Can High Performance Software DSM Systems Designed With InfiniBand Features Benefit from PCI-Express?

R. Noronha and D. K. Panda

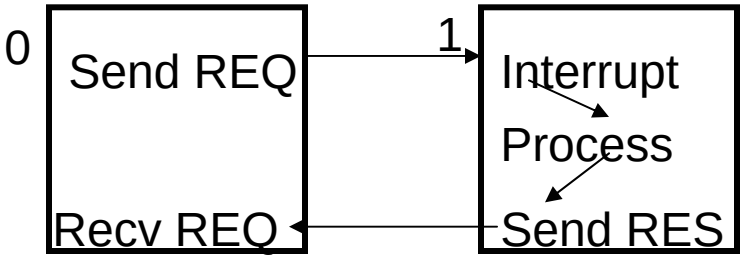
Network Based Computing Laboratory (NBCL)

The Ohio State University

Presentation Outline

- **Introduction and Motivation**
- I/O Interconnection Technologies
- DSM Protocols
- Experimental Results
- Conclusions and Future Work

Introduction



- DSM Protocols
  - Communication intensive
  - Wait time increased by asynchronous request-response model
  - Possible to push request processing to the network
- InfiniBand
  - Low Latency (< 5 µs)
  - Aggregate bandwidth upto 1 (GB/s) with PCI-X 133 MHz
  - Limited by the shared nature of the PCI-X architecture
- PCI-Express
  - Serial point-to-point links
  - Improved latency
  - Higher bandwidth (upto 4 GB/s aggregate)

GB/s = GigaBytes/s MB/s=MillionBytes/s

Motivation

- What is the impact of PCI-Express on DSM systems ?
- DSM protocols and PCI-Express in general
  - Benefit from better latency and aggregate bandwidth
  - Potential for lower wait times
  - Can exploit improved performance of InfiniBand
  - Though still the problem of the asynchronous handler
- Synchronous protocols like NEWGENDSM (previously presented in DSM'04) in particular
  - More tied to network/interconnection performance
  - Benefit from the improved bus performance in particular
  - Impact of asynchronous handler reduced
  - More Potential Benefit

Presentation Outline

- Introduction and Motivation
- I/O Interconnection Technologies
- DSM Protocols
- Experimental Results
- Conclusions and Future Work

I/O Interface Technologies

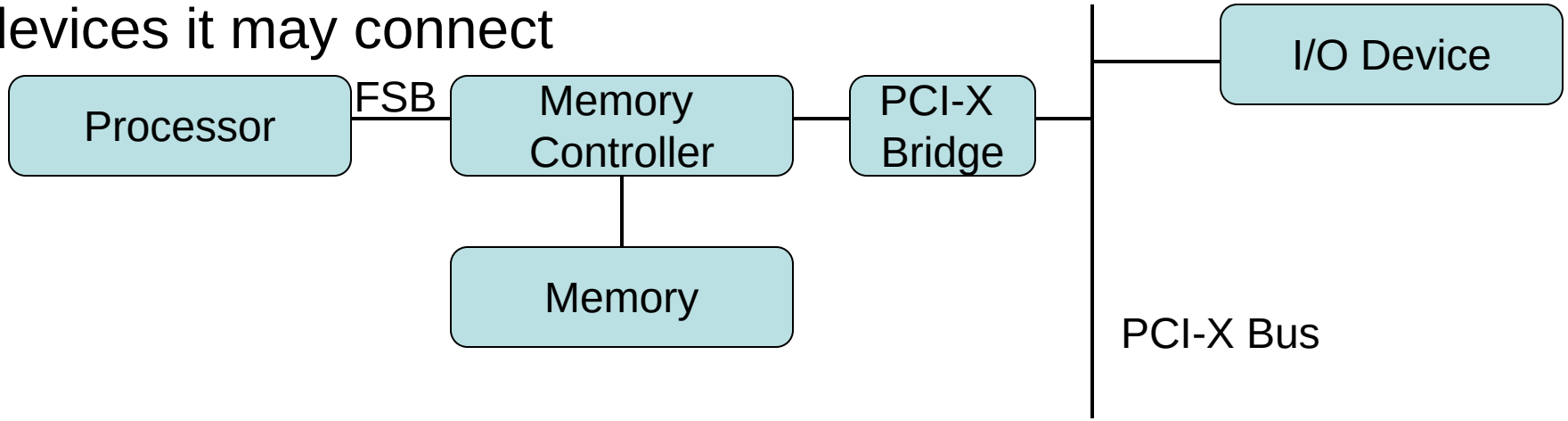
- PCI
  - Dominant Technology
  - Connects I/O and memory
  - Low Bandwidth (132 MB/s)
- Successors
  - PCI-X
  - PCI-Express

GB/s = GigaBytes/s MB/s=MillionBytes/s

PCI-X

- Shared architecture
- 64-bits, 133 MHz most popular
- Aggregate Bandwidth up to 1 GB/s
- Scales in terms of bandwidth or number of devices it may connect

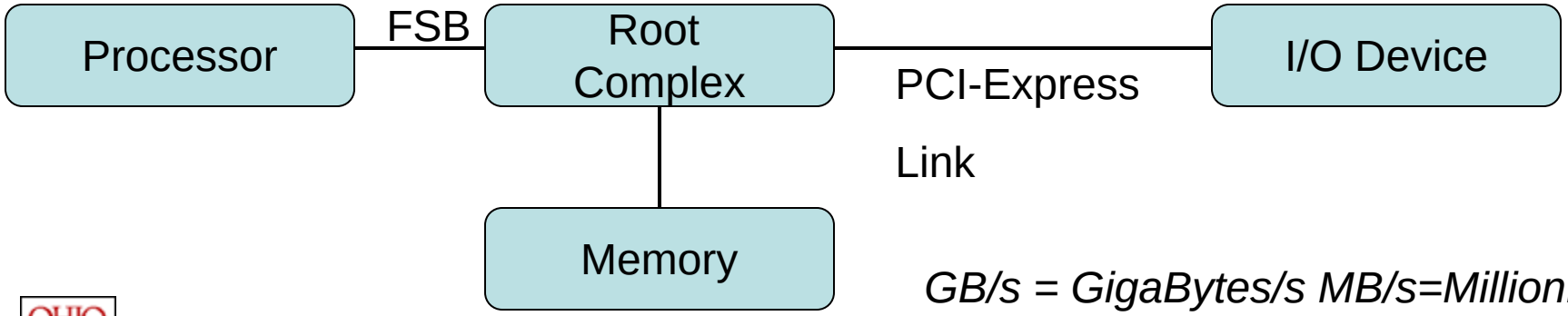
Interface	Peak Bandwidth (MB/s)	Cards/Bus
PCI(66 MHz)	532	1-2
PCI-X(66 MHz)	532	4
PCI-X (133 MHz)	1066	1-2
PCI-X (266 MHz)	2132	1



GB/s = GigaBytes/s MB/s=MillionBytes/s

PCI-Express

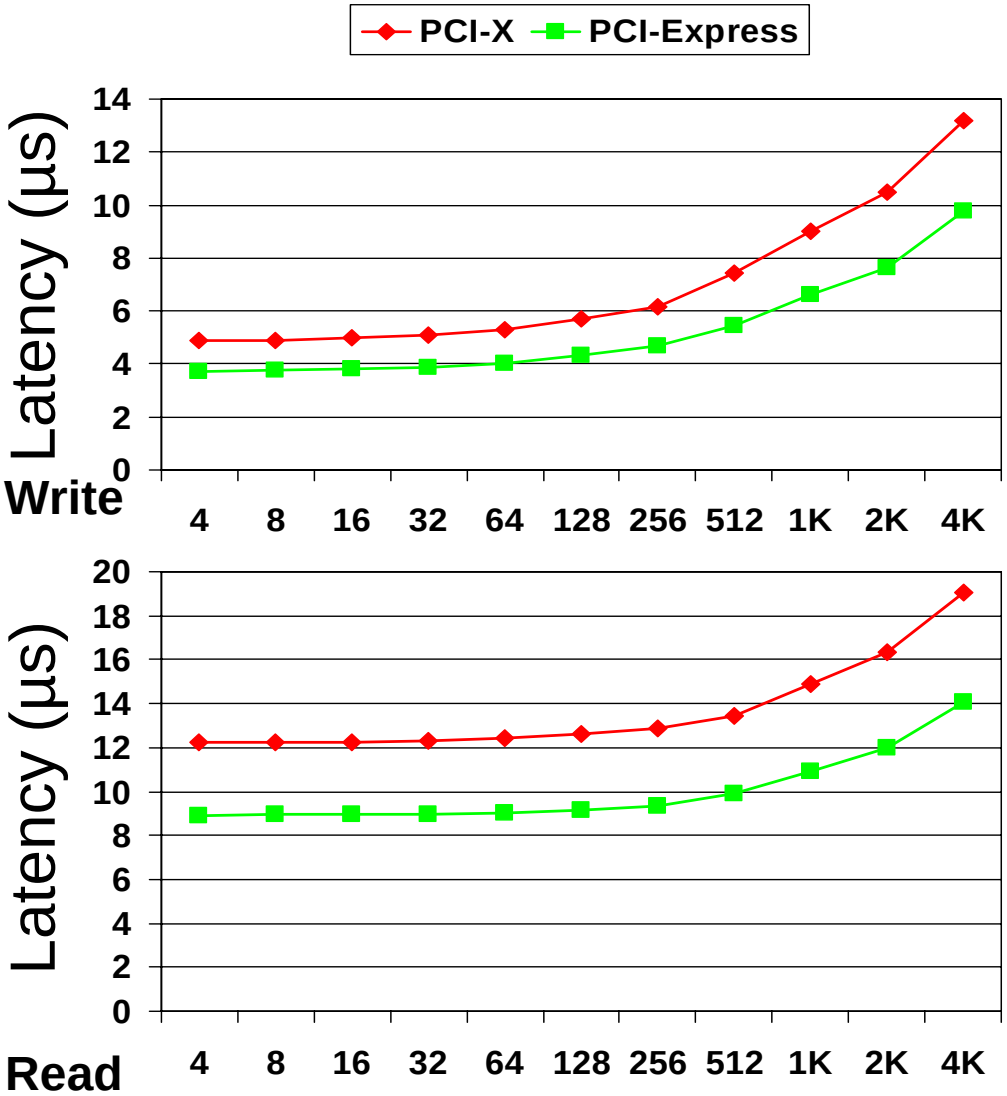
- Serial point-to-point links (lanes) between memory and I/O devices
- Bandwidth scaling achieved by increasing lanes
- Device scaling achieved through point-to-point links
- Aggregate bandwidth upto 4 GB/s (x8)



GB/s = GigaBytes/s MB/s=MillionBytes/s

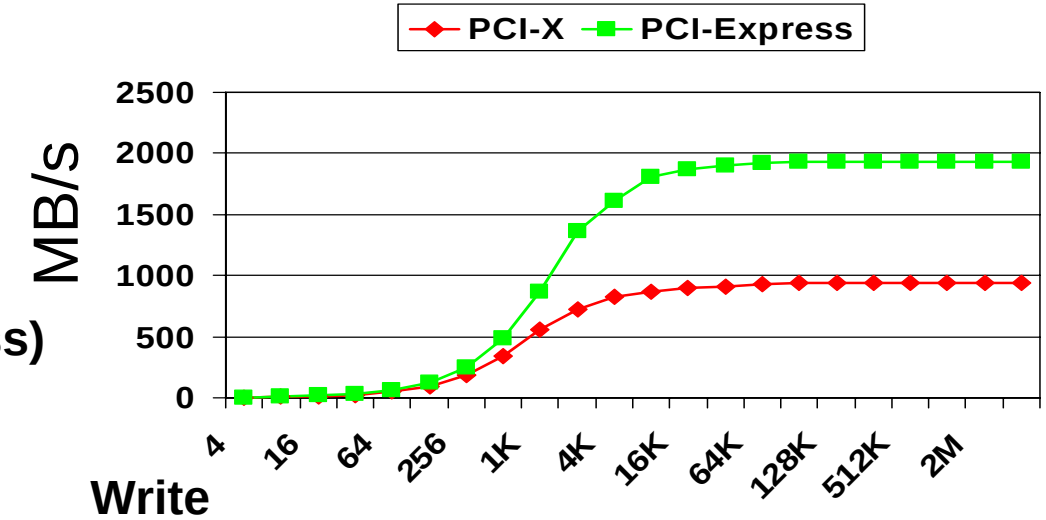
RDMA Latency

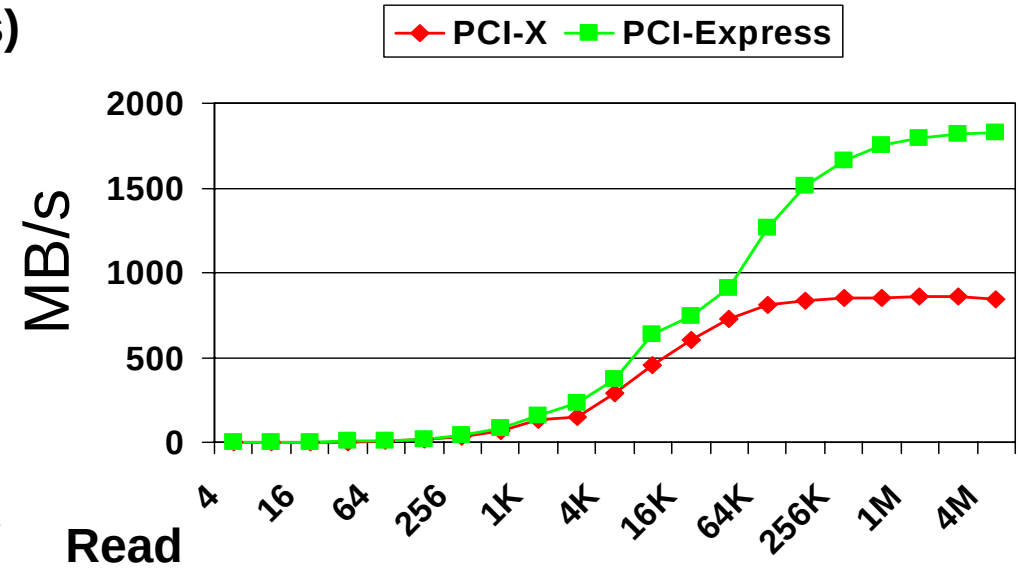
- For a 4 byte message
  - RDMA Write
    - 4.88 µs (PCI-X)
    - 3.73 µs (PCI-Express)
  - RDMA Read
    - 12.23 µs (PCI-X)
    - 8.92 µs (PCI-Express)



RDMA Bi-directional Bandwidth

- **For large messages**
  - **RDMA Write**
    - **946 MB/s (PCI-X)**
    - **1932 MB/s (PCI-Express)**
  - **RDMA Read**
    - **840 MB/s (PCI-X)**
    - **1840 MB/s (PCI-Express)**





GB/s = GigaBytes/s MB/s=MillionBytes/s

Presentation Outline

- Introduction and Motivation
- I/O Interconnection Technologies
- DSM Protocols
- Experimental Results
- Conclusions and Future Work

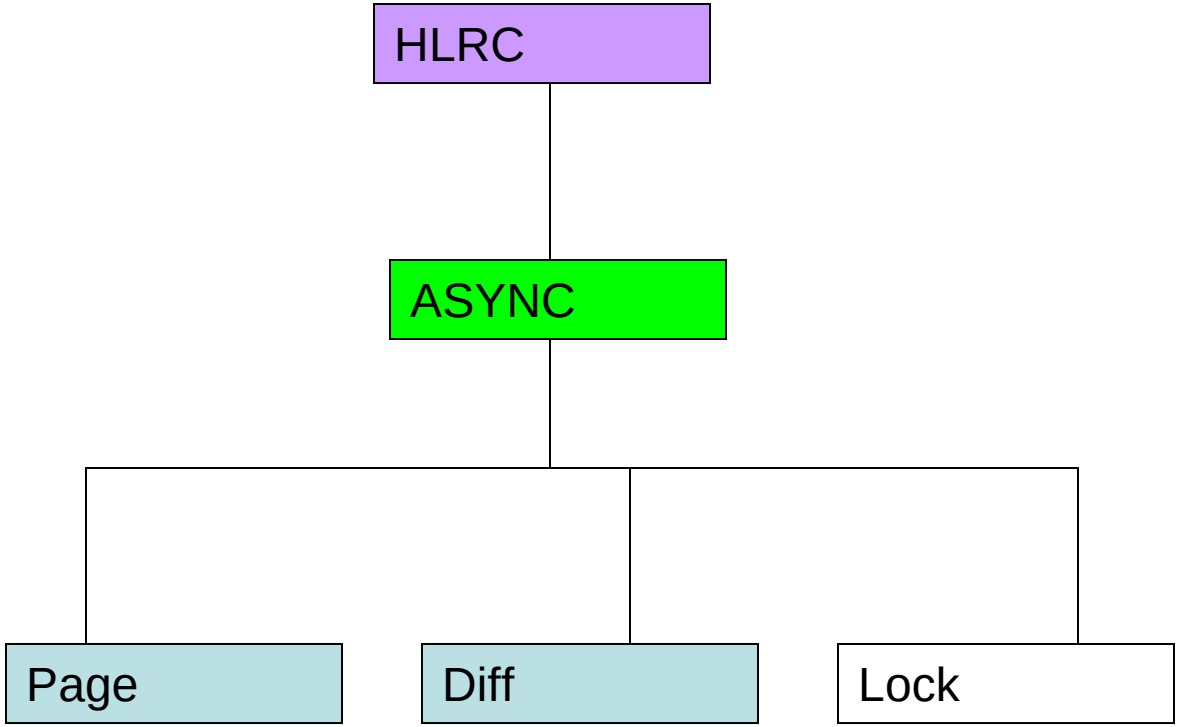
DSM Protocols

- HLRC
  - HLRC/VIA (Rutgers)
  - Home Based Lazy Release Consistency Model
  - Page Based DSM System
    - Internal basic operations
      - Page
      - Diff
      - Lock
  - Use interrupts
  - Referred to as ASYNC

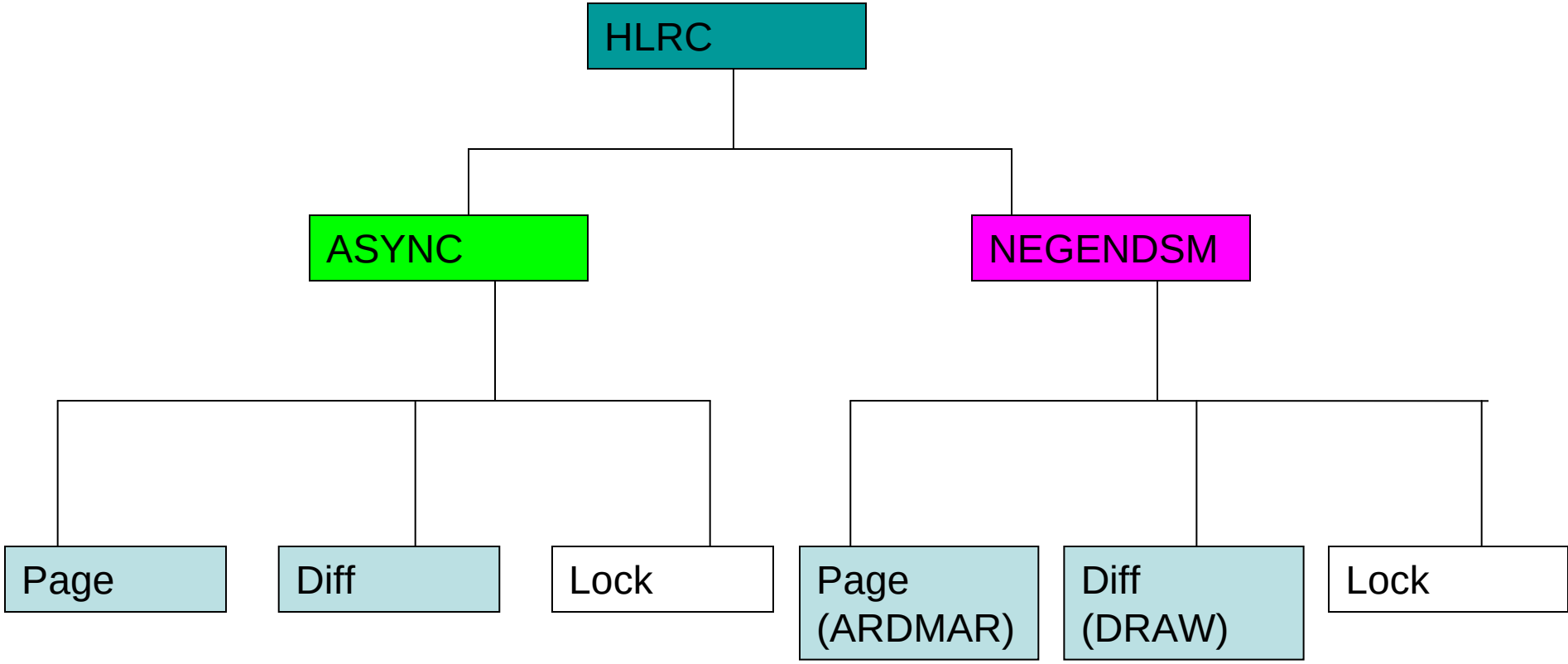
NEWGENDSM

- Design consists of 2 protocols
  - ARDMAR (Atomic and RDMA Write → page fetch)
  - DRAW (Diff using RDMA Write)
- ARDMAR is a synchronous protocol
- DRAW is a hybrid protocol
- NEWGENDSM = ARDMAR + DRAW
- Work presented in DSM'04
  - "Designing High Performance DSM Systems using InfiniBand Features"

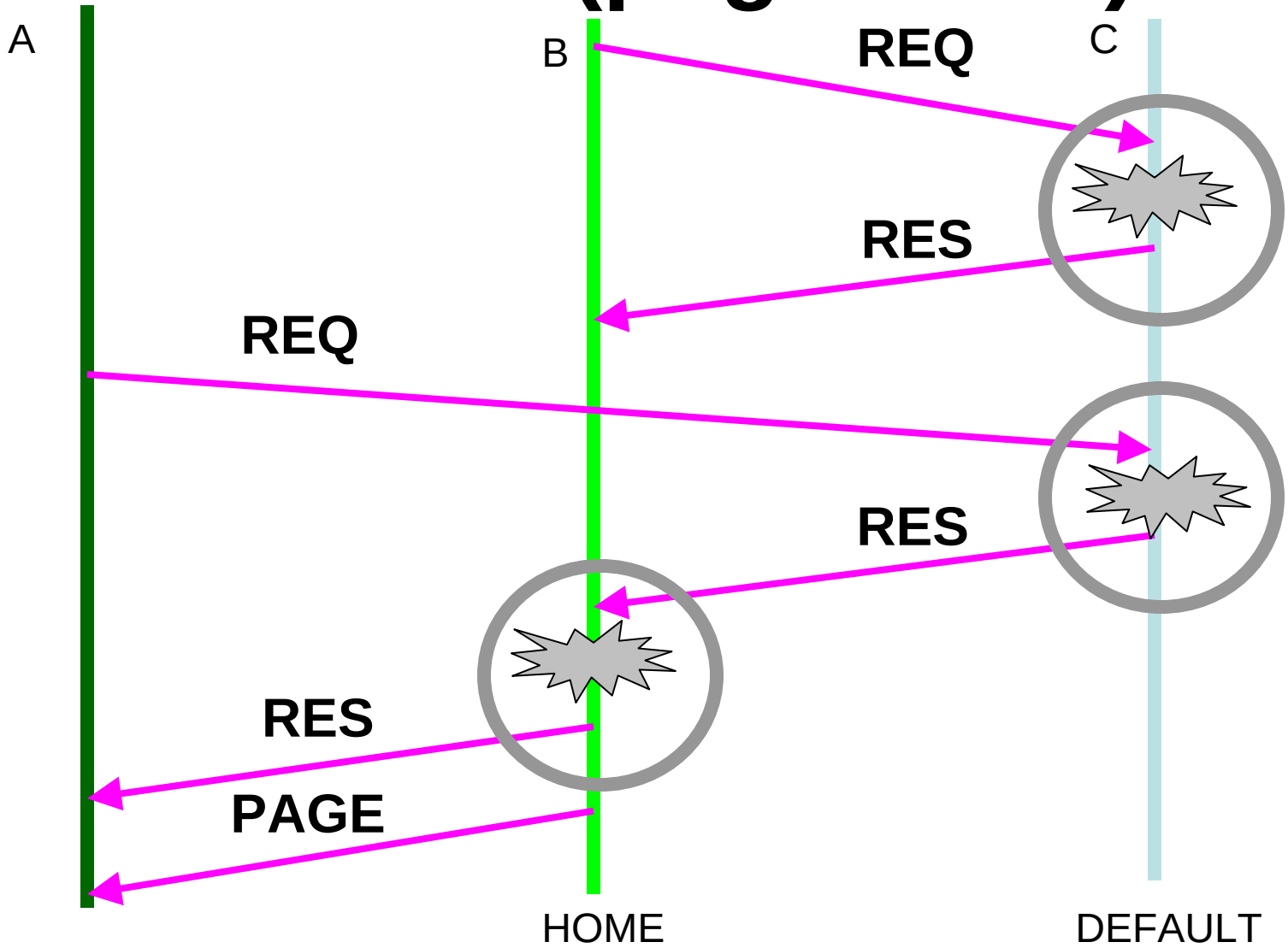
HLRC Design



NEWGENDSM


HLRC
ASYNC
NEGENDSM
Page
Diff
Lock
Page (ARDMAR)
Diff (DRAW)
Lock

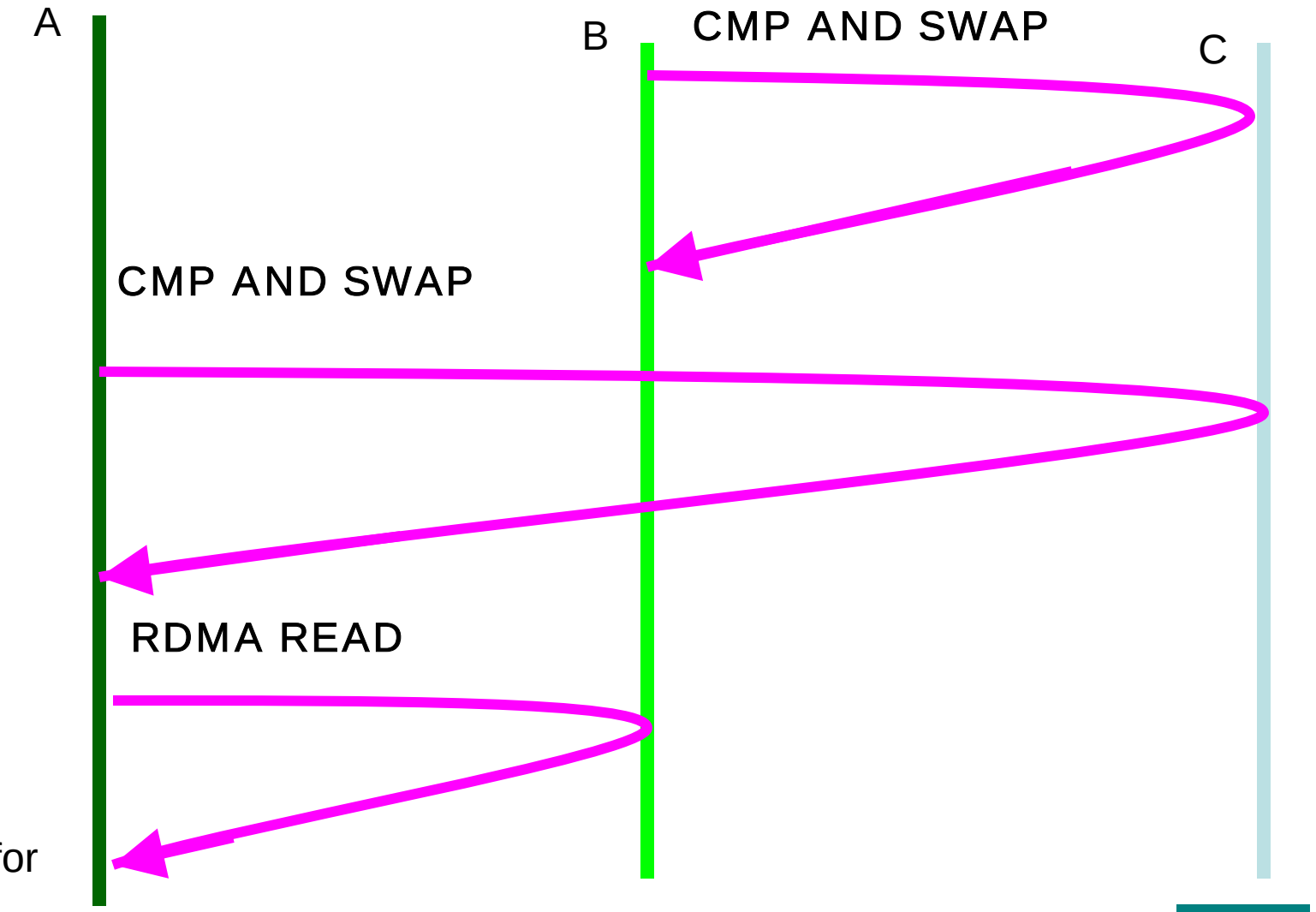
ASYNC (page fetch)



Home for page 2

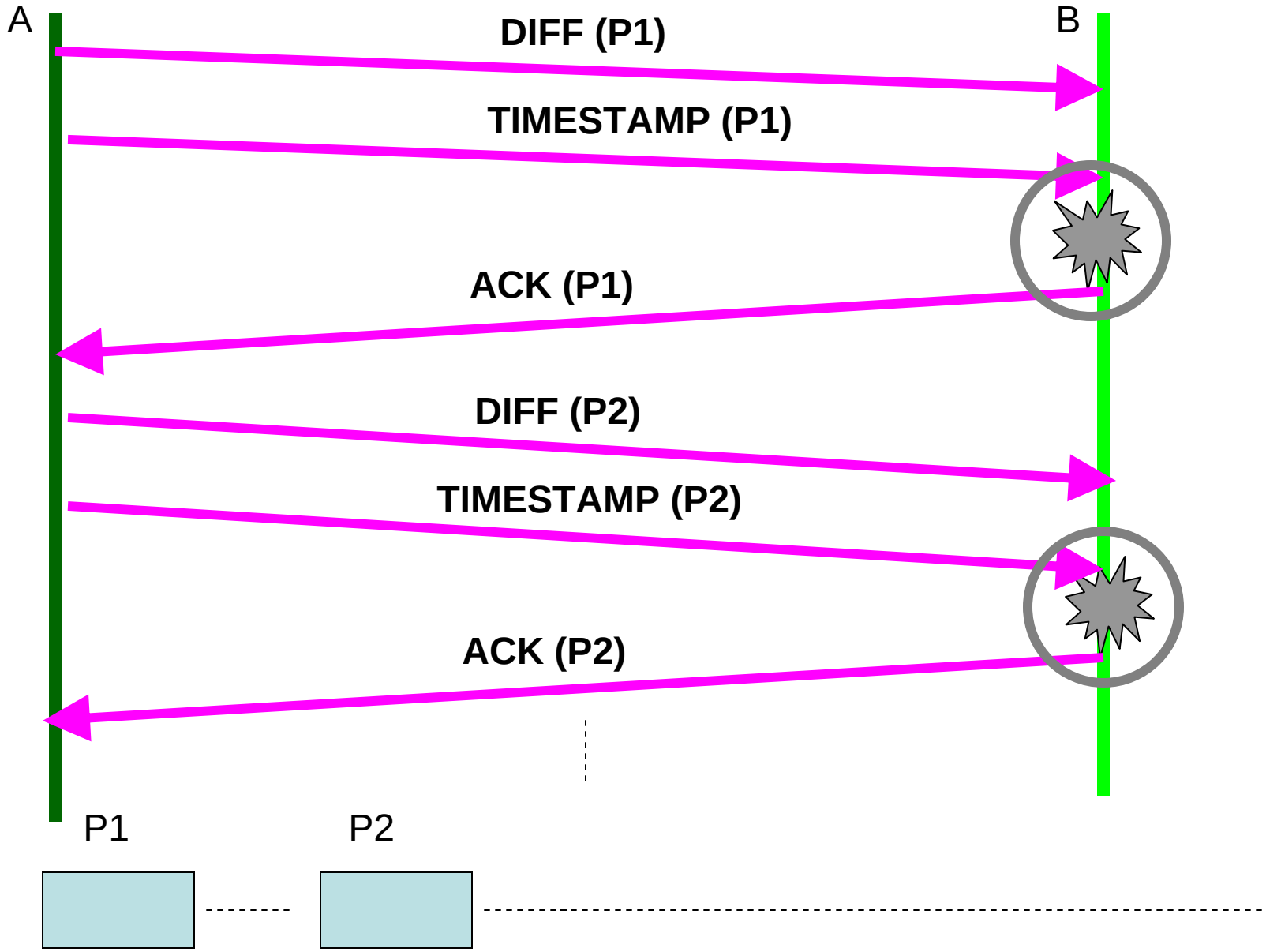
B B B

ARDMAR (Atomic and RDMA Read)

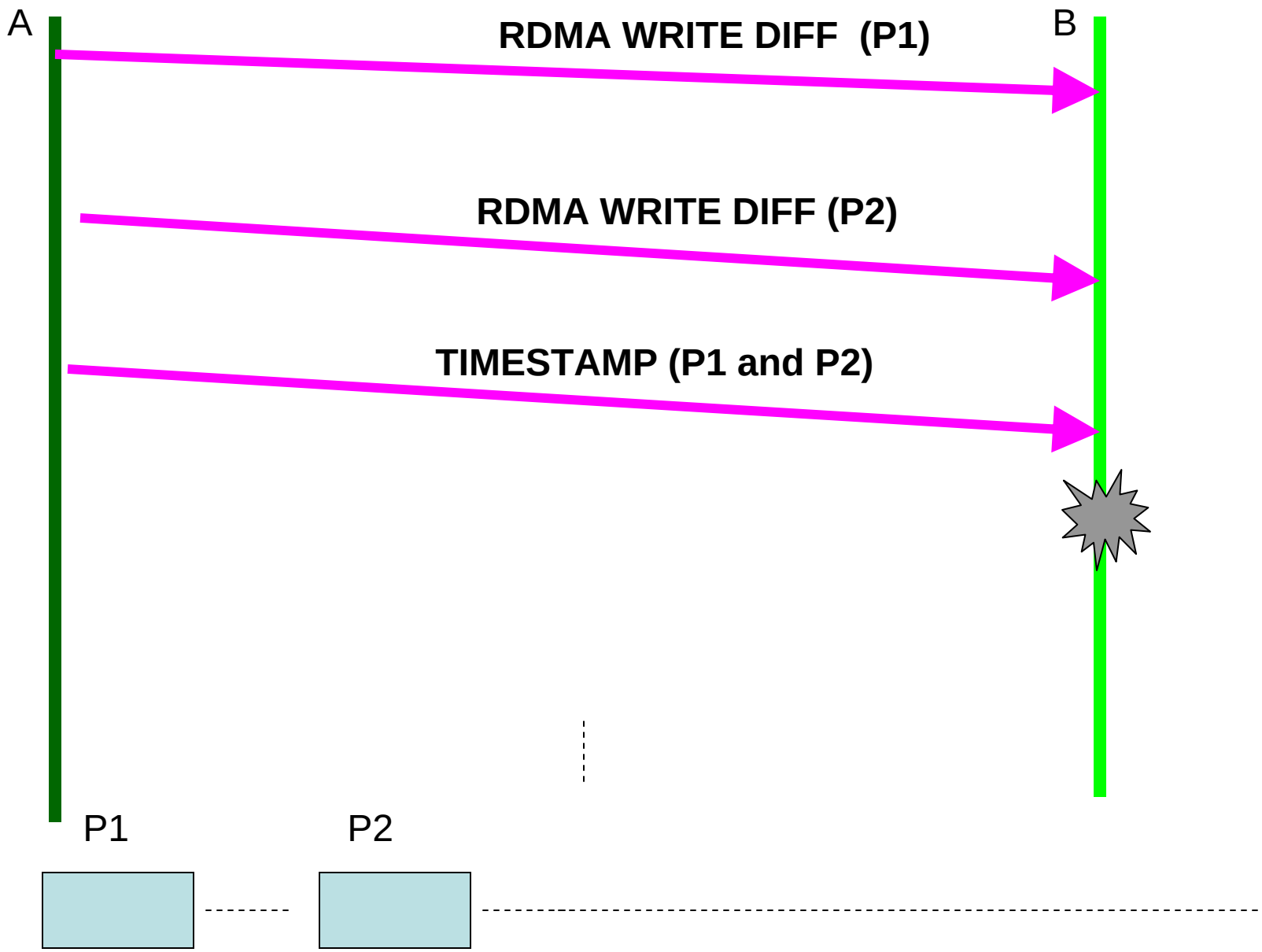


Home for page 2

ASYNC (diff)


A
B
DIFF (P1)
TIMESTAMP (P1)
ACK (P1)
DIFF (P2)
TIMESTAMP (P2)
ACK (P2)
P1
P2

DRAW


A
B
RDMA WRITE DIFF (P1)
RDMA WRITE DIFF (P2)
TIMESTAMP (P1 and P2)
P1
P2

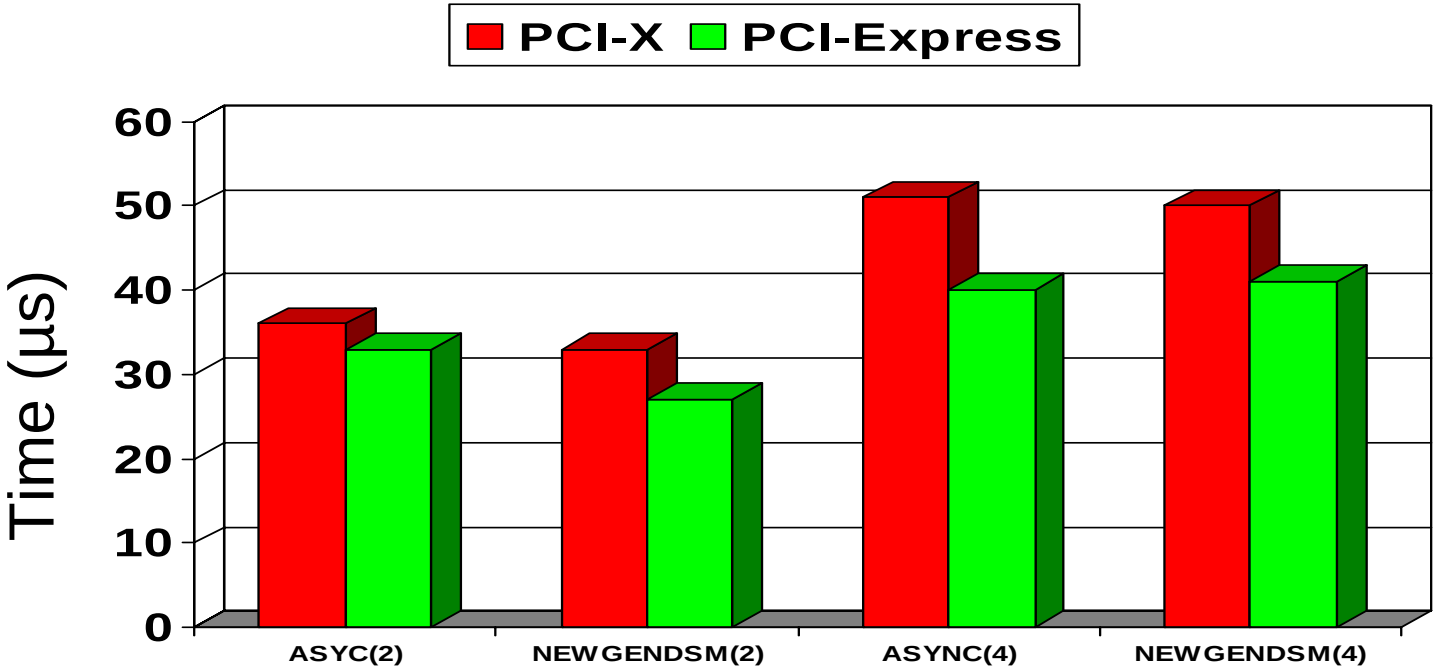
Presentation Outline

- Introduction and Motivation
- I/O Interconnection Technology
- DSM Protocols
- Experimental Results
- Conclusions and Future Work

Experimental Setup

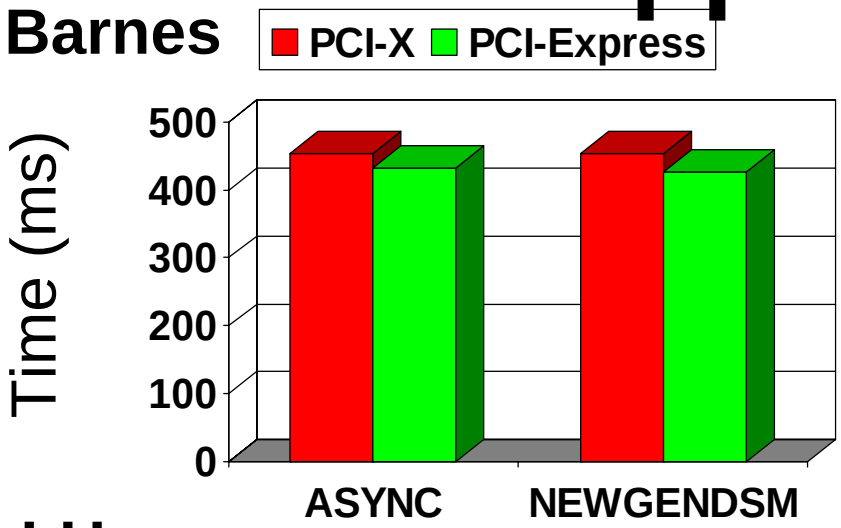
- 4 node cluster
  - Xeon 3.4 GHz (EM64T)
  - 512 MB memory
  - 133 MHz 64-bit PCI-X bus
  - PCI-Express x8
  - InfiniBand network
- InfiniScale MTS2400 24 port 4X switch
- Mellanox InfiniHost MT23108 DualPort 4X HCA's (PCI-X)
- Mellanox InfiniHost MT25208 DualPort 4X HCA's (PCI-Express)

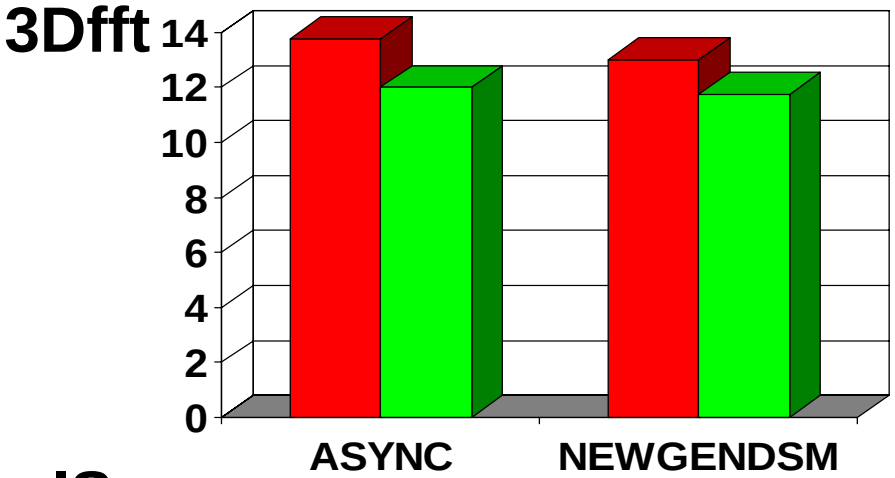
Page Fetch

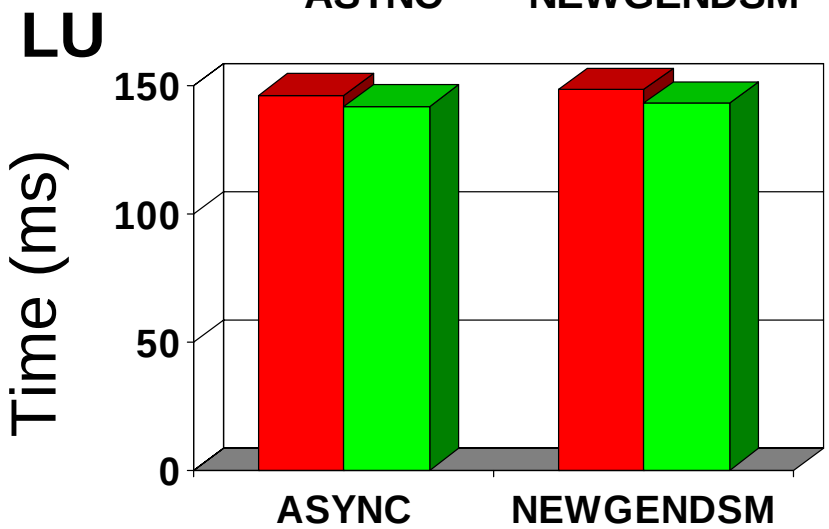


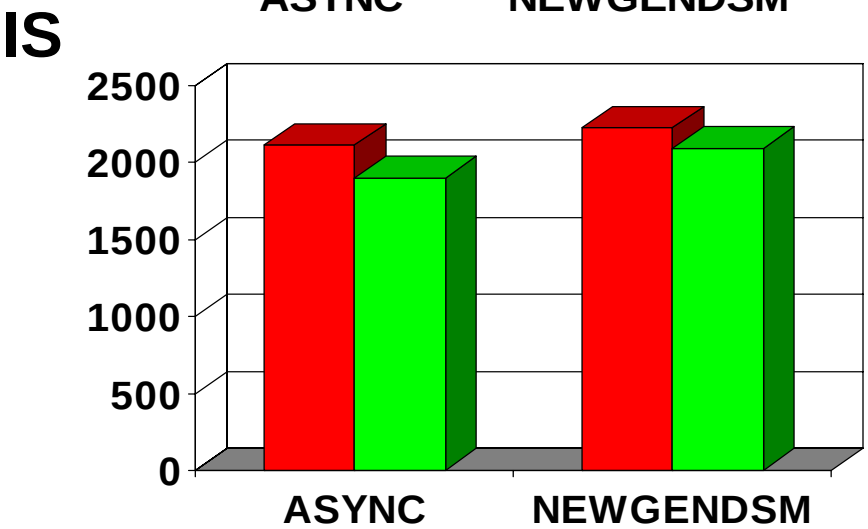
- Average time to fetch a page when a number of nodes are accessing it
- 6-15% improvement at 2 nodes with PCI-Express
- 18-21% improvement at 4 nodes with PCI-Express

Applications



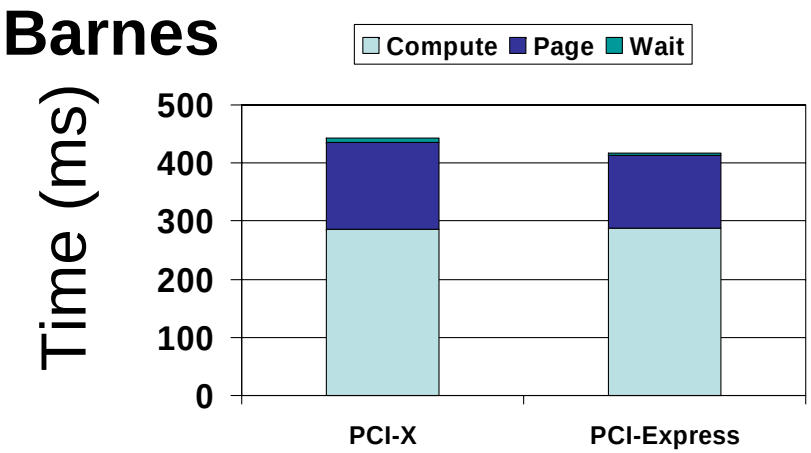


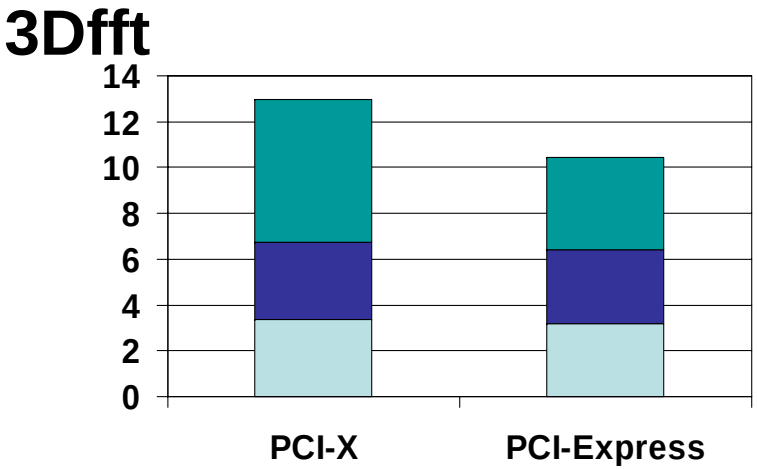


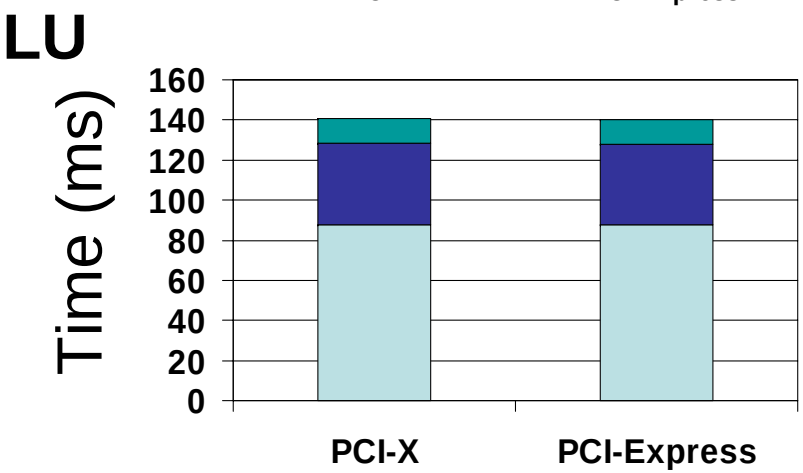


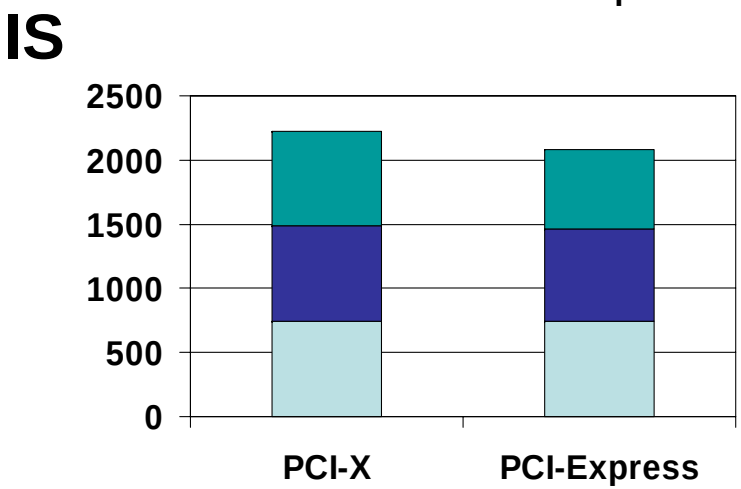
- 13% improvement for 3Dfft with PCI-Express using NEWGENDSM

Application Timing Breakdown









- NEWGENDSM
- Reduction in page and wait times

Conclusions and Future Work

- Examined the impact of PCI-Express on DSM protocols
- Up to 13% improvement in application performance with NEWGENDSM at four nodes
- Plan to study the impact of PCI-Express on DSM systems for a larger cluster size

Web Pointers

NBC-LAB

Group Homepage: http://nowlab.cis.ohio-state.edu

Emails: {noronha, panda}@cse.ohio-state.edu

- Backup Slides

Application Characteristics

- Micro-benchmarks (modified from TreadMarks suite)
  - Page →
    - Average time to fetch a page from a home node when a number of nodes are accessing it
- Applications from SPLASH-2 suite (Barnes, 3Dfft, LU, IS)
  - Barnes
    - N-body simulation using the hierachial Barnes-Hut Method
    - Contains two main arrays one for bodies and the other the cells
    - Sharing patterns are irregular and true
    - Large diff traffic exchanged at barriers
  - 3DFFT
    - Performs three dimensional Fast Fourier Transform
    - Exchanges large volume of messages per unit time
    - Uses a large number of locks and barriers for synchromization
  - LU
    - Factors a dense matrix into the product of a lower and upper triangular matrix
    - Uses blocking to exploit temporal locality on individual sub-matrix elements
    - Sends a large number of diffs
    - Exchanges maximum amount of data traffic
  - IS
    - Implements bucket sort
    - Global array containing the buckets
    - Local array which the local node uses to sort its data
    - After each iteration, each node places its data in the global array and copies the data relevant to it into its local array
    - Large number of diffs are exchanged at intervals

Application Sizes

Application	Parameter	Size
Barnes	Bodies	4096
3Dfft	Grid size	4
LU	Matrix Dimension	16
IS	Number of Keys	2^8