Evaluating Lustre on SDR and DDR InfiniBand

W. Yu, R. Noronha, S. Liang and D. K. Panda
Department of Computer Science and Engineering
The Ohio State University
E-mail: panda@cse.ohio-state.edu
http://www.cse.ohio-state.edu/~panda

Presentation Outline

- Overview of InfiniBand
- Sequential IO over SDR & DDR IBA
- Parallel IO over SDR & DDR IBA
- Lustre/Gen2 Design and Current Status
- Conclusions
- Future Research Plan

InfiniBand and Features

- An emerging open standard for high performance interconnect
- High Performance Data Transfer
  - IPC and I/O
  - Low latency (~1.0-3.0 microsec), High bandwidth (~10-20 Gbps) and low CPU utilization (5-10%)
- Flexibility for WAN communication
- Multiple Transport Services
  - Reliable Connection (RC), Unreliable Connection (UC), Reliable Datagram (RD), Unreliable Datagram (UD), and Raw Datagram
  - Provides flexibility to develop upper layers
- Multiple Operations
  - Send/Recv
  - RDMA Read/Write
  - Atomic Operations (very unique)
    - high performance and scalable implementations of distributed locks, semaphores, collective communication operations

Rich Set of Features of IBA (Cont'd)

- Range of Network Features and QoS Mechanisms
  - Service Levels (priorities)
  - Virtual lanes
  - Partitioning
  - Multicast
    - allows to design a new generation of scalable communication and I/O subsystem with QoS
- Protected Operations
  - Keys
  - Protection Domains
- Flexibility for supporting Reliability, Availability, and Serviceability (RAS) in next Generation Systems with IBA features
  - Multiple CRC fields
    - error detection (per-hop, end-to-end)
  - Fail-over
    - unmanaged and managed
  - Path Migration
  - Built-in Management Services

InfiniBand – SDR and DDR

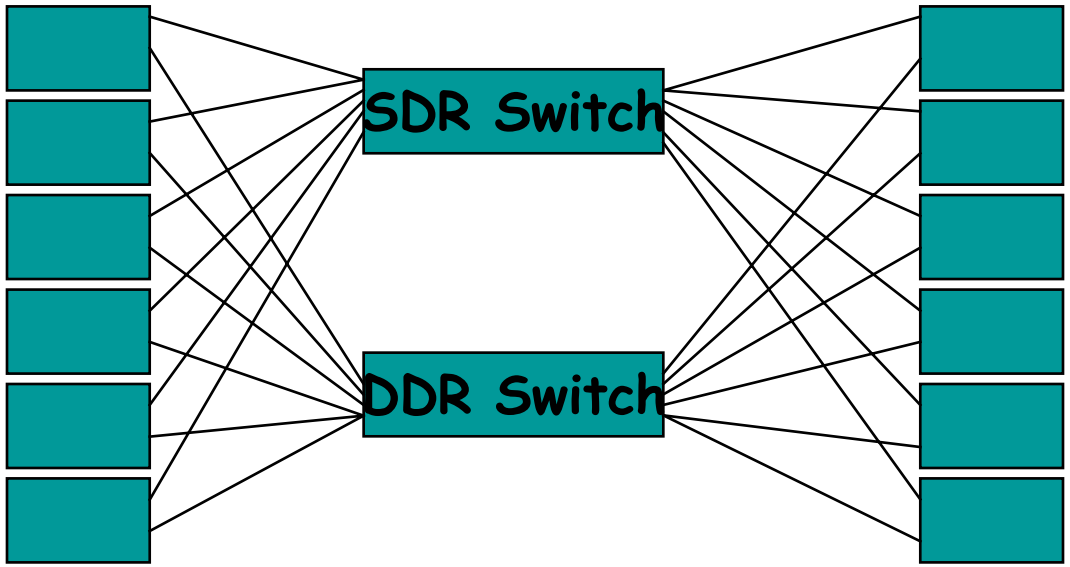
- Blends with emerging network interfaces
  - PCI-Express
  - Hypertransport
- HCAs and switches are available in
  - Single Data Rate (SDR, 4X) – 10 Gbps
  - Double Data Rate (DDR, 4X) – 20 Gbps

Objectives of this Study

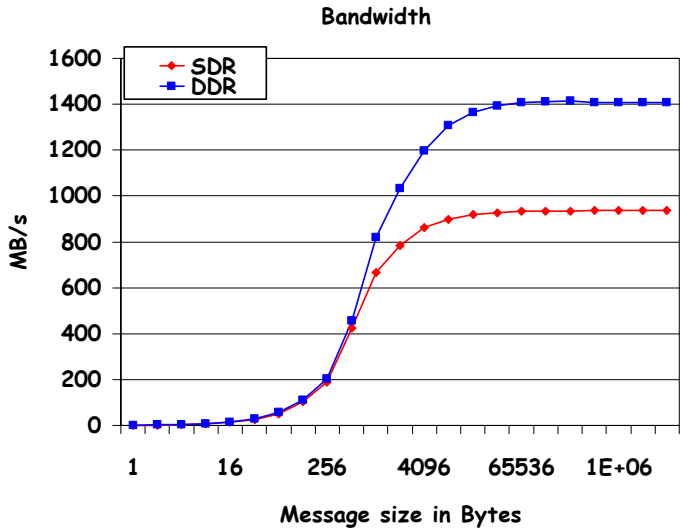
- How well Lustre perform with InfiniBand SDR?
- How much benefits can be achieved with InfiniBand DDR compared to SDR?

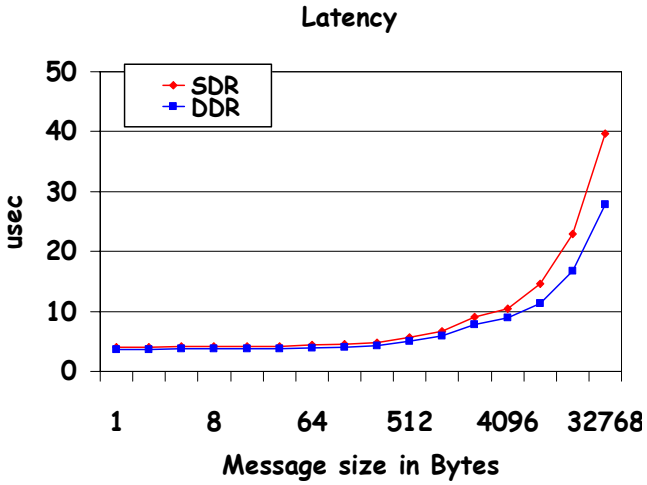
Testbed Configuration

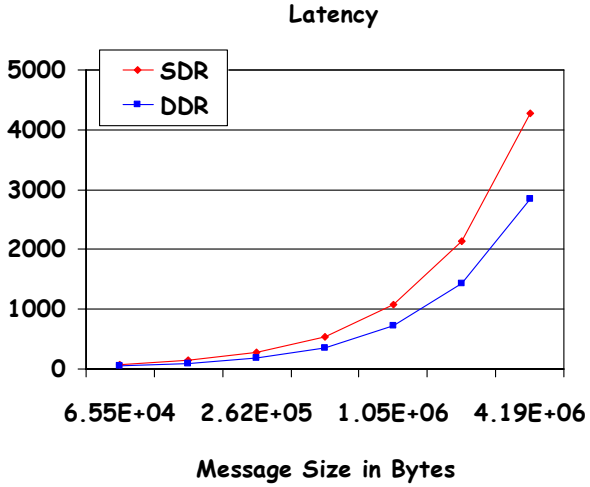
- 16-nodes Blade-server-based cluster with EM64T
- Dual SMP 3.4GHz, 2MB cache and 2 GBytes of memory
- Dual PCI-Express x8 interfaces
- SDR and DDR Switches and HCAs
- Donated to OSU by Intel, Appro and Mellanox



SDR/DDR Performance (Perf_main)







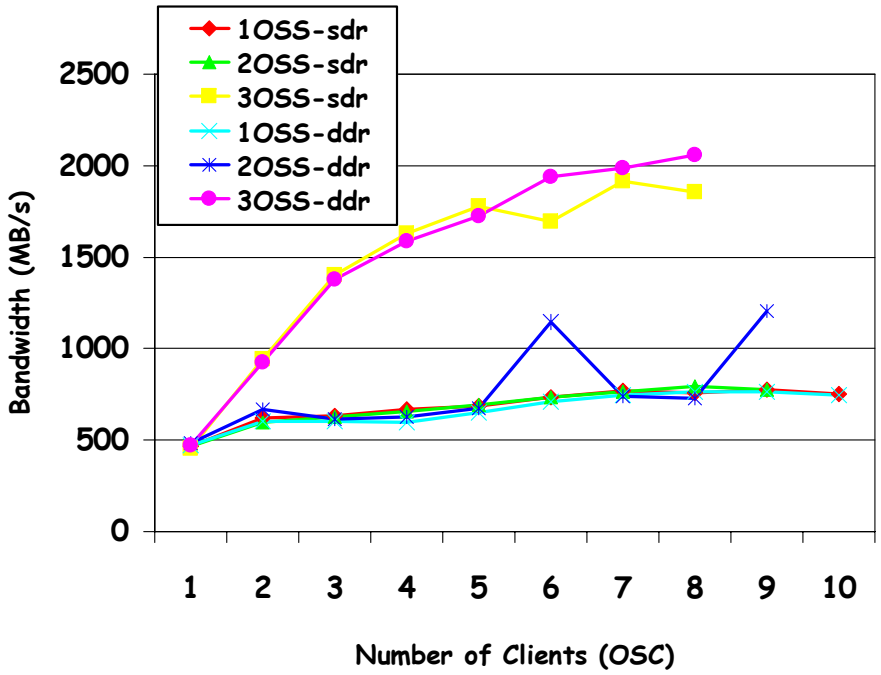
Presentation Outline

- Overview of InfiniBand
- Sequential IO over SDR & DDR IBA
- Parallel IO over SDR & DDR IBA
- Lustre/Gen2 Design and Current Status
- Conclusions
- Future Research Plan

Experimental Setup and Sequential Benchmarks

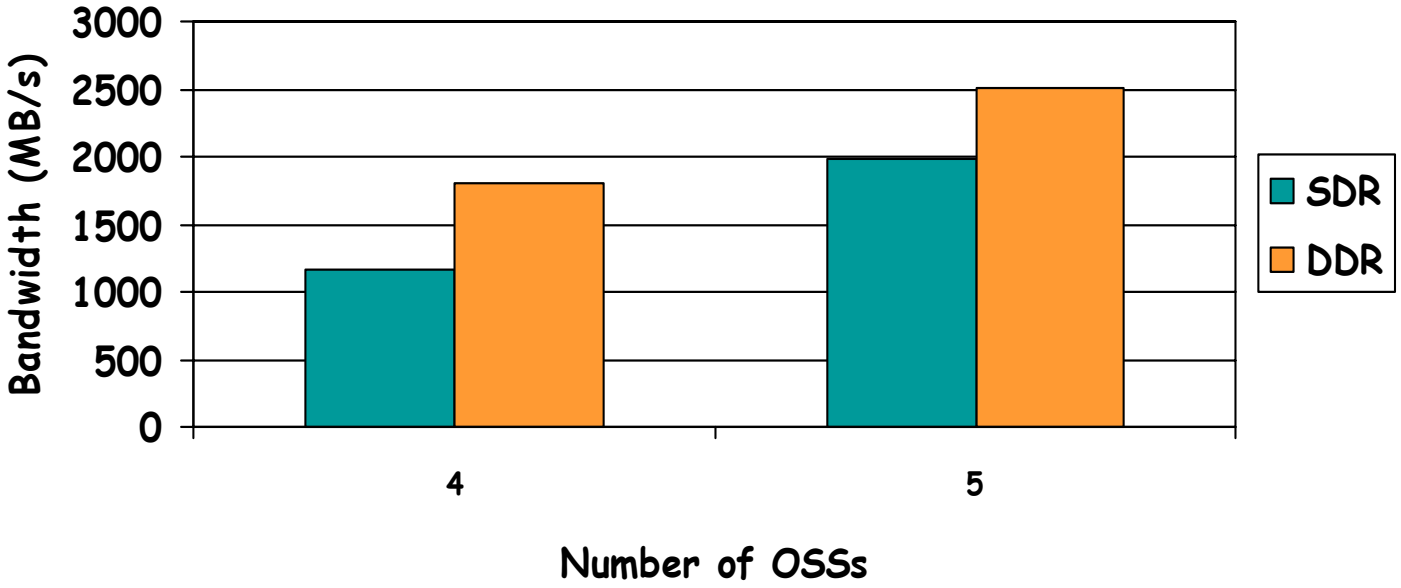
- Experimental Setup
  - Lustre 1.4.6.1
  - One Meta-Data Server
  - Varying numbers of storage server (OSS) and clients (OSC)
    - Up to 12 blades working
  - OSS Configuration
    - Use tmpfs with 1.5Gbytes as storage on each OSS
- Sequential Benchmarks
  - IOzone
    - A popular file system benchmark
    - Analyze file system and I/O performance in various operations
    - Use test file of 128M and record of 512K as default size
  - Fileop
    - A meta-data intensive benchmark
    - Operations include create, stats, readdir, link, unlink and delete
  - Postmark
    - Simulate workloads for Internet email and news servers
    - Measure the transaction rate for a pool of small files
    - Transactions are mostly read/append and create/delete

IOzone Performance - Flush Write



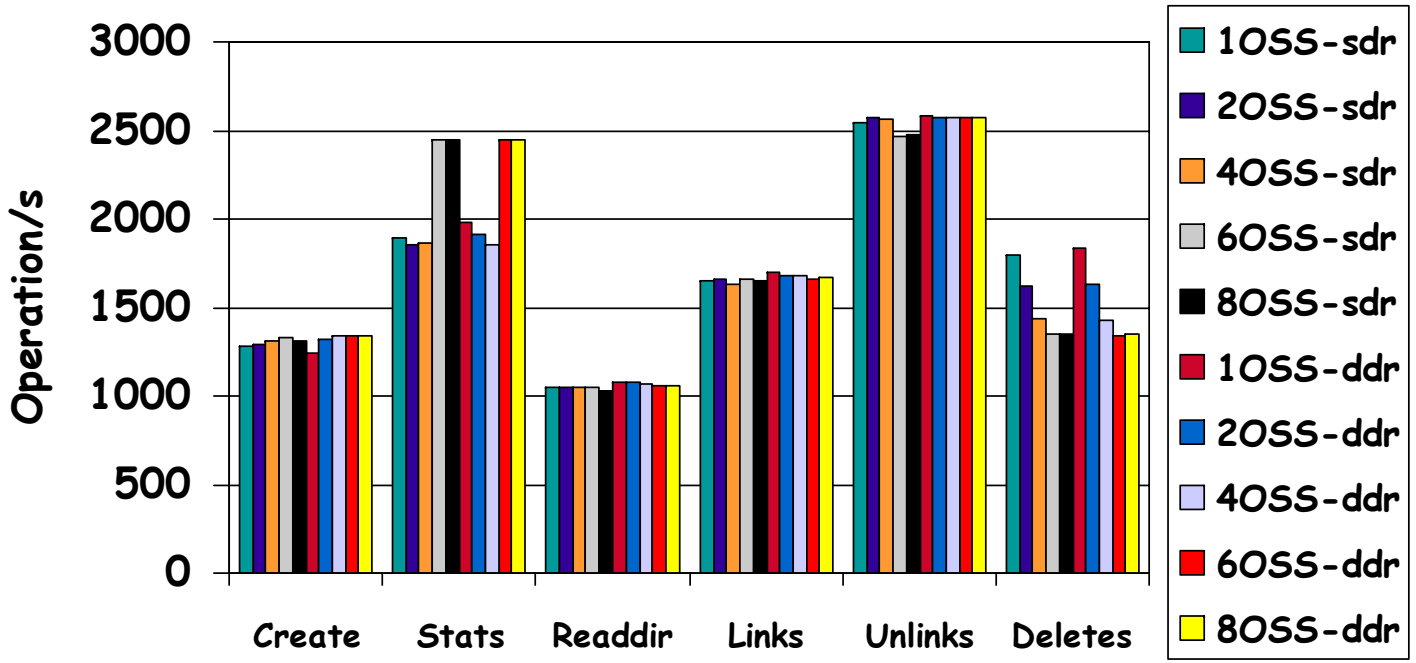
- Up to 10 OSCs perform IO from 1 to 3 OSSs
- DDR shows the benefits with higher number of OSCs
  - More clients exert heavier load on the OSSs, where high bandwidth of DDR helps.

IOZone Flush Write Performance with 4/5 OSSs and 6 Clients



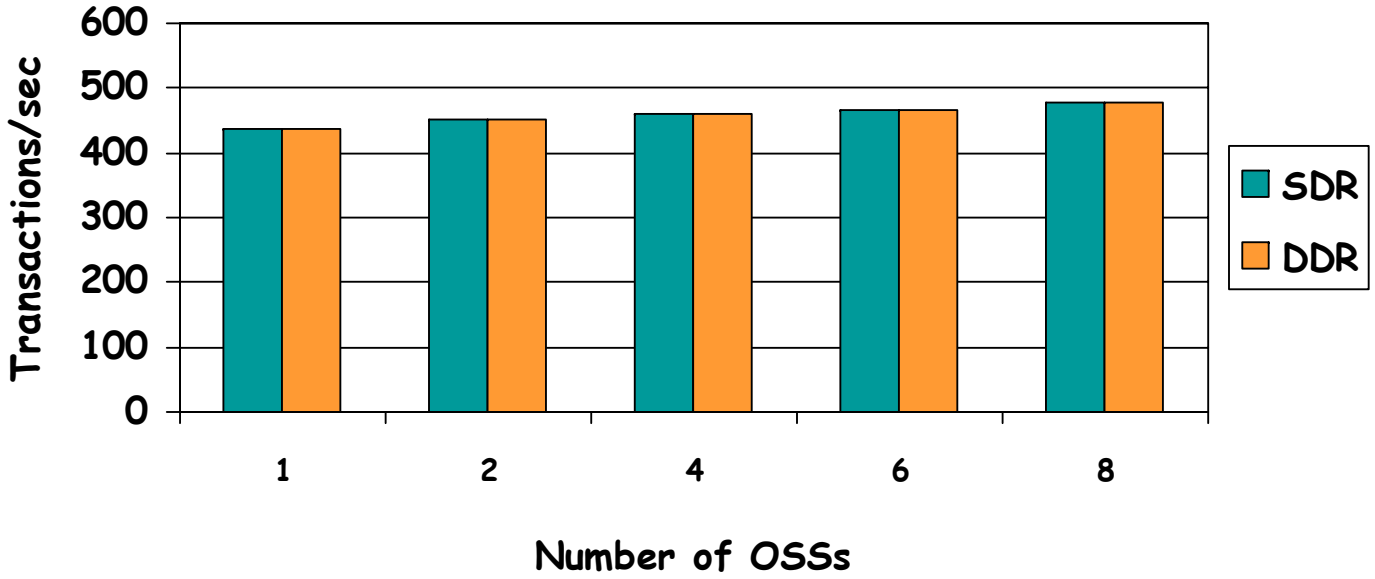
With higher number of OSSs and Clients, DDR-based Lustre performs better compared to SDR-based

Fileop Performance



- Results from creating 10-ary tree of height two
- Single client with varying number of OSSs
- Similar performance because of low pressure on network

Postmark Performance



- Results from 100,000 transactions on 4096 files
- Single client with varying number of OSSs
- Small file operations, thus do not lead to benefit

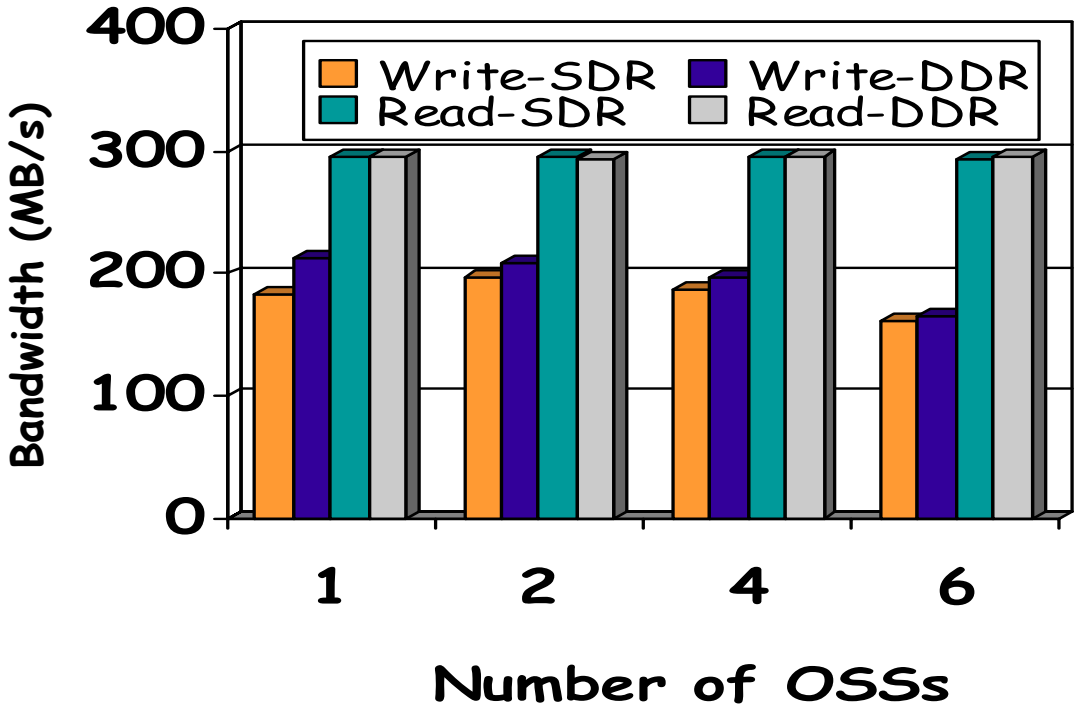
Presentation Outline

- Overview of InfiniBand
- Sequential IO over SDR & DDR IBA
- Parallel IO over SDR & DDR IBA
- Lustre/Gen2 Design and Current Status
- Conclusions
- Future Research Plan

Parallel IO Benchmarks

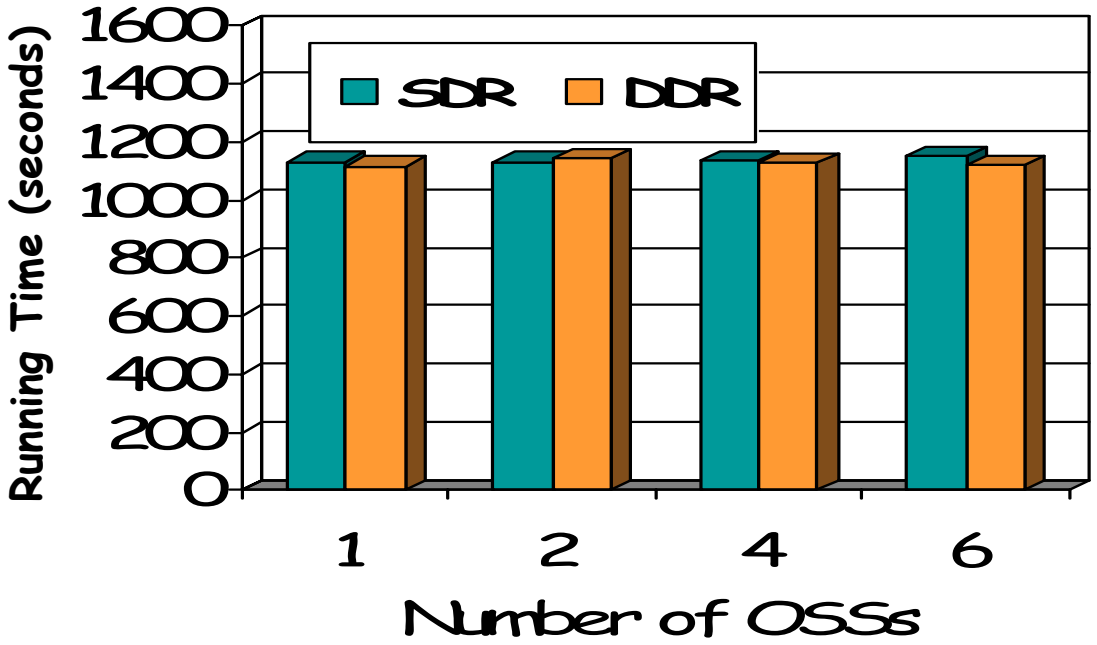
- MPI-Tile-IO
  - A tile reading MPI-IO application from ANL
  - Simulates the workload in some visualization and numerical applications
  - 4 client processes render graphical display of 1024 x 768 pixels, each of 46 bytes
  - 4 tiles in the X dimension and 4 tiles in the Y dimension
  - OSSs use a disk based file system (ext3)
- BT-IO
  - Developed at NASA Ames for NAS BT benchmark
  - Tests the speed of parallel IO capability of HPC applications
  - Data undergoes complex decompositions and distribution, which incurs intensive IO accesses
  - 4 clients processes run class B
  - OSSs use a disk based file system (ext3)

Performance of MPI-Tile-IO



- 4 client processes
- Write Bandwidth improved by 15.8% with one OSS
- Varying improvement with other OSSs
- Read bandwidth is comparable, due to file is caching

Performance of BT-IO



- 4 client processes run class B
- Execution time is comparable for SDR and DDR

Presentation Outline

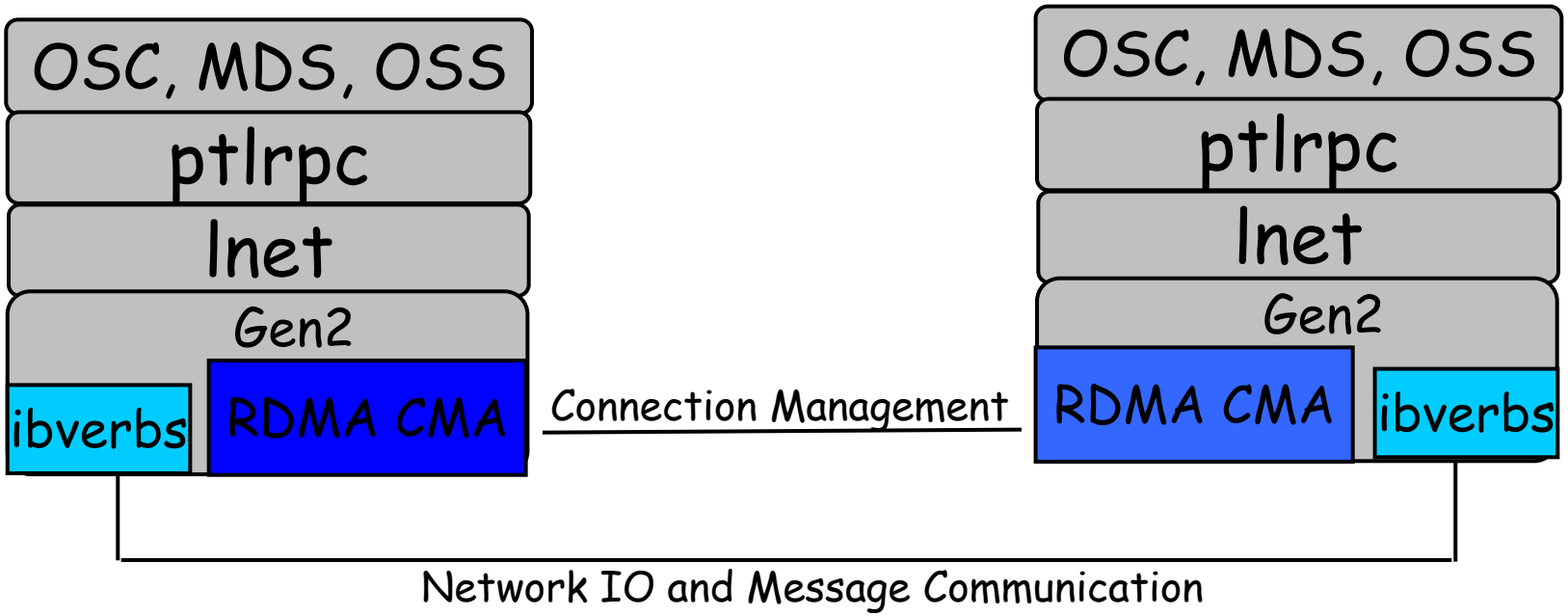
- Overview of InfiniBand
- Sequential IO over SDR & DDR IBA
- Parallel IO over SDR & DDR IBA
- Lustre/Gen2 Design and Current Status
- Conclusions
- Future Research Plan

Lustre/Gen2

- OpenIB Gen2 (New Open Fabrics)
  - www.openib.org
  - Ready to go in production with wide deployment bases
- Current Lustre over InfiniBand
  - Three different stacks for different vendors
    - OpenIB (Gen1), SilverStorm, Voltaire
  - Being tested with OpenIB/Gen2
- OSU Lustre/Gen2
  - Linux-2.6.15.4smp, patched with Lustre support
  - OpenIB SVN revision 5528

Lustre/Gen2 Software Stacks

- Components
  - Connection Management with RDMA CMA
  - Network IO and Communication with ibverbs
  - Support only a single HCA, to be optimized with FMR



Lustre/Gen2 Design

- Kernel Threads
  - Scheduling thread
    - Handle send and receive completion
  - Connection thread
    - Manage connection establishment and teardown
- Experiences with RDMA CMA
  - Desirable to post receive before CMA *ESTABLISHED* event
  - Race condition between *ESTABLISHED* event and incoming receives
    - Server, for example, may be waiting for ESTABLISHED event but gets a receive interrupt as the other side reaches the establishment stage earlier and posts a send
    - Mark connection as ready-to-receive while waiting for ESTABLISHED event

Lustre/Gen2 Status

- Implementation Status
  - Currently working over IA32 with separate MDS, OSS and OSC
  - To be tested over EM64T/Opetron soon, mostly should be fine
  - To merge with Gen2 release 1.0
  - To optimize and test over larger scale systems
- Performance with IOzone (512MBytes file size)
  - Tested over 2.4GHz Xeon nodes, 1GBytes memory, PCI-X and SDR HCA
  - Very preliminary results
  - Write bandwidth:
    - 20MB/sec with disk-based backend storage
  - Read bandwidth:
    - 450MB/sec with disk-based backend storage
    - Small file size, may have cache effects

Presentation Outline

- Overview of InfiniBand
- Sequential IO over SDR & DDR IBA
- Parallel IO over SDR & DDR IBA
- Lustre/Gen2 Design and Current Status
- Conclusions
- Future Research Plan

Conclusions

- Evaluated Lustre over SDR and DDR HCAs on a 12-node InfiniBand cluster
- Sequential IO experiments show that DDR HCAs can provide benefits to Lustre file system when larger number clients exert heavy IO load
- Parallel IO experiments also indicate DDR HCAs is beneficial compared to SDR-based Lustre configuration
- MPI-Tile-IO bandwidth improvement varies from 5-15.8%
- Need to carry out studies in larger-scale systems
  - OSU is deploying a 64-node InfiniBand DDR cluster
- Prototyped Lustre over OpenIB Gen2 stack
- Initial performance numbers are comparable with Lustre over OpenIB Gen1

Presentation Outline

- Overview of InfiniBand
- Sequential IO over SDR && DDR IBA
- Parallel IO over SDR && DDR IBA
- Lustre/Gen2 Design and Current Status
- Future Research Plan

Future Research Plan

- Scalable MPI-IO over Lustre
  - Optimize the current UFS-based ADIO for Lustre
  - Or design an ADIO device over liblustre for MPI-IO
- High Performance Parallel Checkpoint/Restart over Lustre
  - Check pointing/Restart (CR) has been incorporated into OSU MVAPICH2
    - Q. Gao, W. Yu, W. Huang and D. K. Panda, Application-Transparent Checkpoint/Restart for MPI Programs over InfiniBand, Int'l Conference on Parallel Processing (ICPP), June 2006
    - Working prototype was demonstrated at OpenIB-March '06
  - Will be released in upcoming MVAPICH2 release
  - Heavy IO pressure placed on the file system with CR
  - Intend to design a CR-oriented optimized IO path with Lustre

Future Research Plan (Cont'd)

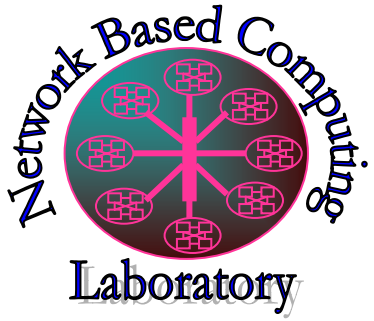
- Cooperative caching for Lustre
  - Use high speed networks for fast global memory caches
  - Explore file delegation for cooperative read/write into global caches
  - Intend to disk-cache file objects for failure recovery
  - Leverage InfiniBand RDMA and atomic capabilities where possible

Acknowledgements

- Current Funding support by

- Current Equipment support by

Web Pointers



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

Other Storage and File system Projects
(GFS over InfiniBand, NFS-RDMA over InfiniBand)

http://nowlab.cse.ohio-state.edu/projects/clust-storage/

http://nowlab.cse.ohio-state.edu/projects/nfs-rdma/