

•
•
•
•
•
•
•
•
•
•
•
•
•
•
•

Challenges for Future Interconnection Networks: Power, Reliability and Performance Scalability?

Panel at HOTI14



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>





Power Consumption



- Will power consumption mainly be the problem of the processor and memory system designers or will it mainly be the network designer's problem?
 - Dell PowerEdge 1950 server with dual dual-core Intel processors, memory, disk, etc.
 - 670 watts
 - InfiniBand DDR (20 Gbps) NIC adapters
 - 3.0-3.5 watt (single-port)
 - 7.8-10.3 watt (dual-port)
 - InfiniBand DDR switch
 - 34 watt for 24 ports 4X DDR (20 Gbps) or 8 ports 12X DDR (60 Gbps)
 - 2500 watt for 288 port -> 8.6 watt/port
 - Myrinet 10GigE PCI-Express NIC
 - 8.3 watt
 - Around 20-30 watt per node for network (NIC + switch port)
 - Power problem is more critical for processor and memory system designers
- 





Reducing Clock Frequency for Network



- What's wrong with simply turning down the clock frequency and going serial as a means of confronting the power problem in networks?
- Inter-node network power consumption is still not critical
- Modern networks actually use multiple parallel serial links
 - InfiniBand 4X using four 1X link (2.5 Gbps each)
 - InfiniBand 12X using 12 1X link (2.5 Gbps each)



Reliability

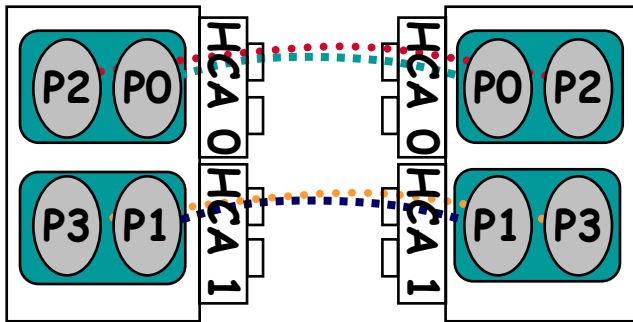
- What critical reliability assumptions we now take for granted are likely not to hold for future interconnection networks and what might be the impact?
- Reliability in a large-scale high-speed network is always a big concern
- InfiniBand provides **Reliable Connection (RC)** transport service where the HCA and switch provide support for CRC, packet drops, etc.
 - Delivers ~2.7 microsec MPI-level latency with this support
 - Also provide support for **Automatic Path Migration (APM)**
 - Overhead in establishing connections and scalability problem (20K-50K nodes)
- InfiniBand's **Unreliable Datagram (UD)** provides scalability but not reliability
- **Reliable Datagram (RD)** transport provides advantages from both sides
 - Not supported yet in current generation InfiniBand network
 - Needs to be supported
- Upper-level Software infrastructure is needed to handle
 - Checkpoint-restart, process migration, congestion management, ..

Factors in Determining Latency and Throughput

- Ultimately, what will determine latency and throughput performance: the interfaces to the network or the network fabric itself?
- Both (network interface and network fabric) are important
- Modern interfaces like PCI-Express and Hypertransport are helping tighter integration between memory and network interface
 - reducing latency and enhancing throughput

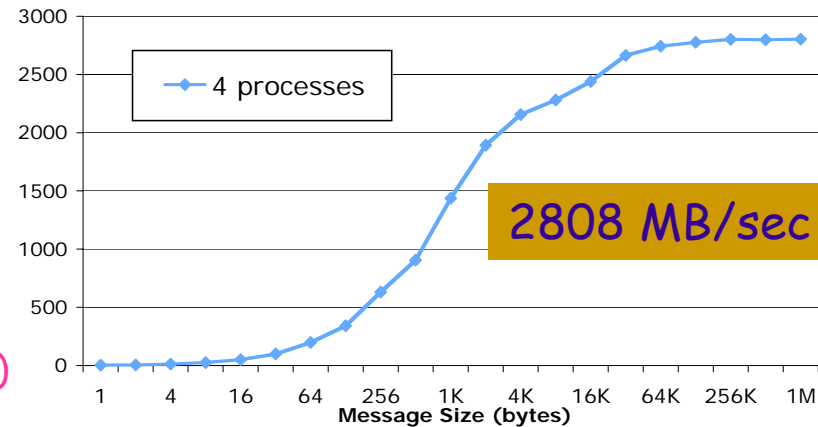
MPI over InfiniBand Performance

(Dual-core Intel Bensley Systems with PCI-Express and Dual-Rail DDR InfiniBand)

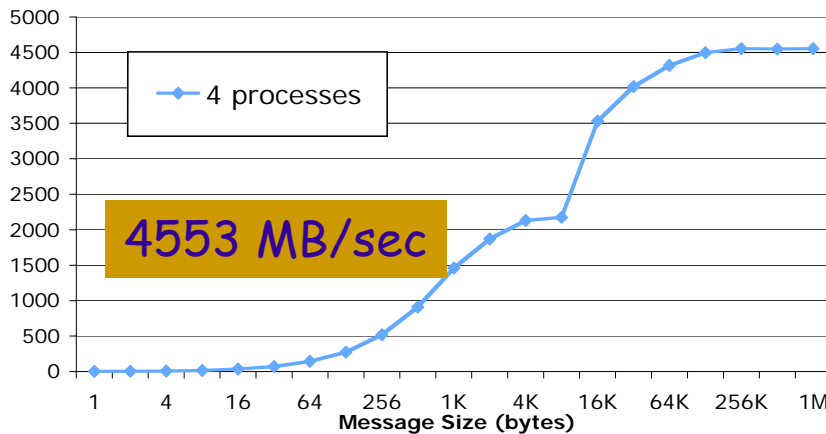


4-processes on each node concurrently communicating over Dual-rail InfiniBand DDR (Mellanox)

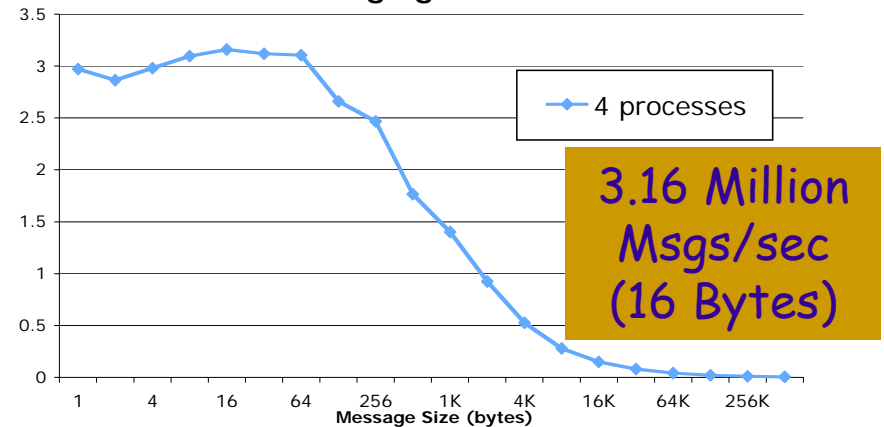
Uni-Directional Bandwidth



Bi-Directional Bandwidth



Messaging Rate



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



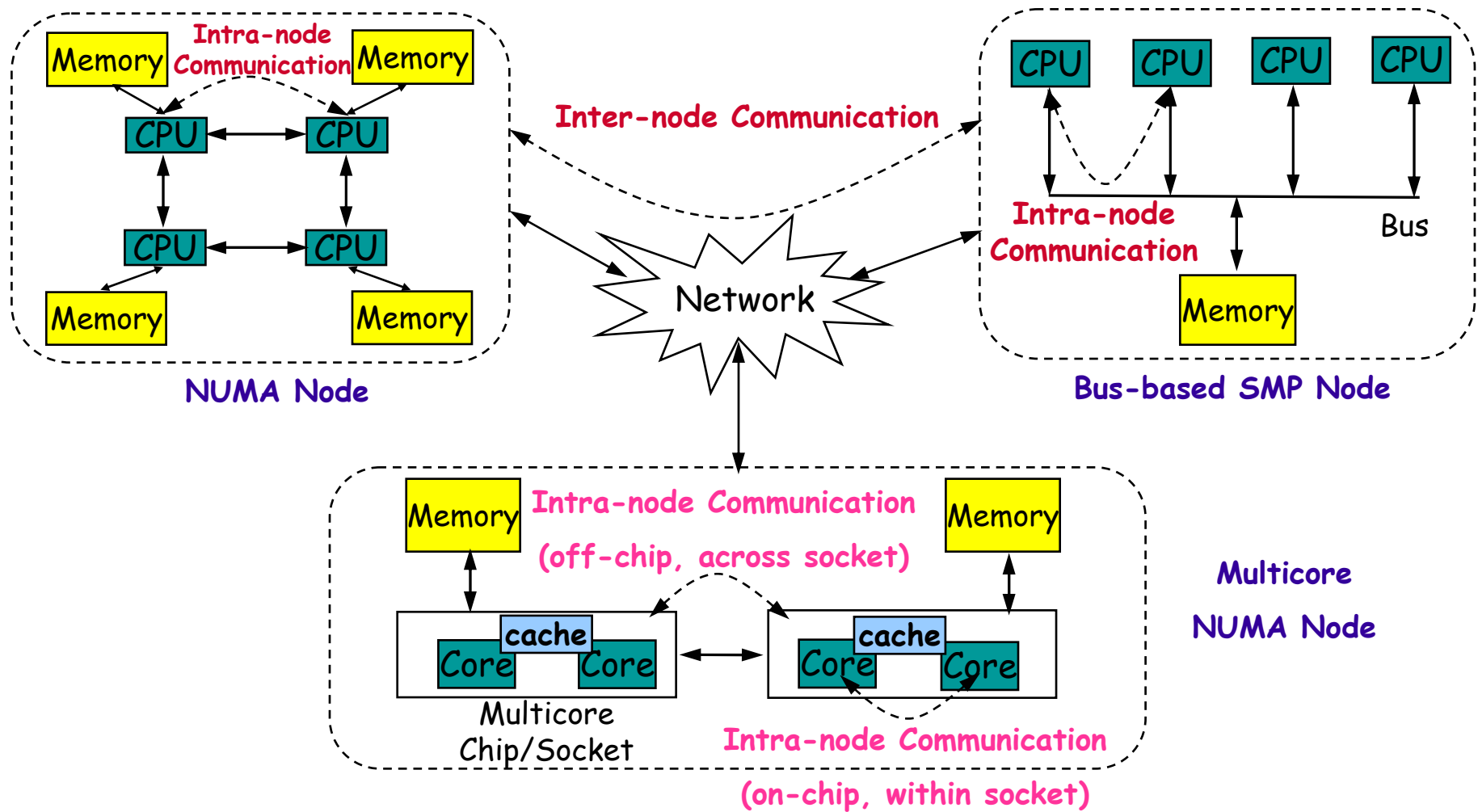
On-chip Networks - The wave of the Future



- 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



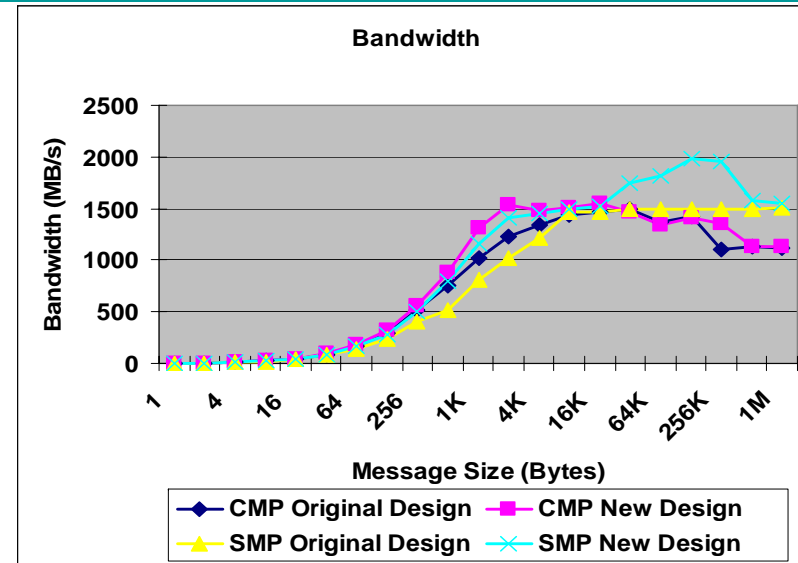
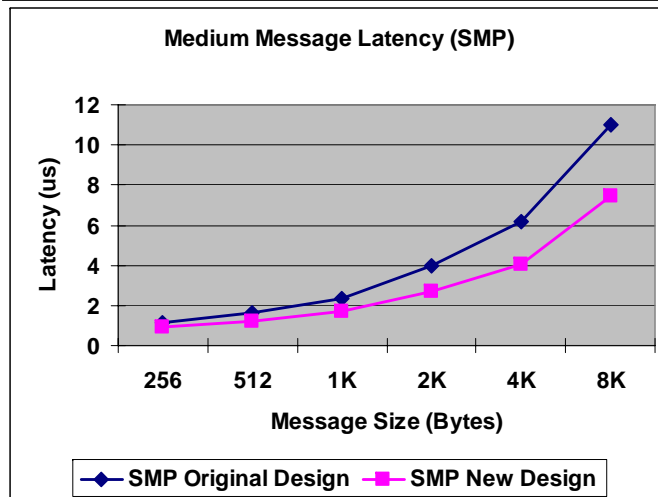
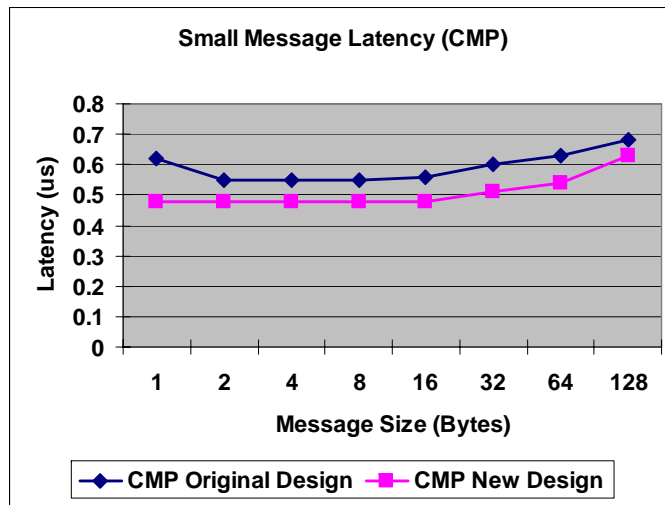
Emerging Clusters with Multi-Level Memory Hierarchy and Different Communication Bandwidth



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 395 organizations in 30 countries
 - Empowering many TOP500 clusters including the 9K processor Sandia Thunderbird cluster (6th)
 - 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

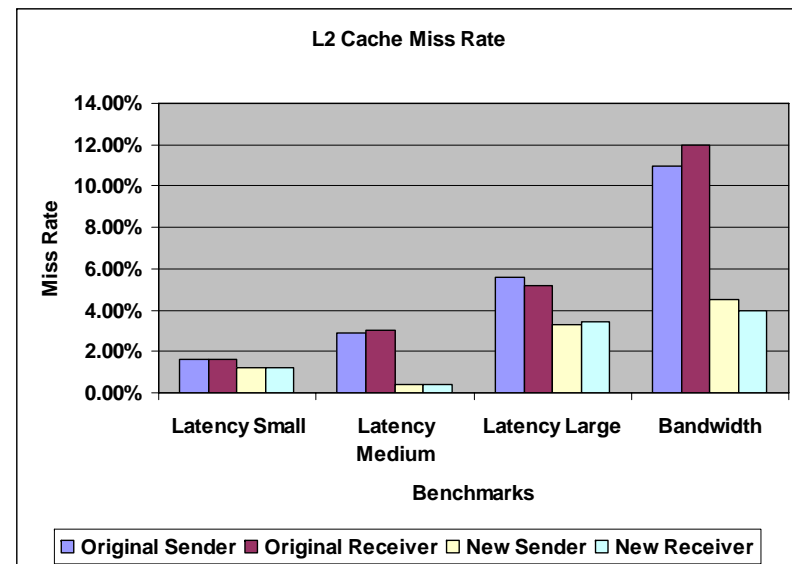
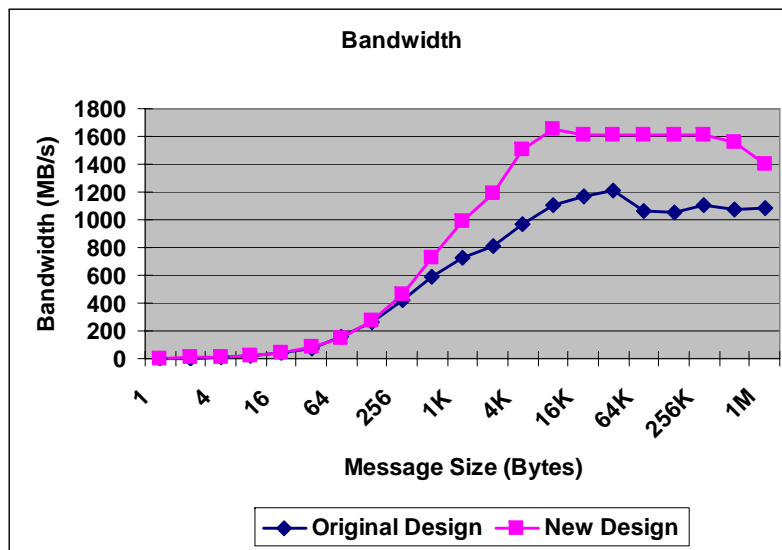
Enhanced MPI Point-to-point Communication Performance on Multicore NUMA Cluster



- CMP (on-chip, within socket) latency is improved by up to 12%
- SMP (off-chip, same node) latency is improved by up to 30%
- Bandwidth is improved by up to 25%

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

Performance on NUMA Cluster



- 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

•
•
•

Open Issues for Enhancing Communication Performance with Next Generation Networks and Multicore Computing Systems?

- 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