# Network-Centric Designs for High-End I/O and File Systems

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

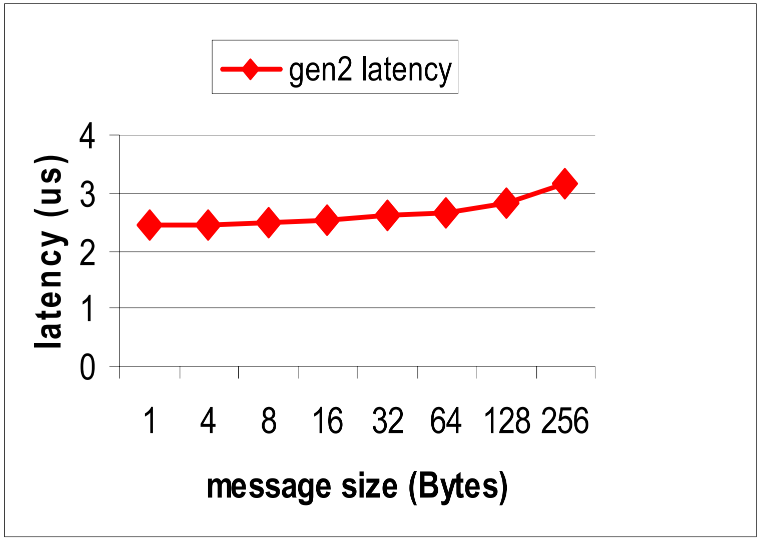
E-mail: panda@cse.ohio-state.edu
http://www.cse.ohio-state.edu/~panda

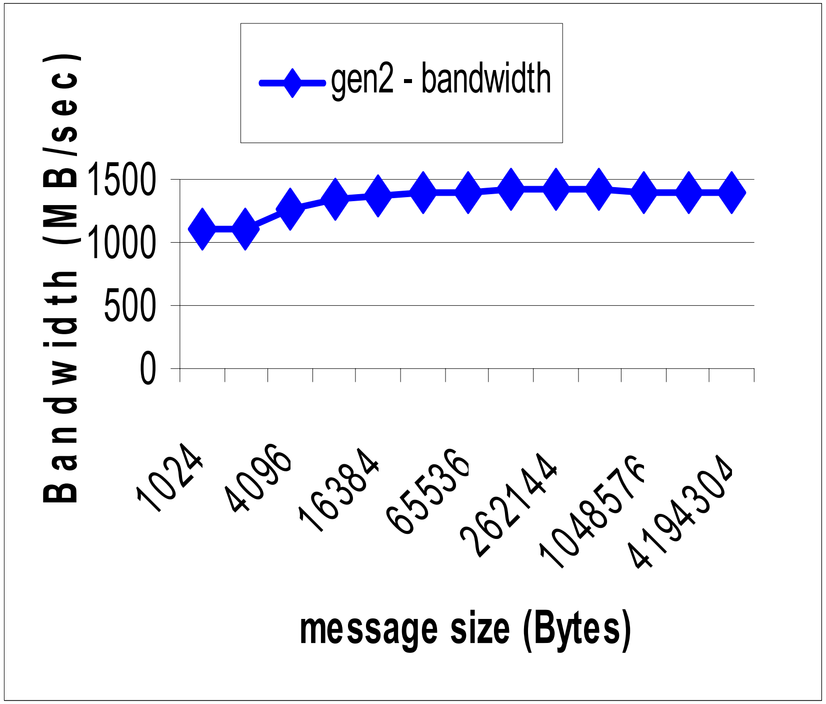
# Trends in Networking Technologies

- Significant growth in commodity systems area networking technology during the last three years
  - InfiniBand, Quadrics, Myrinet
  - Emerging 10GigE
- PCI-Express and Hyper transport interfaces
  - Allow tight integration of NICs to memory
- Performance
  - Low latency (~2us)
  - High bandwidth (~10-15Gbps)
    - InfiniBand 12X is coming (~30Gbps)
  - Low CPU overhead
    - less than 2-3% with InfiniBand

# InfiniBand 4X DDR with OpenIB Gen2 stack

2.43



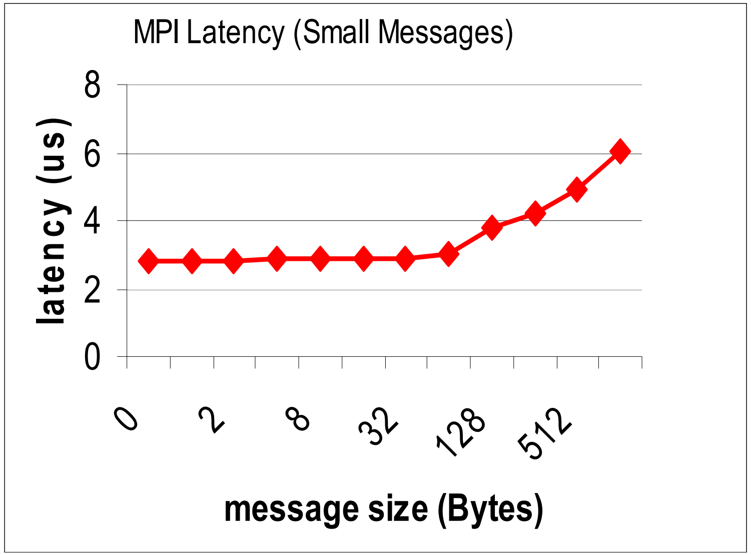


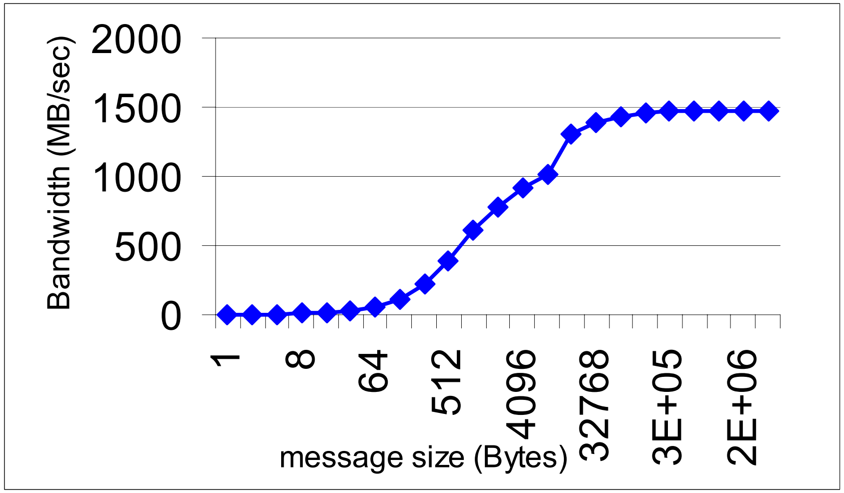
~1500

# MVAPICH-Gen2 with InfiniBand 4X DDR: MPI-Level Performance

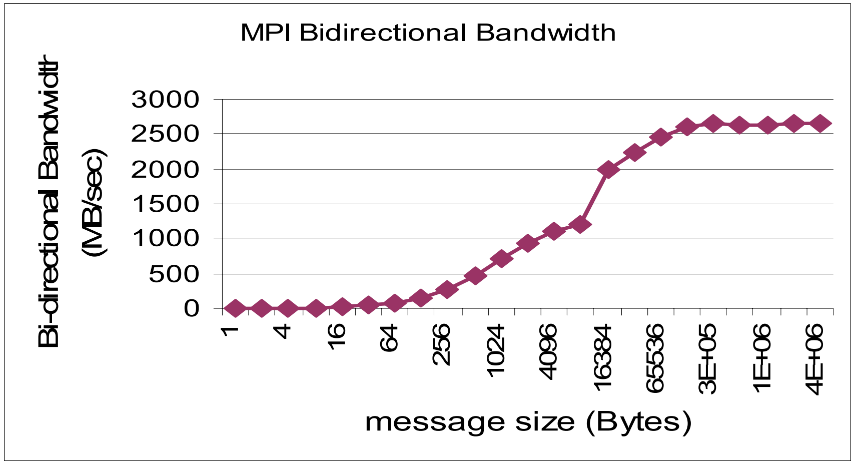
http://nowlab.cse.ohio-state.edu/projects/mpi-iba/

2.84





1458

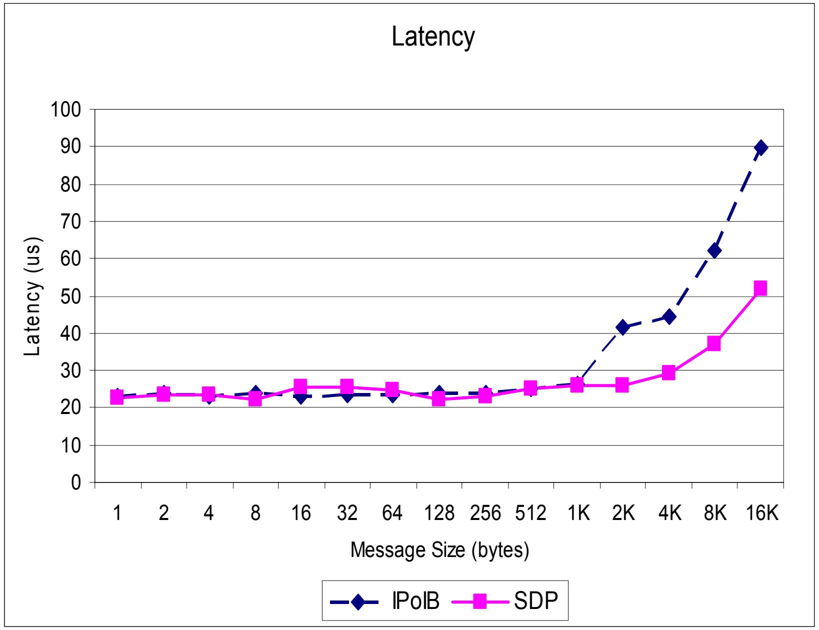


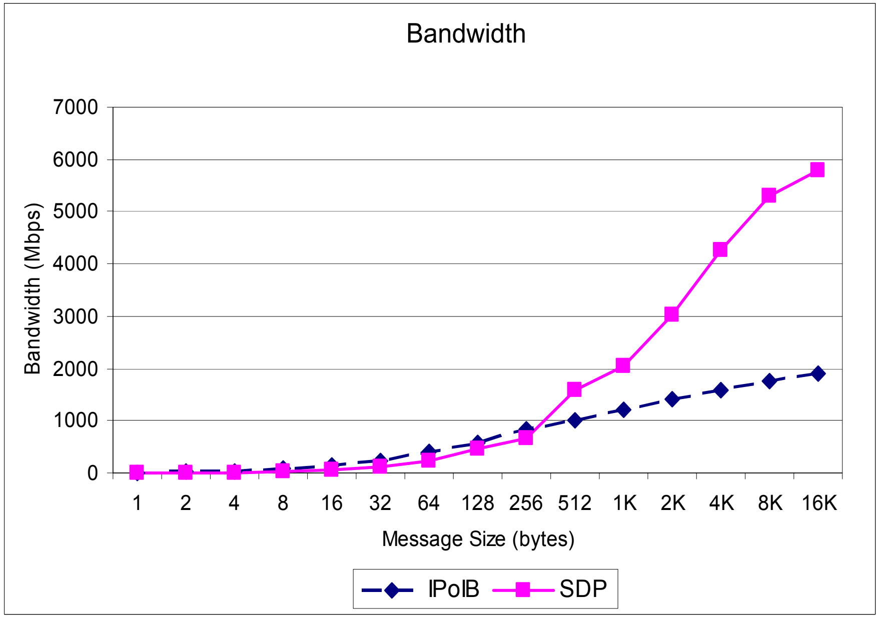
2646

- Single port results only
- Dual port results will be better

08/15/05

# SDP vs. IPoIB: Latency and Bandwidth (InfiniBand SDR with PCI-Express)





SDP enables high bandwidth (up to 750 MBytes/sec or 6000 Mbps), low latency (21 $\mu s$) message passing

# Impact on Network Performance on I/O and File Systems

- Access to remote memory is an order of magnitude better than local disk access
- The gap is steadily increasing
- Following architectural frameworks are getting more realistic
  - Remote-memory
    - NBD using remote swap memory
  - Network RAM
- Will have impact on caching, pre-fetching, …

# Mechanisms in Modern Networks

- Remote DMA (RDMA)
  - Read
  - Write
- Gather/Scatter support with RDMA
- Network-level Atomic operations
  - Fetch and add
  - Compare and Swap
- Network-level broadcast and multicast operations
- Processing offload to NIC and Intelligent NIC
  - Programmable Interface (Myrinet and Quadrics)
  - Integrated MMU (Quadrics)
  - TCP/IP offload (10GigE-Chelsio)
- RDMA is also getting popular in the WAN context
  - RDDP/iWARP protocols
- How to take advantage of these features in designing next generation architecture for I/O and File Systems?

# Fault-Tolerance and RAS Capabilities in Networks

- InfiniBand provides sophisticated Subnet Management
  - Detection of failures (NIC, cable, and switch)
  - Recovery
- Automatic Path Migration (APM)
- Hardware and software support for RAS
- Other networks are also providing some of these features
- How to design fault-tolerant I/O and File Systems with these features while providing Scalability?

# Current State

- Advances in networking (performance and mechanisms) are being integrated into different I/O and File Systems
  - NFS
  - Distributed/Parallel File Systems (PVFS, PVFS2, GPFS, Panasas, Lustre)
  - Storage Area Network, Network Attached Storage
  - iSCSI (iSER)

# Future Research Challenges

- New architectures for I/O and File Systems by taking Network Performance and Mechanisms
- Networked File/Object Accesses
- Networked Block Accesses
- Fault-tolerant designs for I/O and File Systems

# New architectures for I/O and File Systems

- How to take advantage of the following:
  - ~1 microsec node-to-node latency
  - 30~60 Gbps bandwidth
  - Network-level mechanisms
    - RDMA
    - Atomic
    - Multicast
  - Intelligent NICs

# Networked File/Object Accesses

- Parallelizing data movement is now the trend, with either file-based or object-based protocols
  - NFS over RDMA
  - OSD over RDMA
- What about meta-data?
  - Meta-data is important
  - Meta-data is hard to distribute and parallelize
  - Consistence is hard to get right
  - Can we take advantage of the RDMA, atomic, and multicast support mechanisms provided by underlying networks?

# Networked Block Accesses

- Block Accesses can be directed to remote nodes, e.g., NBD and GNBD
- SCSI protocols can be carried over IP (iSCSI) or InfiniBand (iSER)
- Other than GNBD, most of these protocols are typically for single target only
- Parallelize the accesses to block devices
  - Simple early solutions: LVM and RAID
  - More complicated solution: GFS
  - Can we start thinking about a standard for parallelized block accesses?
  - Can such a protocol also be integrated with the emerging RDMA, iWARP and TOE protocols?

# Fault Tolerant I/O and File Systems

- Designing Fault Tolerant IO and File systems
  - Transparent naming, mirroring, network path migration, etc
  - Take advantage of enhanced fault tolerance and RAS support from underlying networks instead of worrying about it in the software layer