





## Some of David's questions and Short answers





- We do not need to develop new parallel programming paradigms since the old shared-memory and message-passing paradigms will be good enough
  - We have enough programming models already ...
  - Do we need more?
  - How well the current programming models work on multicore systems?
  - Do we need better implementations of these models first?

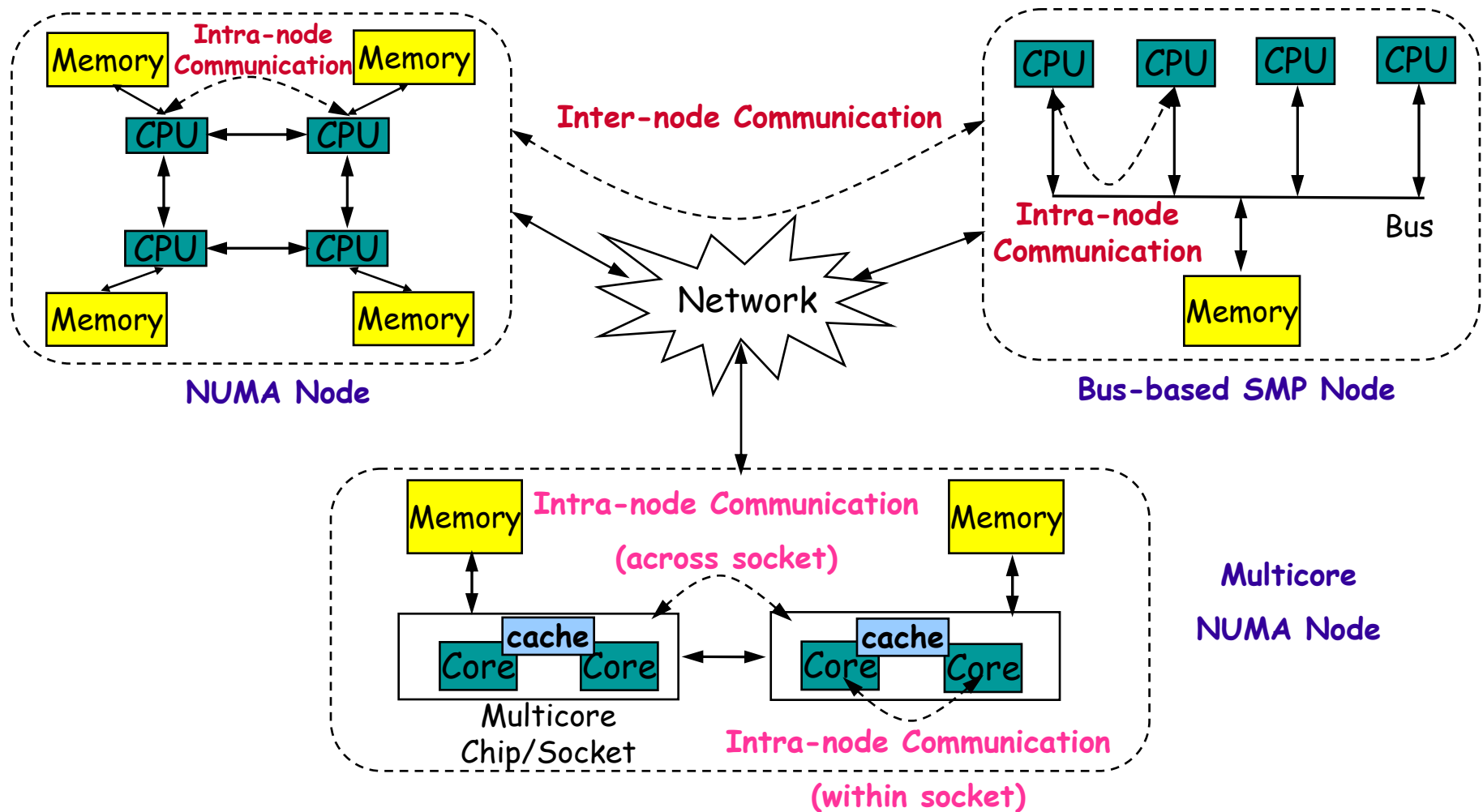


## Some of David's questions and Short answers



- Multicore systems are irrelevant to application software developers because they will be used only for increasing overall system throughput, not for accelerating individual application programs
  - Not really
  - Depends on the **memory-hierarchy** provided by the multicore systems
    - Core-core communication bandwidth
      - Within a socket
      - Across a socket
    - In addition to inter-node communication bandwidth
  - May need better data placement or optimization to extract the maximum performance
- 
- 

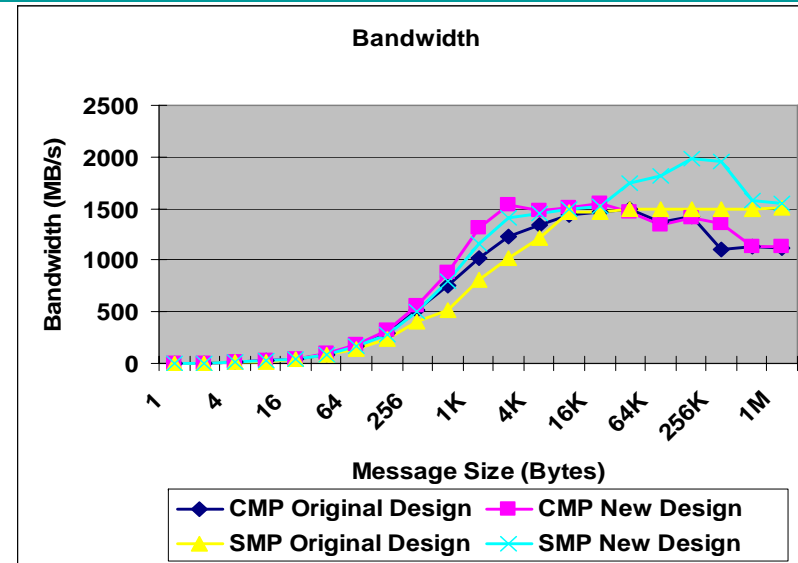
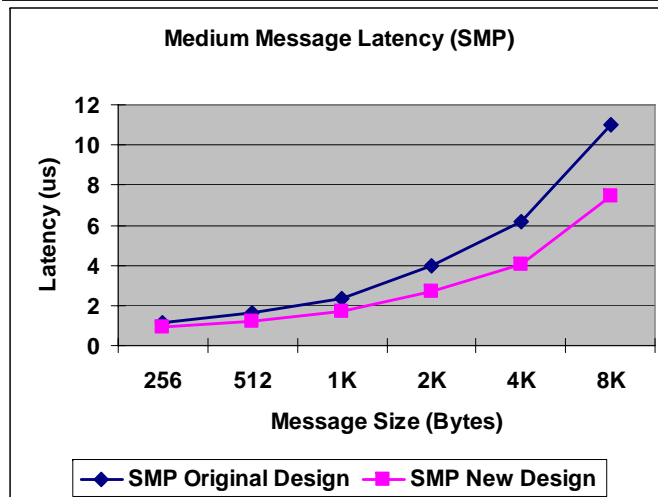
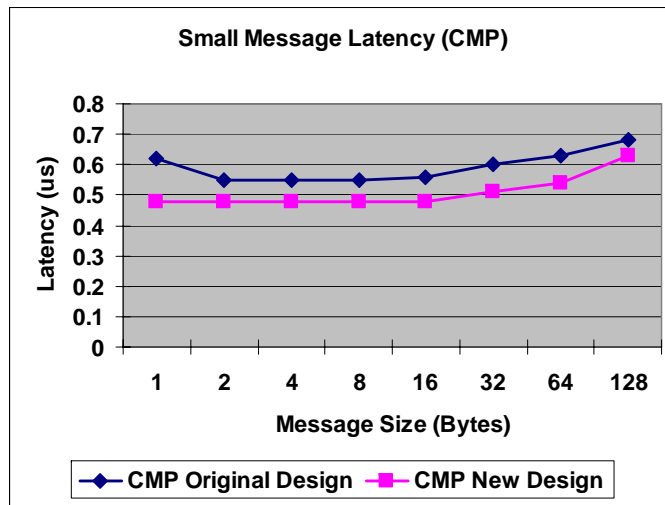
# Emerging Clusters with Multi-Level Memory Hierarchy and Different Communication Bandwidth



## 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 380 organizations in 30 countries
  - Empowering many TOP500 clusters including the 8K processor Sandia Thunderbird cluster (6<sup>th</sup>)
  - 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/mqi-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

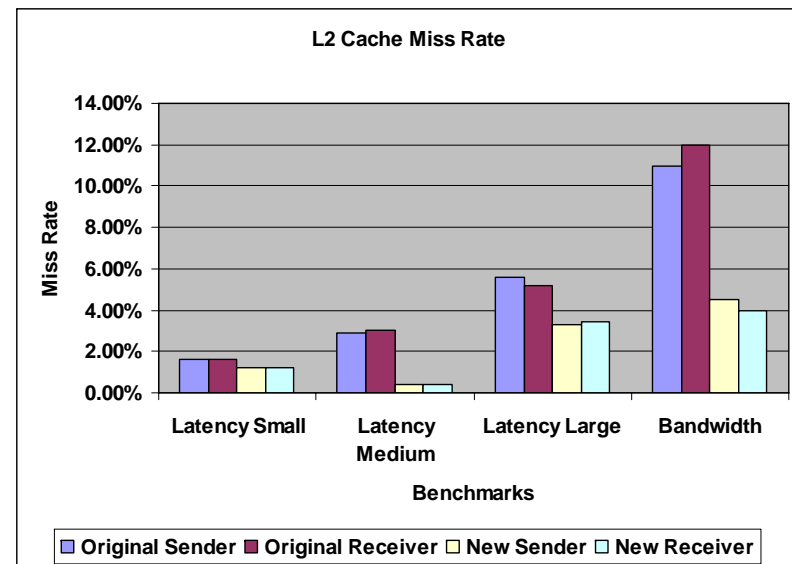
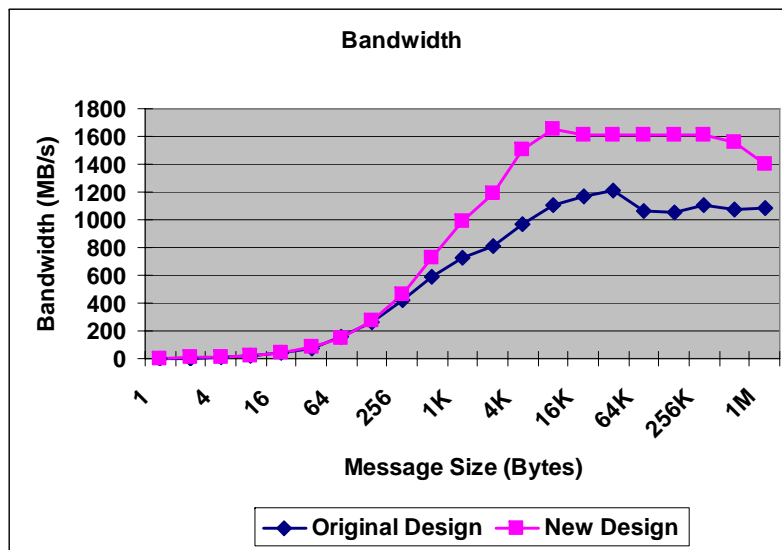
# Enhanced MPI Point-to-point Communication Performance on Multicore NUMA Cluster



- CMP (within socket) latency is improved by up to 12%
- SMP 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 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?



# Designing New Generation Systems with Multicore

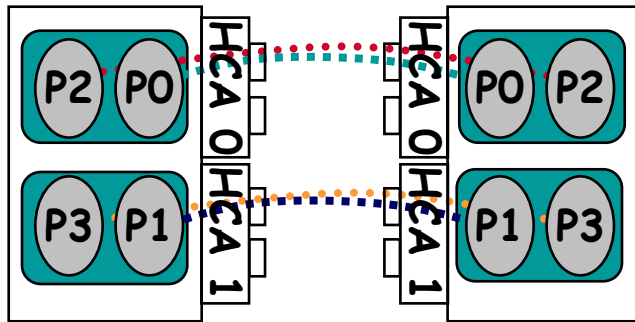


- Cost of high performance interconnect is a major factor in the overall cost
- Designing 1024 processors system
  - Dual processor node: 512 NICs and 512-ports
  - Quad processor node: 256 NICs and 256-ports
  - Dual dual-core processors node: 128 NICs and 128-ports
  - Quad dual-core processors node: 64 NICs and 64-ports
  - Quad quad-core processors node: 32 NICs and 32-ports
- Will see multiple trends
  - Cost-effective multicore systems with less NICs and ports
    - applications requiring low communication performance
  - Multicore systems with multiple rails (NICs per node)
    - Significantly higher performance
    - Fault-tolerance in network fabric
    - Separate rails for communication and I/O



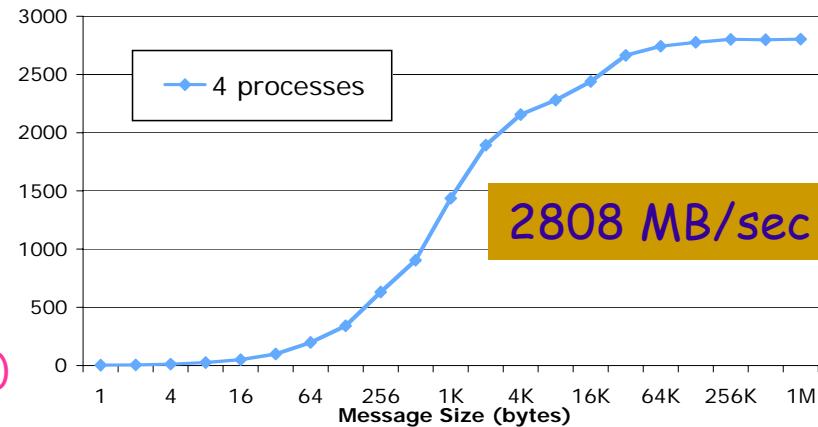
# MPI over InfiniBand Performance

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

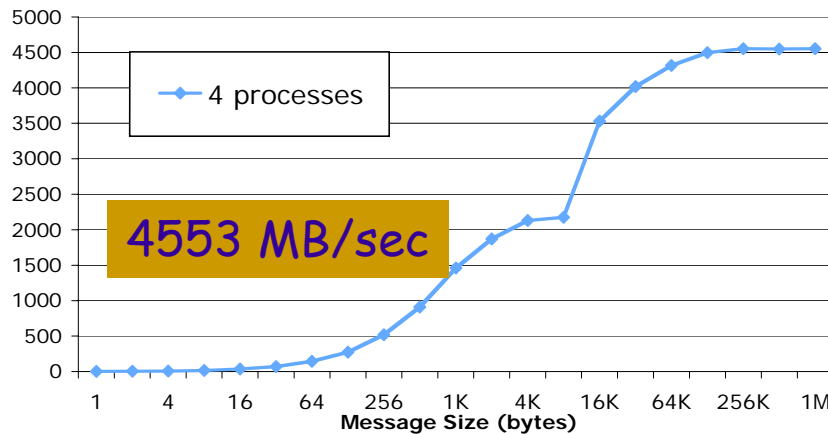


4-processes on each node concurrently communicating over Dual-rail InfiniBand DDR (Mellanox)

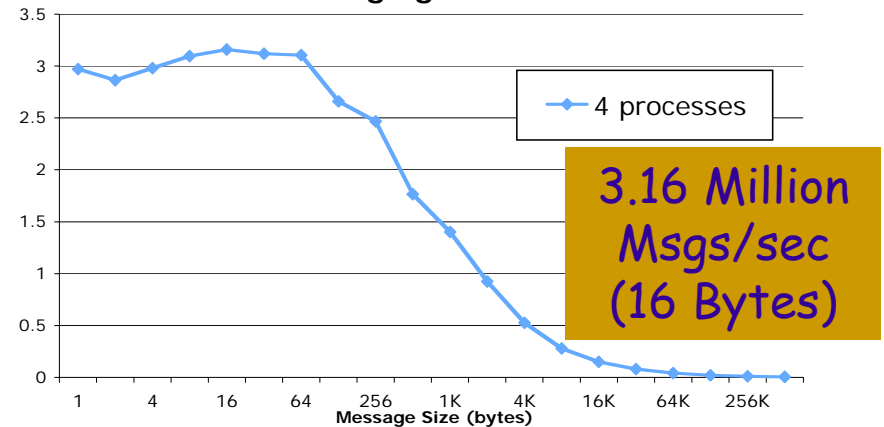
Uni-Directional Bandwidth



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).



# Conclusions



- A new era of parallel processing and cluster computing with multicore processors
- Challenges in optimizing existing programming models and applications to match with the memory hierarchy and architecture of multicore systems
- Provides new opportunities
  - To enhance communication performance
  - Trends for designing next generation clusters with low-cost, performance and fault-tolerance

