
 1. Point-to-point transmission to each destination.
Cost: repeated host to network interface copy for each transmission.
 2. Copy once to interface. Send repeatedly.
- Need for spanning tree protocols. Allows remote hardware to forward packets.

19

2b. EMP: Zero-copy OS-bypass NIC-driven Gigabit Ethernet Message Passing, Supercomputing, November 2001.

- EMP = Ethernet Message Passing, a network communication architecture for Gigabit Ethernet.
- Describes authors' design choices for hardware off-loading.
- Presents experimental results: EMP vs. TCP/IP vs. Myrinet.

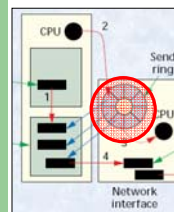
20

6 Design Issues Applied to EMP

1. Data Transfer
 - Direct host to network interface transfer.
 - Does not use DMA area (hence "zero-copy").
2. Address Translation
 - Pinning by Application (Cache coherent DMA).

21

6 Design Issues Applied to EMP



3. Protection (Multiple users on interface)
 - Descriptor management. Store descriptor with location of message in host memory.
 - Cost: limited interface memory.
 - Alternative: Storing descriptors in host's memory would cost a DMA transfer.

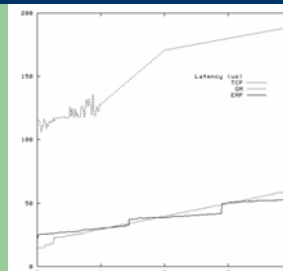
22

6 Design Issues Applied to EMP

4. Control Transfer
 - Traditional Ethernet: I/O interrupt on CPU.
 - InfiniBand: Completion queue. Requires buffer.
 - EMP: polls the descriptors.
5. Reliability
 - ACK with retransmission ability. Supports out of order frames.
6. Multicast: Not supported.

23

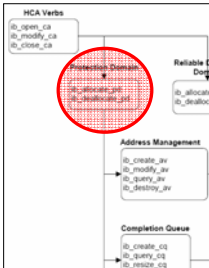
Experimental Results: Latency



- EMP is comparable to Myrinet (GM).
- (Horizontal axis is message size in kilobytes.)

24

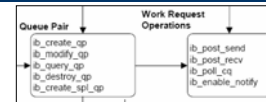
3. Protection



- Has primacy in verb group architecture.
- Protection Domain (PD) required to create any resource on, and do anything with, the HCA.
- Provides security for queue pairs and memory regions.

31

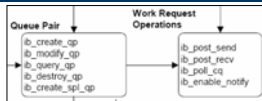
1. Data Transfer in InfiniBand Queue Pairs & Work Requests



- What `ib_create_qp` verb needs to do:
 - Allocate QP resources.
 - Register creation with HCA.
 - Set initial QP states.
 - Return a QP Handle.

32

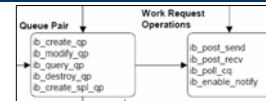
1. Data Transfer in InfiniBand Create Queue Pair, Continued



- Input parameters required for `ib_create_qp` :
 - QP Type (Reliable Connection, datagram, etc)
 - Max # of WQE's.
 - WQE page size.
 - ID of associated Completion Queue.
 - Protection Domain (PD).

33

1. Data Transfer in InfiniBand Work Requests



- Work Request Operations
 - Verify QP state is appropriate
 - Verify that a WQE is available
 - Select WQE type (Send, Recv, RDMA, etc)
 - Inform HCA that work is available for network transmission.

34

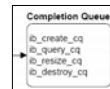
2. Address Translation in InfiniBand

- Primary verb: Register Memory Region (`ib_register_mr`).
- Input: HCA Handle, PD, Virtual Address, Data length, Access control rights.
- Output: Memory region handle, L_KEY (Local access rights), R_KEY (Remote access rights).
- HCA uses virtual address and the KEYS to determine physical memory pages for DMA network operation.

35

4. Control Transfer in InfiniBand Completion Queue Polling

- Completion Queue's contain CQE's.
- CQE must provide following info:
 - Type of operation completed.
 - Length of data.
 - Source QP (for datagram QP's)
 - Communication path to remote end (for UD QP's)
 - Free count of WQE's (for RD QP's). Use to avoid send/receive buffer overflow (flow control).



36

5. Reliability in InfiniBand

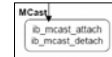
Class of Service	Description
Reliable Connection	acknowledged - connection oriented
Reliable Datagram	acknowledged - multiplexed
Unreliable Connection	unacknowledged - connection oriented
Unreliable Datagram	unacknowledged - connectionless
Raw Datagram	unacknowledged - connectionless

- Options for reliability due to many classes of InfiniBand transmission.

37

6. Multicast in InfiniBand

- Require verbs to add & remove QP's from the multicast group.
- Details are vendor specific.



38

Conclusion: 6 Design Issues for Communication Architecture

Alternatives within each have costs & benefits.

1. Data Transfer
2. Address Translation
3. Protection
4. Control Transfer
5. Reliability
6. Multicast

39