Designing High Performance Communication Middleware with Emerging Multi-core Architectures

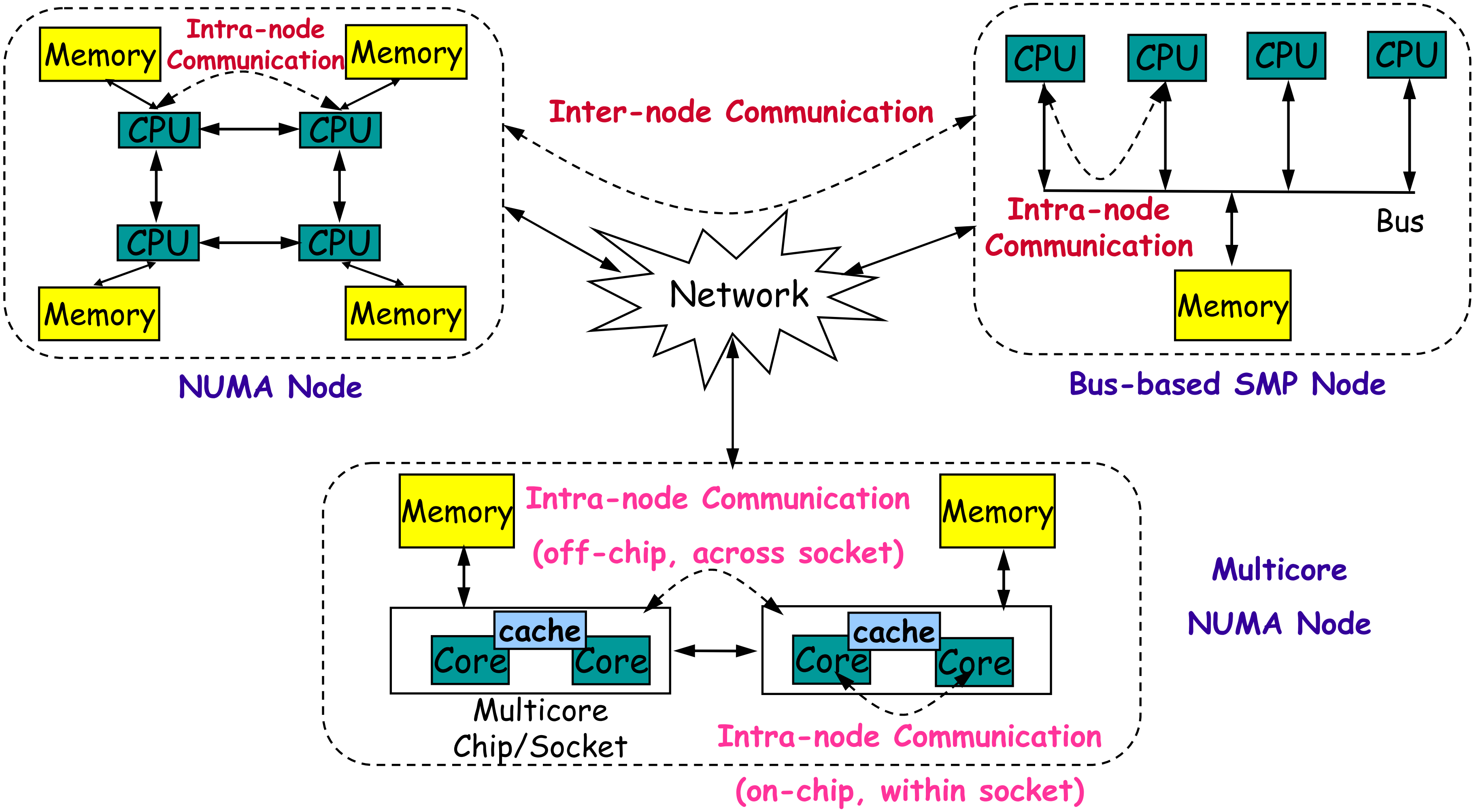
Dhabaleswar K. (DK) Panda
Department of Computer Science and Engg.
The Ohio State University

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

On-chip Networks – The wave of the Future

- Node architectures are changing rapidly
  - Especially with multi-core processors
- Introducing new memory-hierarchy
  - Core-core communication
    - Within a socket (on-chip)
    - Across a socket (off-chip)
  - Inter-node communication (off-chip) bandwidth
- New challenges in designing on-chip networks
- Also need better software (data placement or optimization) to extract the maximum performance

Emerging Clusters with Multi-Level Memory Hierarchy and Different Communication Bandwidth



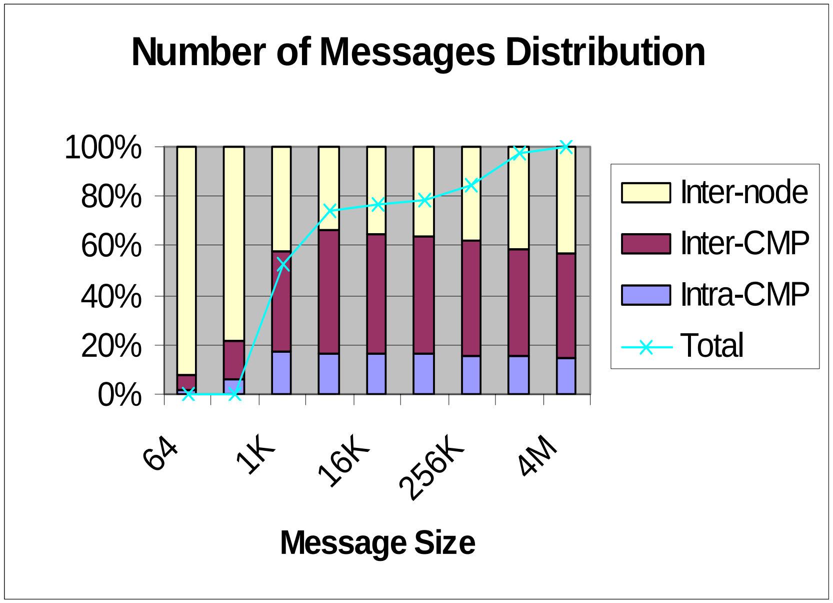
Designing Efficient Communication Middleware

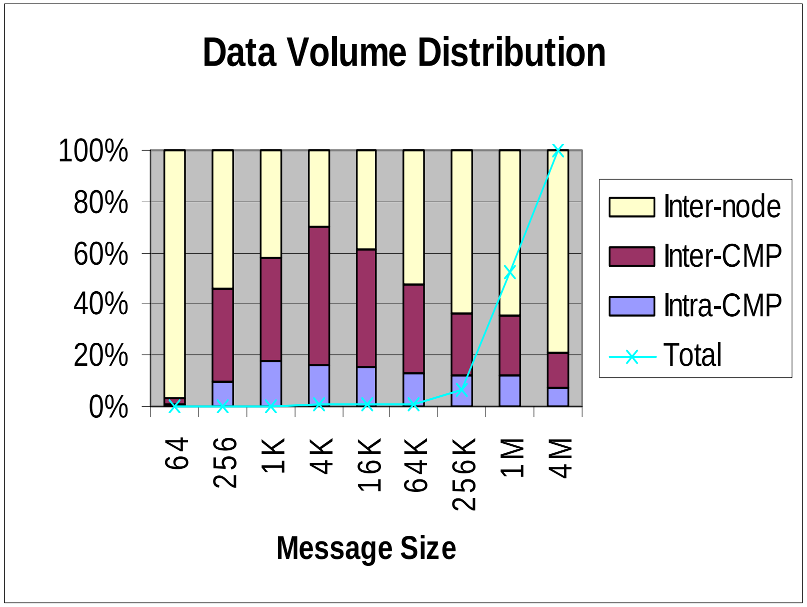
- Criticality of message distribution across the three levels?
- Can we optimize on-chip communication by exploiting multi-core architecture?
- Benefits to the overall application
- A study on
  - MPI Library and applications
  - Intel and Opteron Multi-core systems
  - InfiniBand as Cluster Interconnect

Enhancing MPI Library for Efficient Intra-node (within socket and across socket) Communication

- High Performance MPI Library for InfiniBand Clusters
  - MVAPICH (MPI-1) and MVAPICH (MPI-2)
  - Used by more than 435 organizations in 30 countries
  - Empowering many TOP500 clusters
  - Available with software stacks of many InfiniBand and server vendors including the OpenIB and Open Fabrics Enterprise Distribution (OFED)
  - http://nowlab.cse.ohio-state.edu/projects/mpi-iba/
- Already has good support for intra-node MPI point-to-point communication (with copying) over shared memory
- Recently designed an efficient and scalable scheme with associated data structures for emerging multicore and NUMA systems

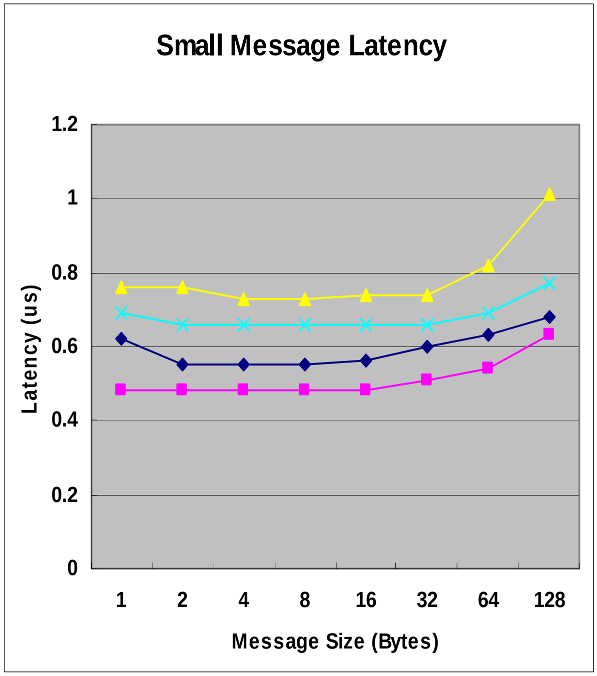
Cumulative Message Distribution

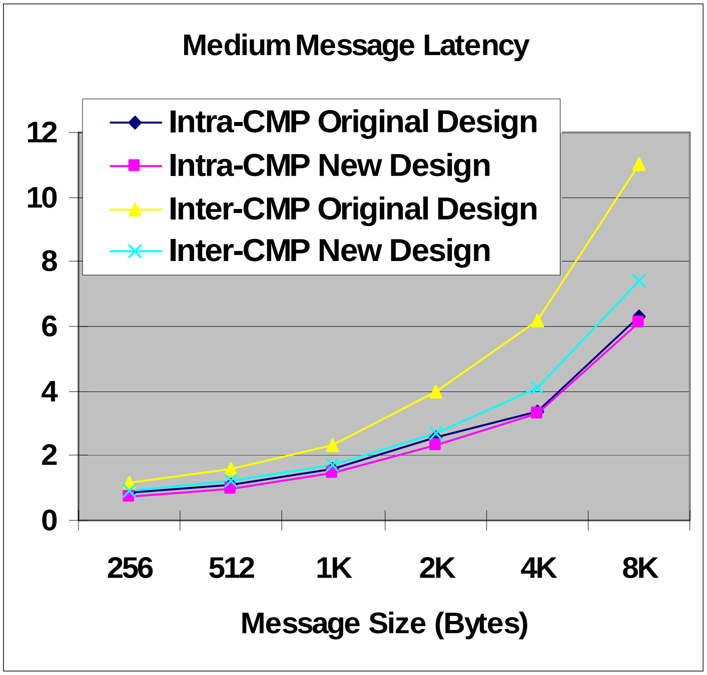


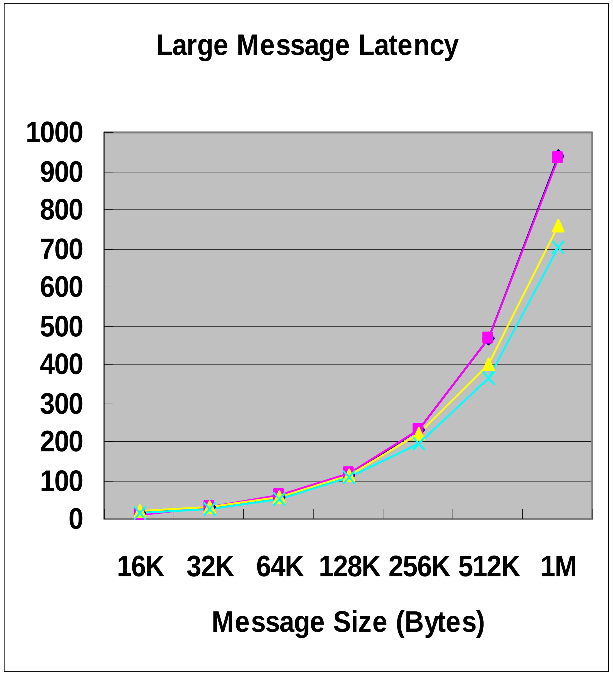


- High Performance Linpack (HPL)
- Dual dual-core Opteron cluster, 16 nodes, 64 processes

Enhanced Message Passing Design for Improving On-Chip and Off-Chip Communication



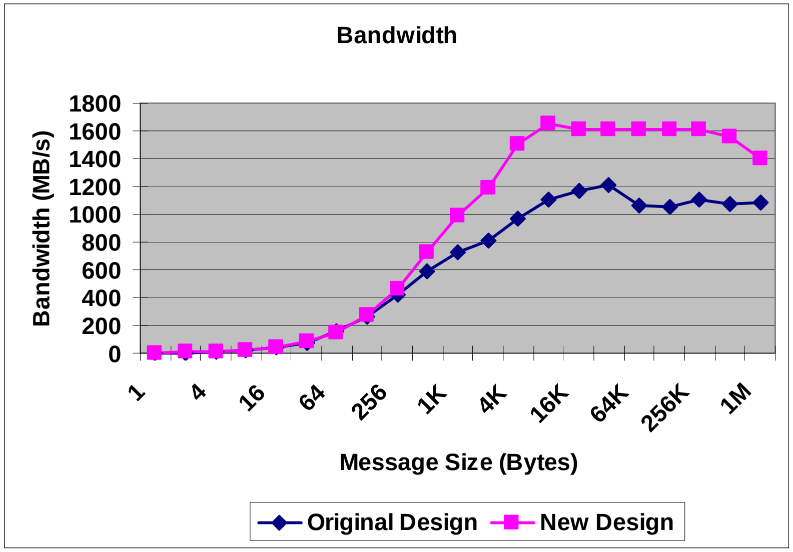


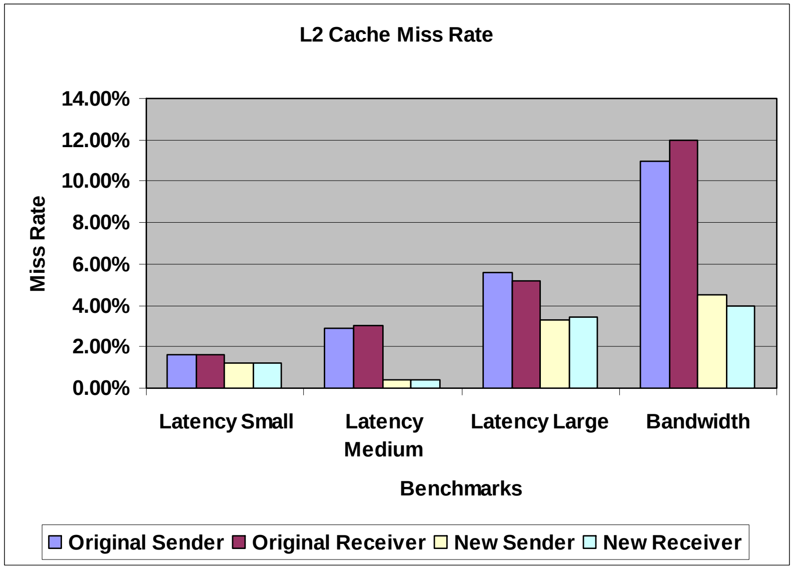


- Dual dual-core AMD Opteron processors, 2.0GHz, 1MB L2 cache
- The new design improves
  - inter-CMP latency for all the messages
  - intra-CMP latency for small and medium messages

L. Chai, A. Hartono and D. K. Panda, Designing High Performance and Scalable MPI Intra-node Communication Support for Clusters, Cluster '06

Performance Improvement in Bandwidth and Cache Miss Rates

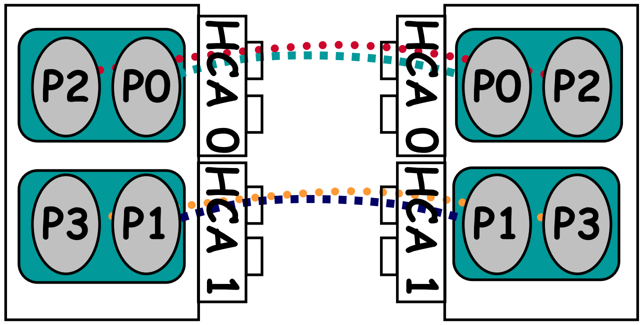




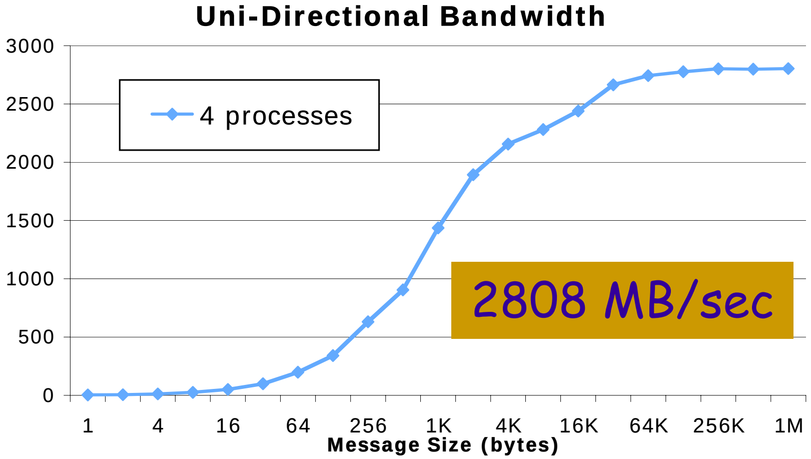
- Bandwidth is improved by up to 50%
- Benefits mainly come from the reduced L2 cache miss rate

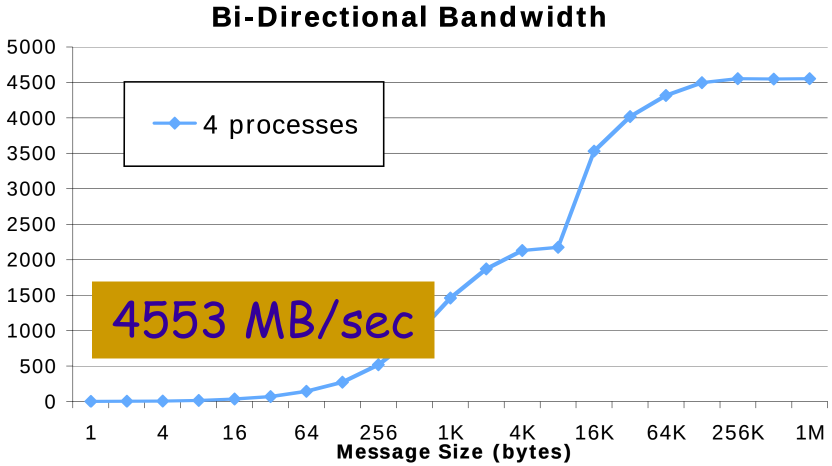
- High performance implementations of programming models needed
- End applications need to optimize data placements for good performance

Inter-node Communication Performance



4-processes on each node concurrently
communicating over Dual-rail InfiniBand DDR (Mellanox)

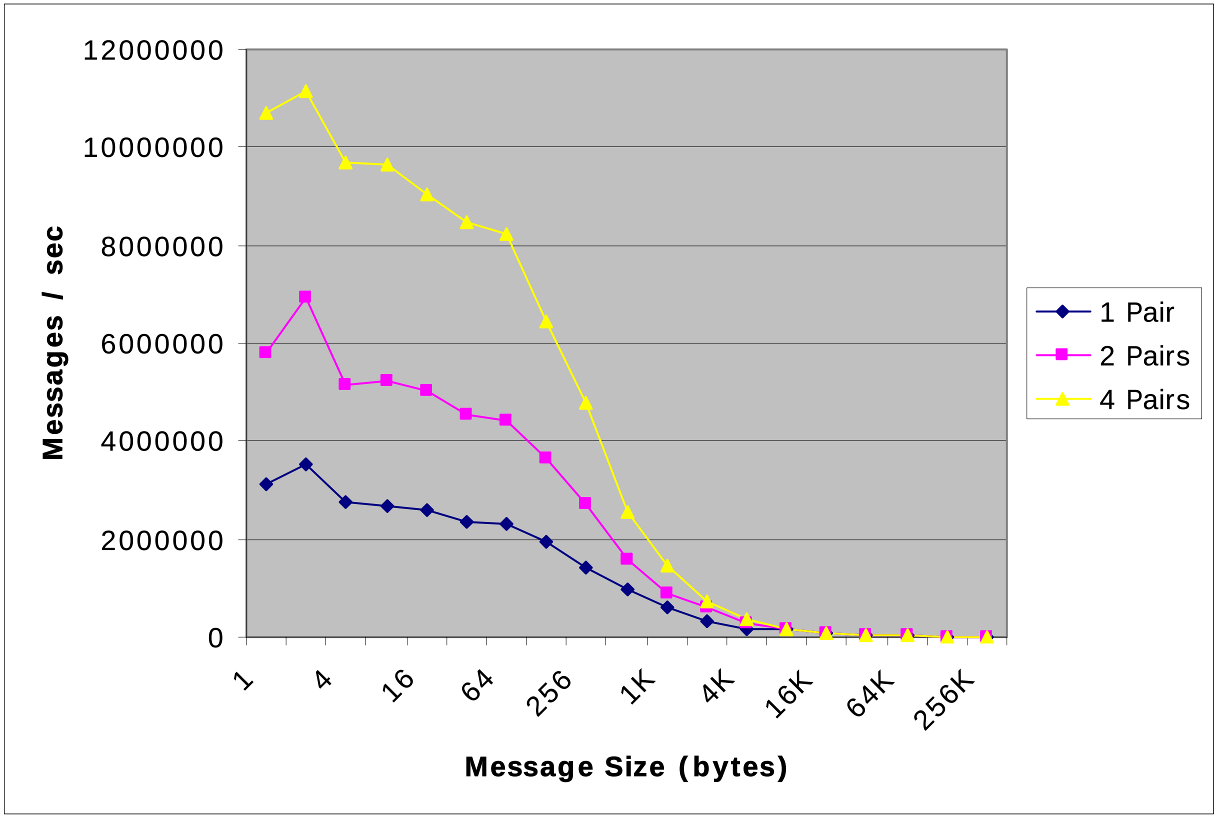




M. J. Koop, W. Huang, A. Vishnu and D. K. Panda, Memory Scalability Evaluation of Next Generation Intel Bensley Platform with InfiniBand, Hot Interconnect Symposium (Aug. 2006).

Inter-node Messaging Rate

- Design based on MVAPICH 0.9.8
- Preliminary performance results
  - Dual dual-core Intel Woodcrest
  - Single DDR card



Open Issues for Enhancing Communication Performance with Multicore and On-Chip/Off-Chip Interconnects

- Can one or more cores be dedicated for communication?
  - to achieve better overlap of computation and communication
- Can one-sided and multi-threading be well supported on multicore systems?
- Can collective communication (broadcast, multicast, all-to-all, all-reduce, etc.) be implemented efficiently and in a non-blocking manner
  - To have better overlap of computation with collective communication
- Can we aim for fine-grain synchronization?
- Will depend on the capability and features of both on-chip and off-chip networks