NETWORK-BASED COMPUTING LABORATORY

Memory Scalability Evaluation of the Next-Generation Intel Bensley Platform with InfiniBand

Matthew Koop, Wei Huang, Ahbinav Vishnu, Dhabaleswar K. Panda

Network-Based Computing Laboratory
Department of Computer Science & Engineering
The Ohio State University

OHIO STATE

Introduction

- Computer systems have increased significantly in processing capability over the last few years in various ways
  - Multi-core architectures are becoming more prevalent
  - High-speed I/O interfaces, such as PCI-Express have enabled high-speed interconnects such as InfiniBand to deliver higher performance

- The area that has improved the least during this time is the memory controller

Traditional Memory Design

- Traditional memory controller design has limited the number of DIMMs per memory channel as signal rates have increased
- Due to high pin count (240) required for each channel, adding additional channels is costly
- End result is equal or lesser memory capacity in recent years

Fully-Buffered DIMMs (FB-DIMMs)

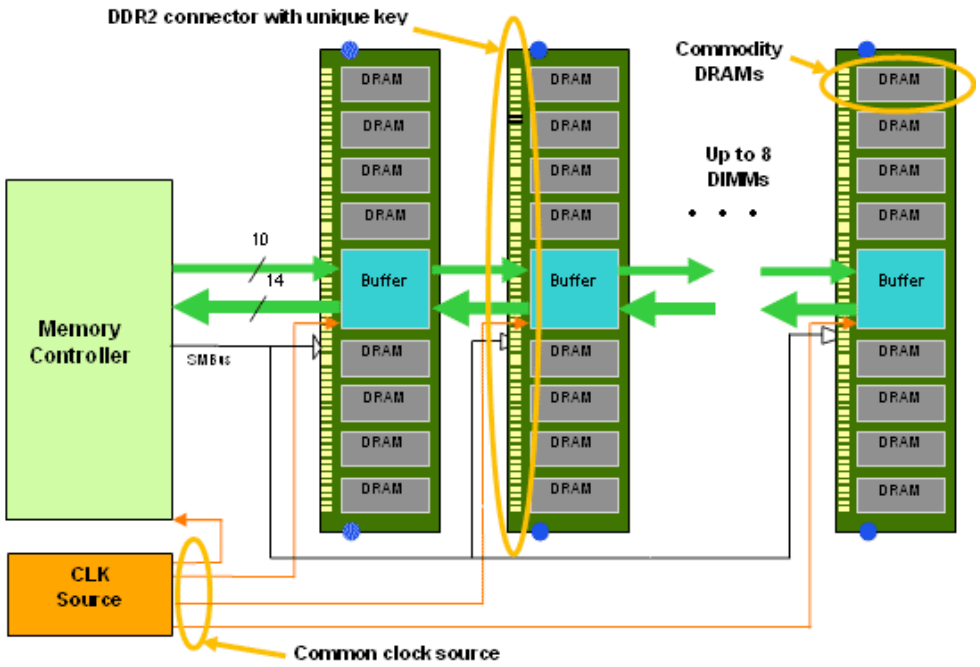


Image courtesy of Intel Corporation

- FB-DIMM uses serial lanes with a buffer on each chip to eliminate this tradeoff
- Each channel requires only 69 pins
- Using the buffer allows larger numbers of DIMMs per channel as well as increased parallelism

Evaluation

- With multi-core systems coming, a scalable memory subsystem is increasingly important
- Our goal is to compare FB-DIMM against a traditional design and evaluate the scalability

- Evaluation Process
  - Test memory subsystem on a single node
  - Evaluate network-level performance with two InfiniBand Host Channel Adapters (HCAs)

Outline

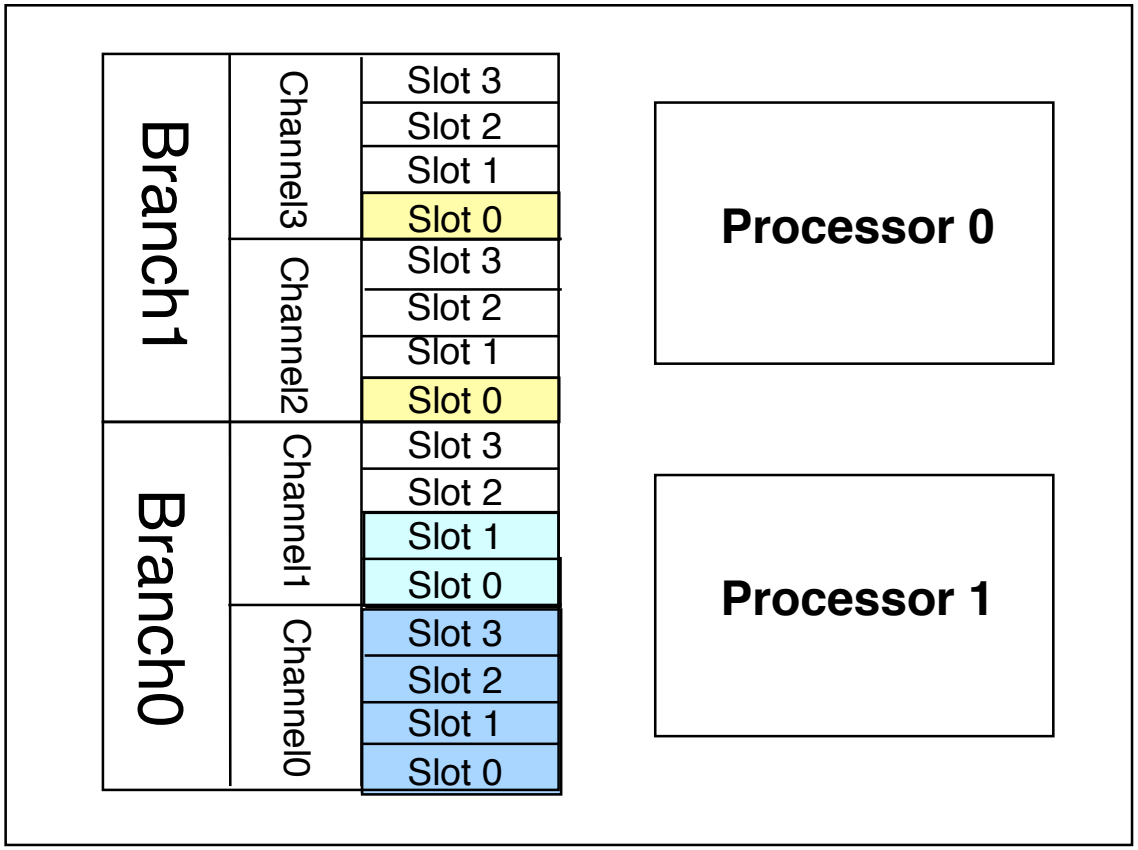
- Introduction & Goals
- Memory Subsystem Evaluation
  - Experimental testbed
  - Latency and throughput
- Network results
- Conclusions and Future work

Evaluation Testbed

- Intel “Bensley” system
  - Two 3.2 GHz dual-core Intel Xeon “Dempsey” processors
  - FB-DIMM-based memory subsystem
- Intel Lindenhurst system
  - Two 3.4 GHz Intel Xeon processors
  - Traditional memory subsystem (2 channels)

- Both contain:
  - 2 8x PCI-Express slots
  - DDR2 533-based memory
  - 2 dual-port Mellanox MT25208 InfiniBand HCAs

Bensley Memory Configurations



- The standard allows up to 6 channels with 8 DIMMs/channel for 192GB
- Our systems have 4 channels, each with 4 DIMM slots
- To fill 4 DIMM slots we have 3 combinations

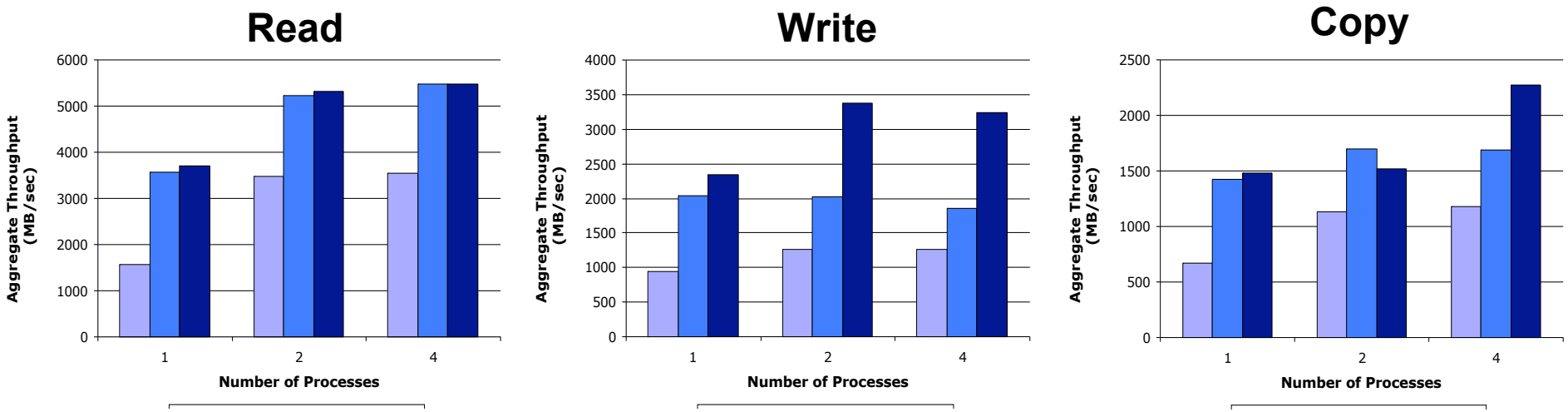
Subsystem Evaluation Tool

lmbench 3.0-a5: Open-source benchmark suite for evaluating system-level performance

- Latency
  - Memory read latency
- Throughput
  - Memory read benchmark
  - Memory write benchmark
  - Memory copy benchmark

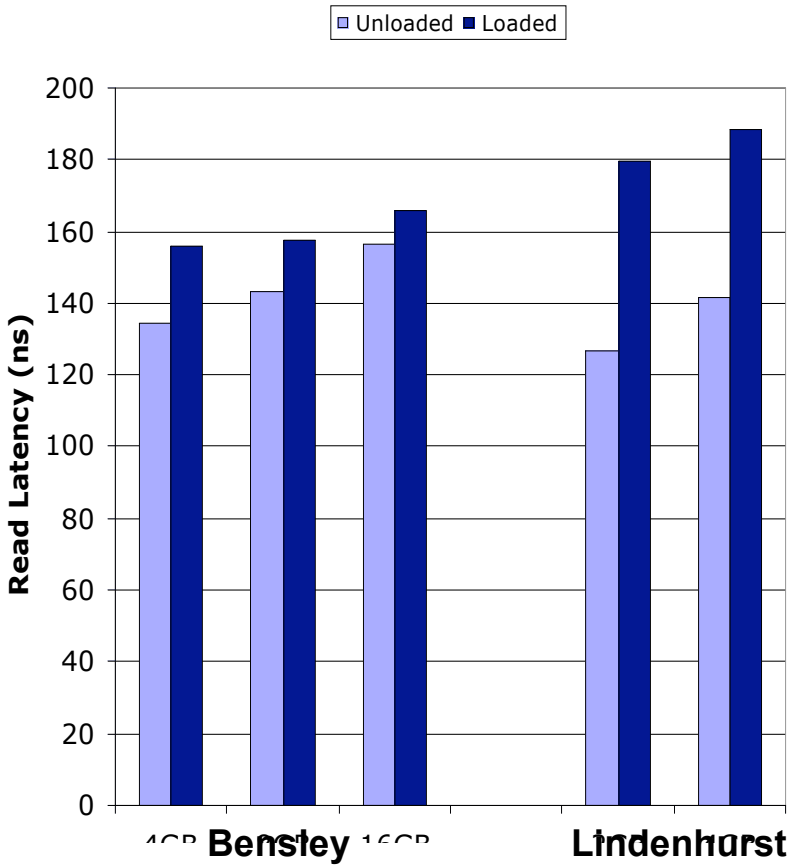
Aggregate performance is obtained by running multiple long-running processes and reporting the sum of averages

Bensley Memory Throughput



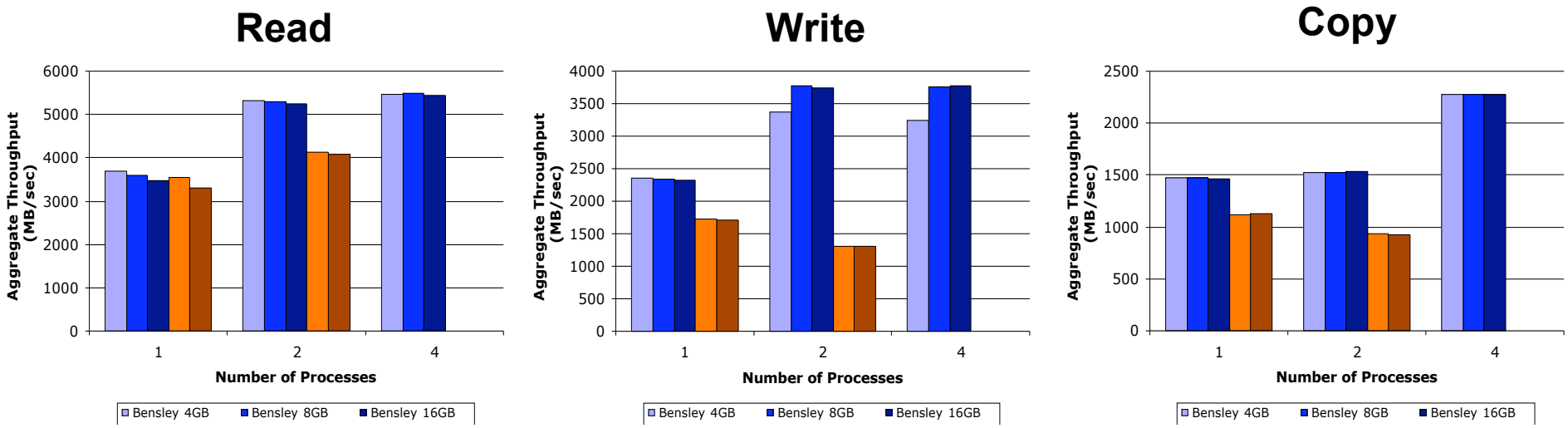
- To study the impact of additional channels we evaluated using 1, 2, and 4 channels
- Throughput increases significantly from one to two channels in all operations

Access Latency Comparison



- Comparison when *unloaded* and *loaded*
- *Loaded* is when a memory read throughput test is run in the background while the latency test is runnning
- From unloaded to loaded latency:
  - Lindenhurst: 40% increase
  - Bensley: 10% increase

Memory Throughput Comparison



- Comparison of Lindenhurst and Bensley platforms with increasing memory size
- Performance increases with two concurrent read or write operations on the Bensley platform

Outline

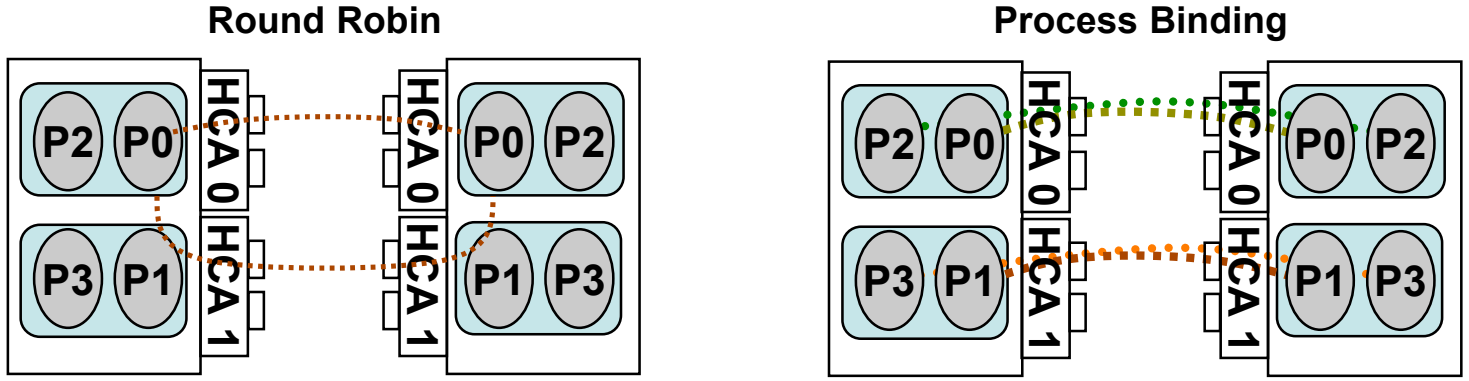
- Introduction & Goals
- Memory Subsystem Evaluation
  - Experimental testbed
  - Latency and throughput
- Network results
- Conclusions and Future work

OSU MPI over InfiniBand

- Open Source High Performance Implementations
  - MPI-1 (MVAPICH)
  - MPI-2 (MVAPICH2)
- Has enabled a large number of production IB clusters all over the world to take advantage of InfiniBand
  - Largest being Sandia Thunderbird Cluster (4512 nodes with 9024 processors)
- Have been directly downloaded and used by more than 395 organizations worldwide (in 30 countries)
  - Time tested and stable code base with novel features
- Available in software stack distributions of many vendors
- Available in the OpenFabrics(OpenIB) Gen2 stack and OFED
- More details at

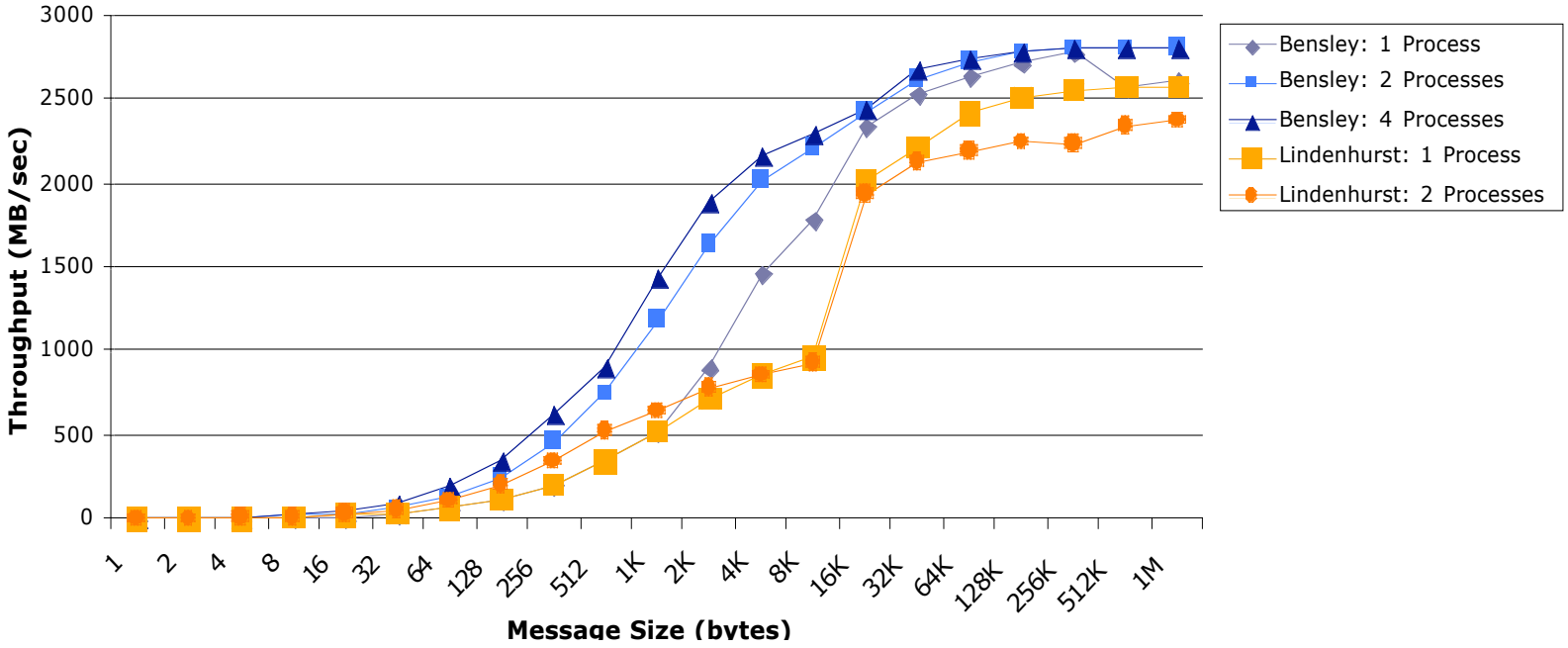
  http://nowlab.cse.ohio-state.edu/projects/mpi-iba/

Experimental Setup



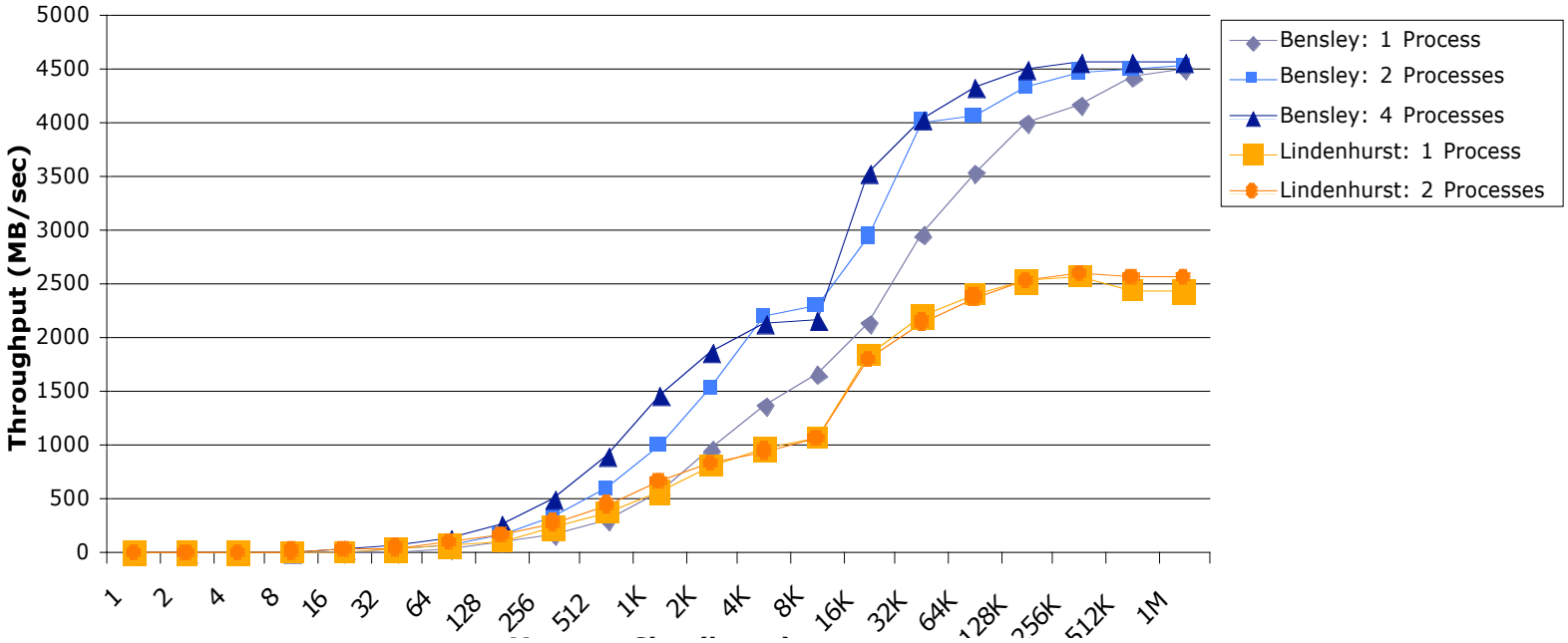
- Evaluation is with two InfiniBand DDR HCAs, which uses the “multi-rail” feature of MVAPICH
- Results with one process use both rails in a *round-robin* pattern
- 2 and 4 process pair results are done using a *process binding* assignment

Uni-Directional Bandwidth



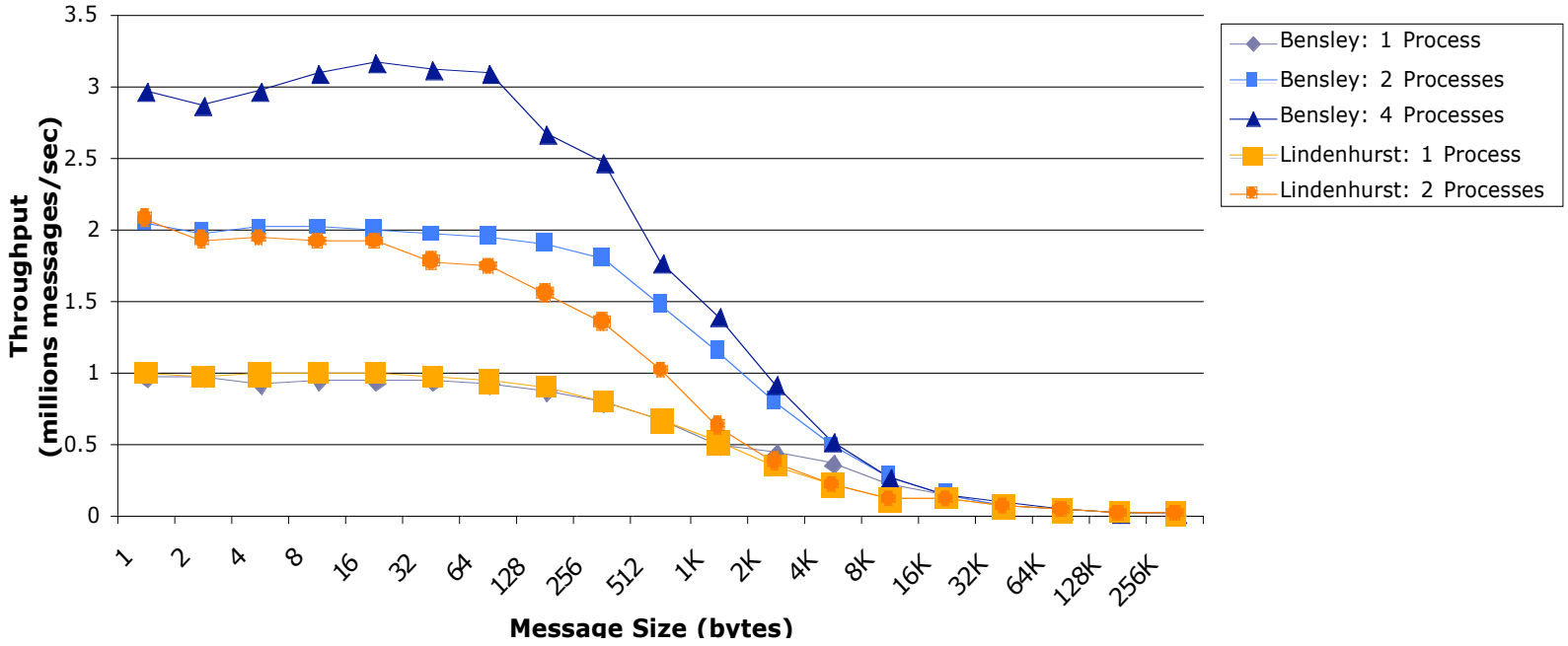
- Comparison of Lindenhurst and Bensley with dual DDR HCAs
- Due to higher memory copy bandwidth, Bensley signficantly outperforms Lindenhurst for the medium-sized messages

Bi-Directional Bandwidth



- At 1K improvement:
  - Lindenhurst: 1 to 2 processes:15%
  - Bensley: 1 to 2 processes: 75%, 2 to 4: 45%
- Lindenhurst peak bi-directional bandwidth is only 100 MB/sec greater than uni-directional

Messaging Rate



- For very small messages, both show similar performance
- At 512 bytes: Lindenhurst 2 process case is only 52% higher than 1 process, Bensley still shows 100% improvement

Outline

- Introduction & Goals
- Memory Subsystem Evaluation
  - Experimental testbed
  - Latency and throughput
- Network results
- Conclusions and Future work

Conclusions and Future Work

- Performed detailed analysis of the memory subsystem scalability of Bensley and Lindenhurst
- Bensley shows significant advantage in scalable throughput and capacity in all measures tested

- Future work:
  - Profile real-world applications on a larger cluster and observe the effects of contention in multi-core architectures
  - Expand evaluation to include NUMA-based architectures

Acknowledgements

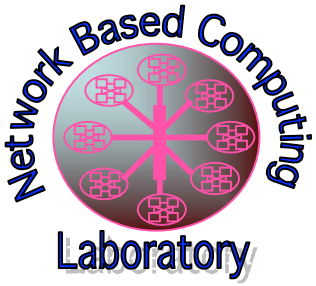
Our research is supported by the following organizations

- Current Funding support by

- Current Equipment support by


Web Pointers

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



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

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