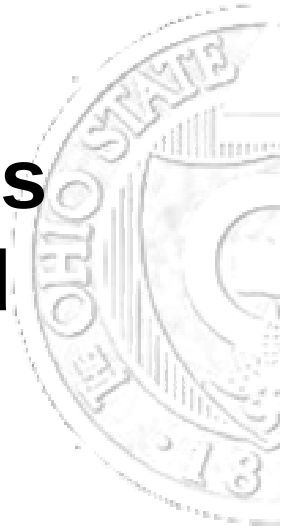
NETWORK-BASED COMPUTING LABORATORY



Can Memory-Less Network Adapters Benefit Next-Generation InfiniBand Systems?

Sayantan Sur, Abhinav Vishnu, Hyun-Wook Jin, Wei Huang

and D. K. Panda

{surs, vishnu, jinhy, huanwei, panda}@cse.ohio-state.edu

Network Based Computing Laboratory

The Ohio State University

OHIO STATE

Presentation Outline

- Introduction and Motivation
- Problem Statement and Approach Used
- Overview of InfiniHost III NIC Architecture
- Micro-benchmark Level Evaluation
- Application Level Evaluation
- Conclusions and Future Work

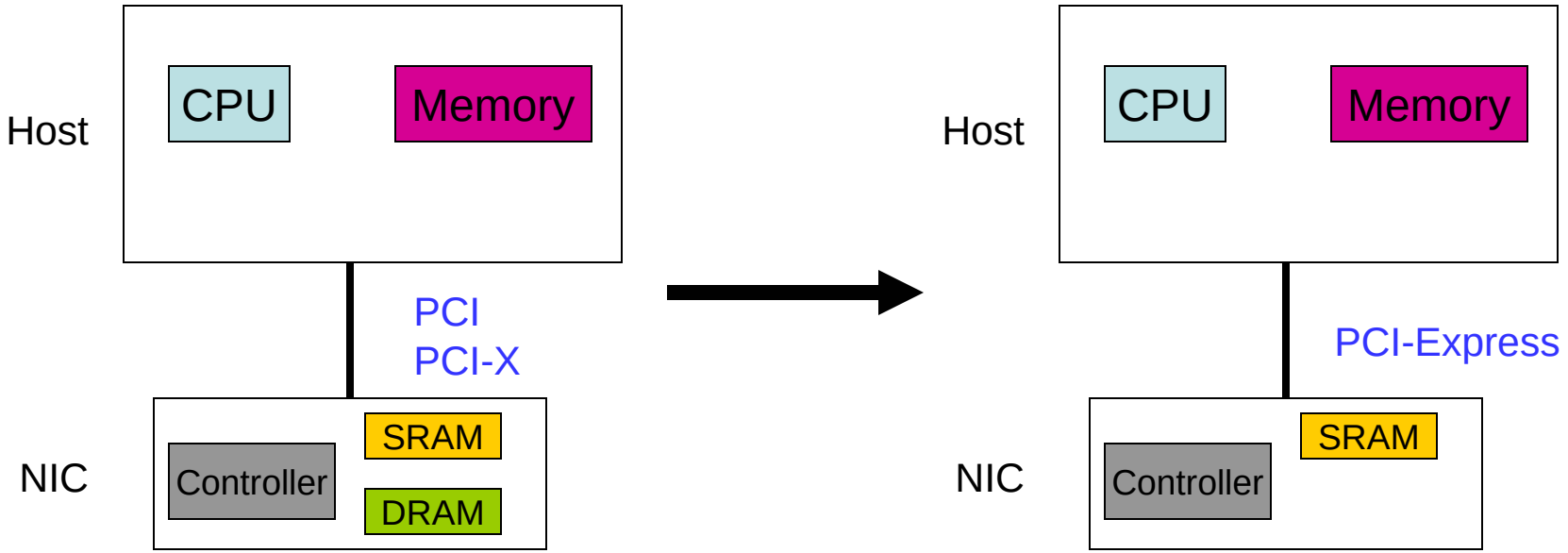
Introduction

- InfiniBand is an industry standard
- Gaining momentum as a high performance interconnect for
  - HPC Clusters
  - Data centers
  - File/Storage Systems
- Many other features
  - One sided operations (RDMA, RDMA scatter/gather, atomics)
  - Hardware multicast
  - Quality of Service
- Recently introduced Double Data Rate (DDR)
- Low latency (~2 us), High bandwidth (~1500 MB/s)

InfiniBand and PCI Express

- Previous generation InfiniBand adapters used
  - PCI
  - PCI-X interfaces
- Latest generation adapters use
  - PCI-Express interface
  - Hyper transport Interface (PathScale)
- Features of PCI-Express
  - Third Generation I/O Bus Architecture
  - Uses point-to-point serial interface
  - Has multiple lanes, each delivering 250 MB/s
  - X8 PCI-Express delivers 8*250 = 2GB/s in each direction
  - Aggregate bandwidth 4 GB/s
  - I/O devices connected directly to memory controller
    - Reduces latency for memory access

Can Network Adapters go Memory-Less with PCI-Express?



Mellanox has introduced such a Memory-Less Network Adapter for PCI-Express based Systems

Pros and Cons of Memory-Less Adapters

- Pros
  - Lesser design complexity of NIC
  - Less power consumption by NIC
  - Overall lower cost for NIC
- Cons
  - Can Memory-Less operation hurt performance for end applications?
  - Will NIC Memory-Less operation result in increased Host memory usage?

Presentation Outline

- Introduction and Motivation
- Problem Statement and Approach Used
- Overview of InfiniHost III NIC Architecture
- Micro-benchmark Level Evaluation
- Application Level Evaluation
- Conclusions and Future Work

Problem Statement

- Can next generation InfiniBand systems take advantage of Memory-Less Adapters?
- What will be the Impact on Application Performance?

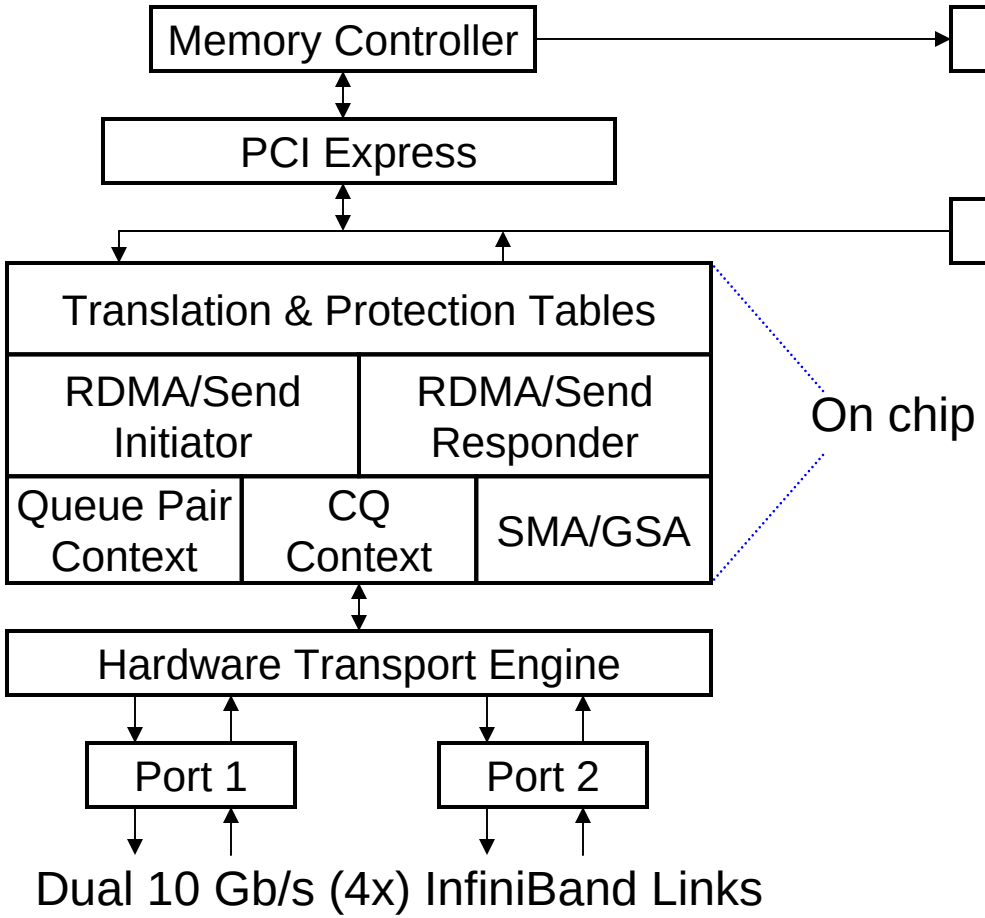
Approach Used

- Micro-benchmarks
  - Design suitable experiments and study the impact of NIC vs. Host memory
    - Experiments focus on critical NIC elements
      - Queue Pair context
      - Virtual-to-physical address translation entries
    - Cache Miss Penalty
    - Effect of Pending I/O Bus Operations
    - Cache misses for Address Translation
    - Host memory usage
- Applications-Level Evaluation

Presentation Outline

- Introduction and Motivation
- Problem Statement and Approach Used
- Overview of InfiniHost III NIC Architecture
- Micro-benchmark Level Evaluation
- Application Level Evaluation
- Conclusions and Future Work

Overview of InfiniHost III HCA



- HCA keeps IBA data structures in InfiniHost Context Memory (ICM)
- ICM can be placed in either NIC or Host memory
- There is an ICM cache “On chip”
- In absence of NIC memory, ICM is located in Host memory

MT25218 HCA Architecture (Courtesy Mellanox)

Presentation Outline

- Introduction and Motivation
- Problem Statement and Approach Used
- Overview of InfiniHost III NIC Architecture
- Micro-benchmark Level Evaluation
- Application Level Evaluation
- Conclusions and Future Work

Micro-benchmark Level Evaluation

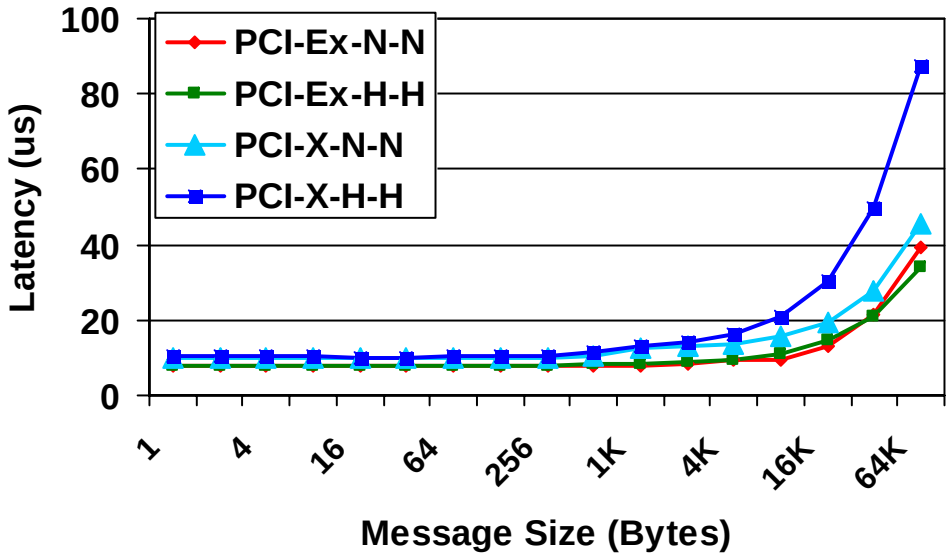
- Host-host and NIC-NIC memory data transfer with PCI-Express and PCI-X
- ICM Cache Miss Penalty
- Effect of Pending I/O Bus Operations
- Cache misses for Address Translation
- Host memory usage

Experimental Platform

- Four dual Intel Xeon 3.2 GHz EM64T
- System Memory: 786MB
- Operating System: RedHat AS3 (Update 2)
- InfiniBand
  - Mellanox MT25128 HCA
    - Can be run in Mem and MemFree modes through firmware modifications
  - Mellanox 24 port switch
  - IB Golden CD version 1.7.0

Host-Host and NIC-NIC with PCI-Express and PCI-X

- Experiment to measure one way memory access latency of Host memory by NIC and NIC memory by NIC itself
- One process has two queue pairs connected to each other
  - Data is moved in two ways
    1) Host to Host memory
    2) NIC to NIC memory
  - The experiment is repeated for both PCI-Express, PCI-X



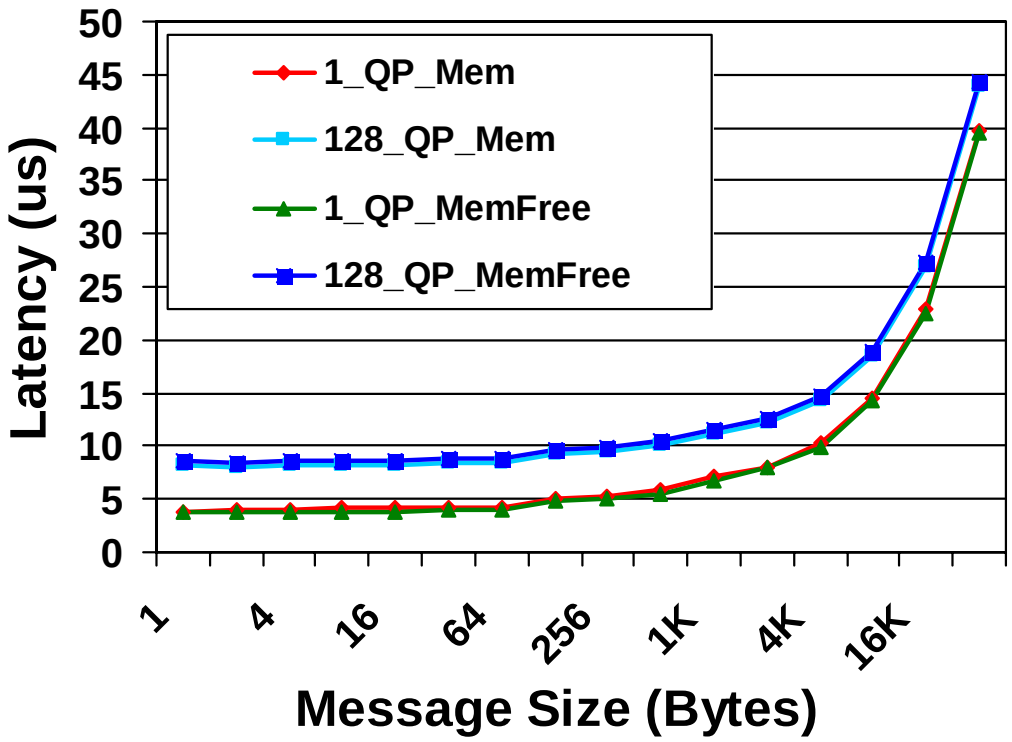
Host/NIC Memory Access Latency

Observation:

For PCI-Express systems, Host and NIC memory access times are almost the same!

ICM Cache Miss Penalty

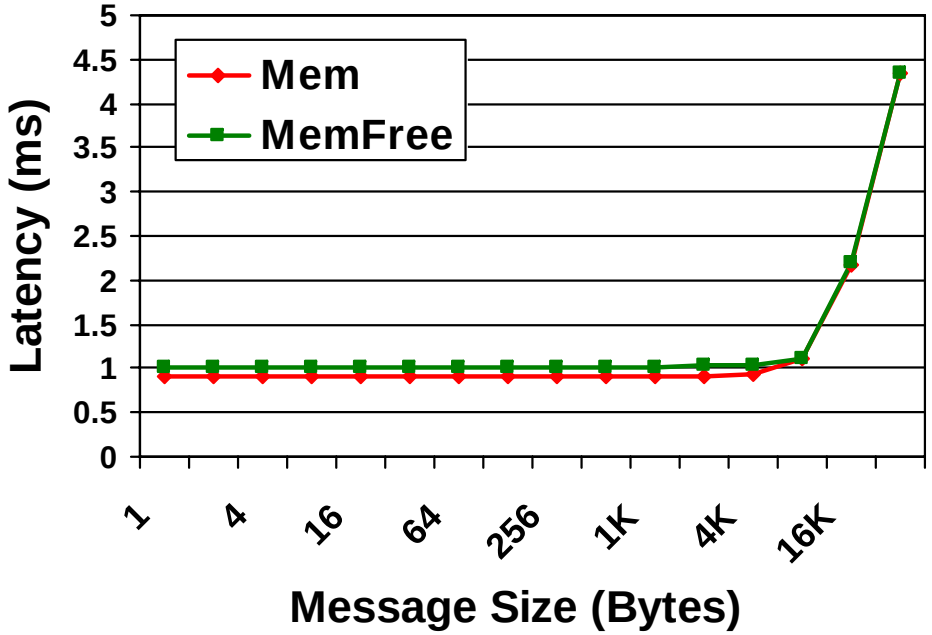
- This experiment causes ICM cache misses for QP Context
- Two processes have *n* (1-128) QPs between themselves
- Ping-pong latency test; QPs accessed in a cyclically
- When *n* is large (e.g. 128), causes ICM cache miss for QP context
- 128 QP ping-pong latency is higher than 1 QP latency, indicating ICM cache misses



In case of both Mem and MemFree cards, increase in ping-pong latency is almost the same!

Effect of pending I/O bus operations

- Experiment to see impact of I/O bus transactions on ICM cache miss penalty
- Two processes scatter data (non-blocking) on128 QPs and wait for completion, QPs are accessed cyclically
- This pattern means when QP context access incurs an ICM cache miss, there are I/O bus transactions



In case of both Mem and MemFree cards, increase in scatter latency is very less (10% for messages up to 8K)!

ICM Cache misses for Address Translation

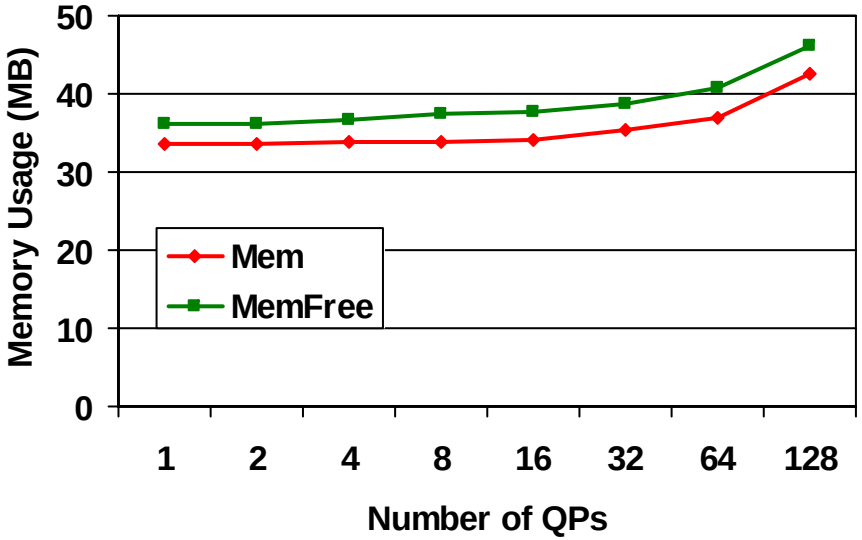
- Two processes are connected over one QP
- Conduct bandwidth test with decreasing percentage of buffer reused
- Since the translation entries are per page, lesser buffer reuse implies more and more misses for Address Translation

Reuse (%)	MemFree	Mem
100	918 MB/s	926 MB/s
75	918 MB/s	925 MB/s
50	919 MB/s	927 MB/s
25	918 MB/s	926 MB/s
0	917 MB/s	922 MB/s

There is almost no difference in the case Mem and MemFree cards!

Host Memory Usage

- In MemFree HCAs the Host memory is used to store ICM
- Need for memory increases as the number of connections mainly due to QP control buffers etc.?
- We are interested in increase of memory usage with number of connections and not the absolute number (depends on various libraries)



The MemFree mode consumes extra host memory but the difference is not much, at the same time allowing more efficient use of memory

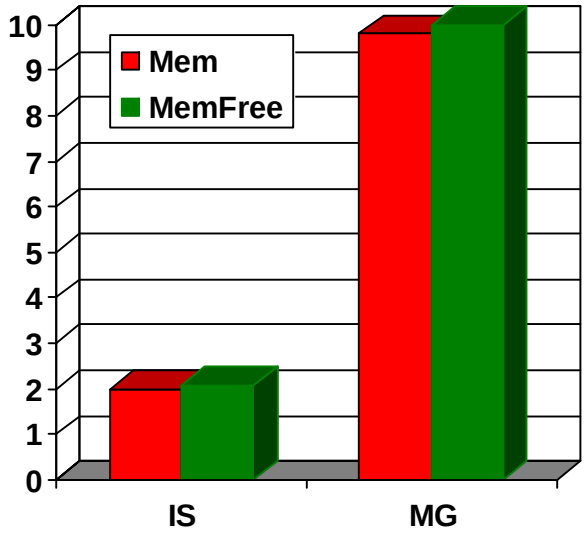
Presentation Outline

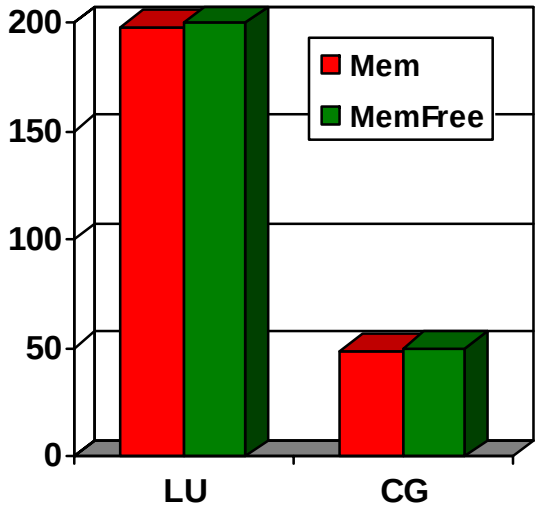
- Introduction and Motivation
- Problem Statement and Approach Used
- Overview of InfiniHost III NIC Architecture
- Micro-benchmark Level Evaluation
- Application Level Evaluation
- Conclusions and Future Work

Use of MVAPICH (OSU MPI for InfiniBand) Library

- Open Source Distributions
  - MVAPICH 0.9.5 (MPI-1)
    - Highly optimized for InfiniBand
    - RDMA-based design for point-to-point and collectives
    - Exploits InfiniBand hardware-level multicast for MPI Broadcast
    - Efficient intra-node shared memory support (bus-based and NUMA-based)
    - Multi-rail support (multiple adapters and multiple ports/adapter)
    - Upcoming support for OpenIB/Gen2 and uDAPL
- Directly downloaded and being used by more than 250 organizations worldwide (across 28 countries)
- Available in the software stack distributions of IBA vendors
- Empowers multiple InfiniBand clusters in the TOP 500 list
- URL: http://nowlab.cse.ohio-state.edu/projects/mpi-iba/

Application Level Evaluation





- NAS benchmarks: Integer Sort (IS), Multi Grid code fragment (MG), LU decomposition (LU), Conjugate Gradient (CG)
- Executed on 8 processes on 4 nodes

Almost same (<1% difference) performance between Mem and MemFree mode of operation

Conclusions

- Carried out an in-depth study of InfiniBand MemFree HCAs
- Designed several new Micro-benchmarks and evaluated them
- Basic ICM cache miss penalty is almost the same in both Mem and MemFree operation
- Slightly higher memory usage, but more efficient in utilizing all available memory in system
- NAS benchmark evaluation reveals almost no difference (<1%) in performance
- MemFree cards can provide good performance with low cost

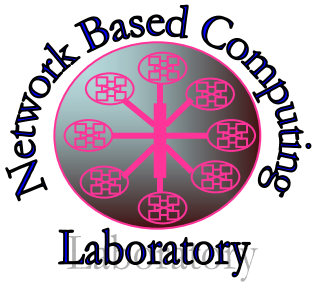
Acknowledgements

Our research is supported by the following organizations

- Current Funding support by

- Current Equipment donations by

Web Pointers



http://www.cse.ohio-state.edu/~panda/

http://nowlab.cse.ohio-state.edu/

MVAPICH Web Page

http://nowlab.cse.ohio-state.edu/projects/mpi-iba/

E-mail: panda@cse.ohio-state.edu