

Evaluating the Impact of RDMA on Storage I/O over InfiniBand

J Liu, DK Panda and M Banikazemi[†]



Computer and Information Science
The Ohio State University

[†]IBM T J Watson Research Center



•
•

Presentation Outline

- **Introduction/Motivation**
- **RDMA Assisted iSCSI Overview**
- **Design and Implementation**
- **Performance Evaluation**
- **Conclusion**



InfiniBand Overview



- Industry standard
- Interconnect for connecting processing nodes and I/O nodes
- High performance
 - Less than 5 μ s latency
 - Over 840MB/s unidirectional Bandwidth
- InfiniBand clusters are becoming increasingly popular





Storage for InfiniBand Clusters



- Local storage
- Network storage
 - Network Attached Storage (NAS)
 - Storage Area Networks (SANs)



⋮

SAN for InfiniBand Clusters

- Fibre Channel (FC)
- SCSI RDMA Protocol (SRP)
- Internet SCSI (iSCSI)



FC and SRP



- **Fibre Channel (FC)**
 - Good performance
 - Requires new hardware (HBAs, switches)
 - Requires separate management infrastructure
- **SCSI RDMA Protocol (SRP)**
 - InfiniBand native protocol
 - No new hardware required
 - Requires implementation from scratch
 - Requires new management infrastructure

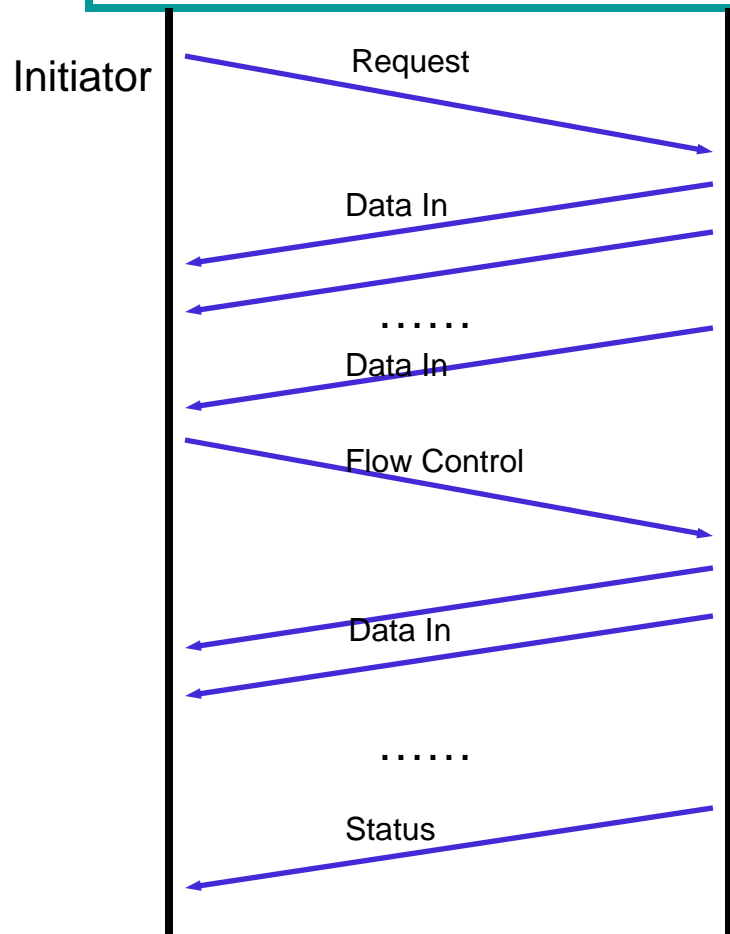


iSCSI



- Uses TCP/IP as the underlying transport layer
 - No additional hardware for hosts (InfiniBand supports IPoIB)
 - Relative less software development effort (Existing management infrastructure in TCP/IP can be reused)
- Performance may be an issue
 - High overhead in the TCP/IP stack

iSCSI Data Transfer: Read

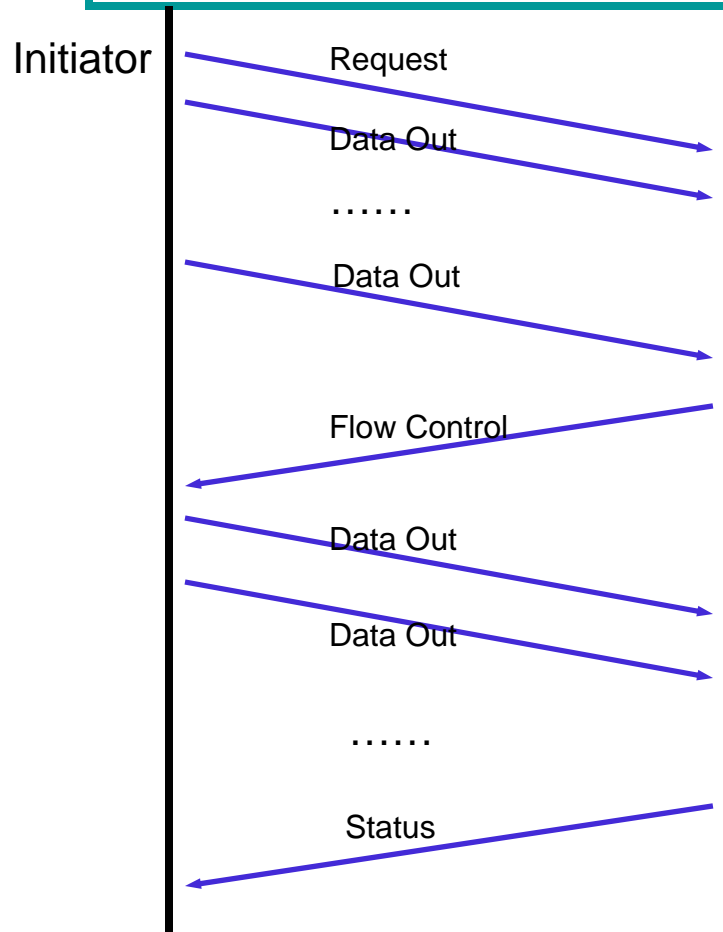


→ TCP/IP Communication

Target

- All communication through TCP/IP
- Multiple data packets may be necessary
- Flow control for data packets may be necessary

iSCSI Data Transfer: Write



→ TCP/IP Communication

- All communication through TCP/IP
- Multiple data packets may be necessary
- Flow control for data packets may be necessary



Problems with iSCSI



- Limited Performance because
 - Protocol overhead in TCP/IP
 - Interrupts are generated for each network packet
 - Extra copies when sending and receiving data





Improving iSCSI Performance



- Eliminating receiver side copies in the TCP/IP stack
- Direct data placement in HBA
 - Special hardware
 - Low compatibility
- iSCSI extension for RDMA
 - Special hardware (RNICs)
 - Standard interface (RDMA over IP)
- Using RDMA over InfiniBand
 - RDMA assisted iSCSI

•
•

Presentation Outline

- Introduction/Motivation
- **RDMA Assisted iSCSI**
- Design and Implementation
- Performance Evaluation
- Conclusion

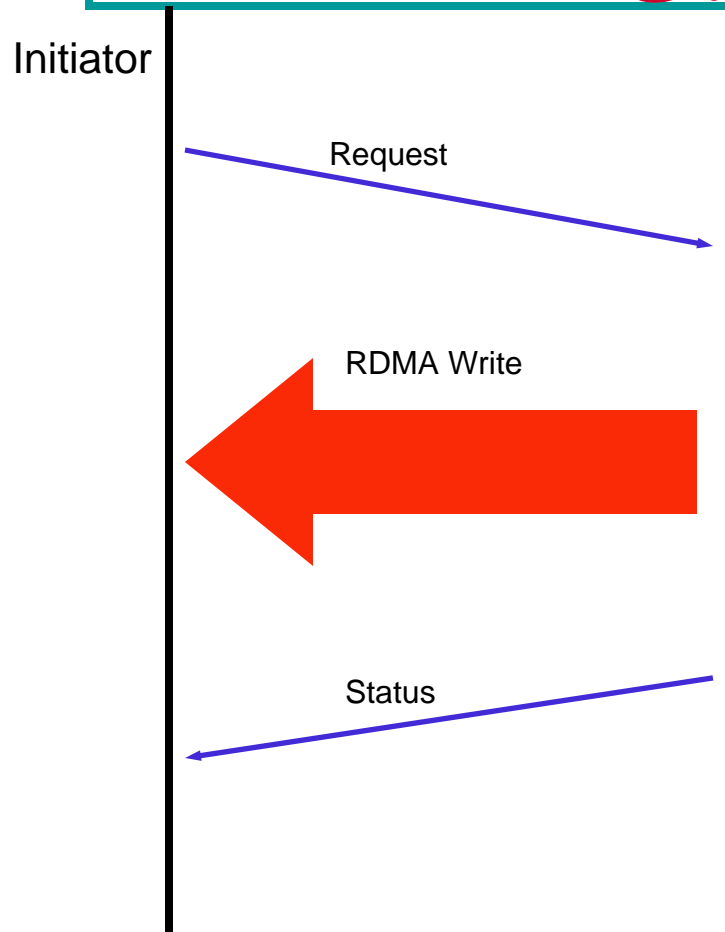


RDMA Assisted iSCSI Overview



- Combining IPoIB and RDMA over InfiniBand in iSCSI
 - iSCSI control PDUs go through TCP/IP
 - iSCSI data transfers use RDMA
 - IPoIB for compatibility
 - RDMA over InfiniBand for performance
- Reusing existing infrastructure
 - Reusing many existing IP based protocols
 - No additional hardware needed for hosts

iSCSI Data Transfer with RDMA: Read

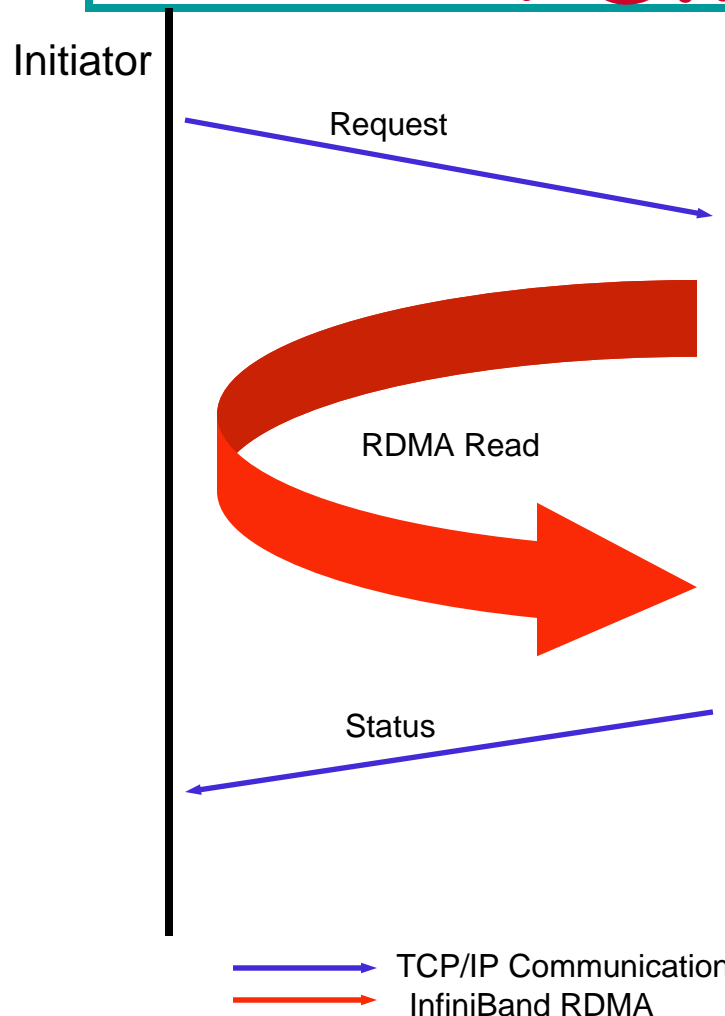


→ TCP/IP Communication
→ InfiniBand RDMA

Target

- Requests carry buffer information (address, length, tags ...)
- All control transfer through TCP/IP
- All data transfer through InfiniBand RDMA
- No need for multiple data packets
- No flow control for data packets necessary

iSCSI Data Transfer with RDMA: Write



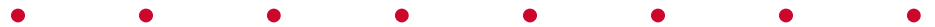
- Requests carry buffer information
- All control transfer through TCP/IP
- All data transfer through InfiniBand RDMA
- No need for multiple data packets
- No flow control for data packets necessary



Advantages of Using RDMA



- **Improved performance**
 - Much less protocol overhead (TCP/IP bypassed for data transfer)
 - No data copies in the protocol
 - Reduced number of interrupts at the client



•
•

Presentation Outline

- Introduction/Motivation
- RDMA Assisted iSCSI
- Design and Implementation
- Performance Evaluation
- Conclusion

•
•
•

Design Issues

- **Memory Registration**
- **Session Management**
- **Reliability and Security**



• • • • • • • • •



Memory Registration



- Memory needs to be registered before it can be used for RDMA in InfiniBand
- Memory registration cost is high
 - Targets can usually pre-registered all the memory to avoid the cost
 - More difficult for hosts (clients) because they do not have total control over the memory buffers



Techniques to Improve Memory Registration



- **Memory Registration Cache (MRC)**
 - Maintains a “cache” of registered buffer and de-register buffer in a lazy manner
 - Depends on buffer reuse
 - May not be effective for storage buffers
- **Fast Memory Registration (FMR)**
 - Divide registration into two steps (preparation and mapping)
 - Only mapping appears in the critical path
 - Supported in some InfiniBand implementations
- **Zero-Cost Kernel Memory Registration (ZKMR)**
 - Map physical memory to virtual address space in kernel and pre-register the mapped virtual address space
 - Do “virtual” to “mapped” address translation during communication (very fast)
 - Some limitations





Session Management



- Use TCP/IP exclusively for LOGIN phase
 - Reusing existing bootstrapping and target discovering protocols
- LOGIN phase negotiate the use of InfiniBand RDMA
 - Fall back on the original iSCSI if RDMA cannot be used



Reliability and Security



- Reliability
 - TCP/IP checksum may be insufficient for some applications
 - iSCSI supports CRC
 - No need to use CRC in RDMA assisted iSCSI because InfiniBand has end-to-end CRC
- Security
 - RDMA assisted iSCSI can take advantage of existing authentication protocols
 - However, IPSec cannot be used directly



Implementation



- Linux 2.4.18
- Based on Intel v18 iSCSI implementation
- InfiniBand Access Layer and IPoIB
- Ram disk based target implementation
- Kernel SCSI driver at the client

•
•

Presentation Outline

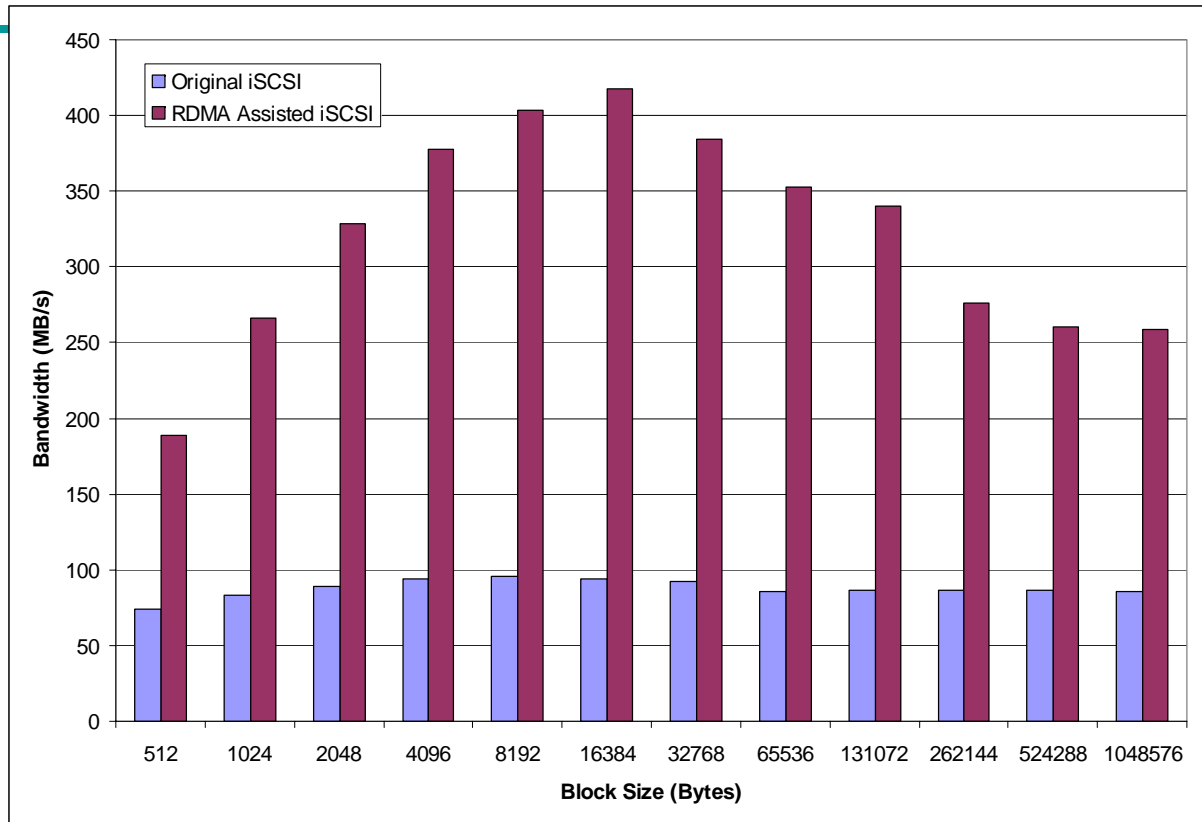
- Introduction/Motivation
- RDMA Assisted iSCSI
- Design and Implementation
- Performance Evaluation
- Conclusion

•
•
•

Experimental Testbed

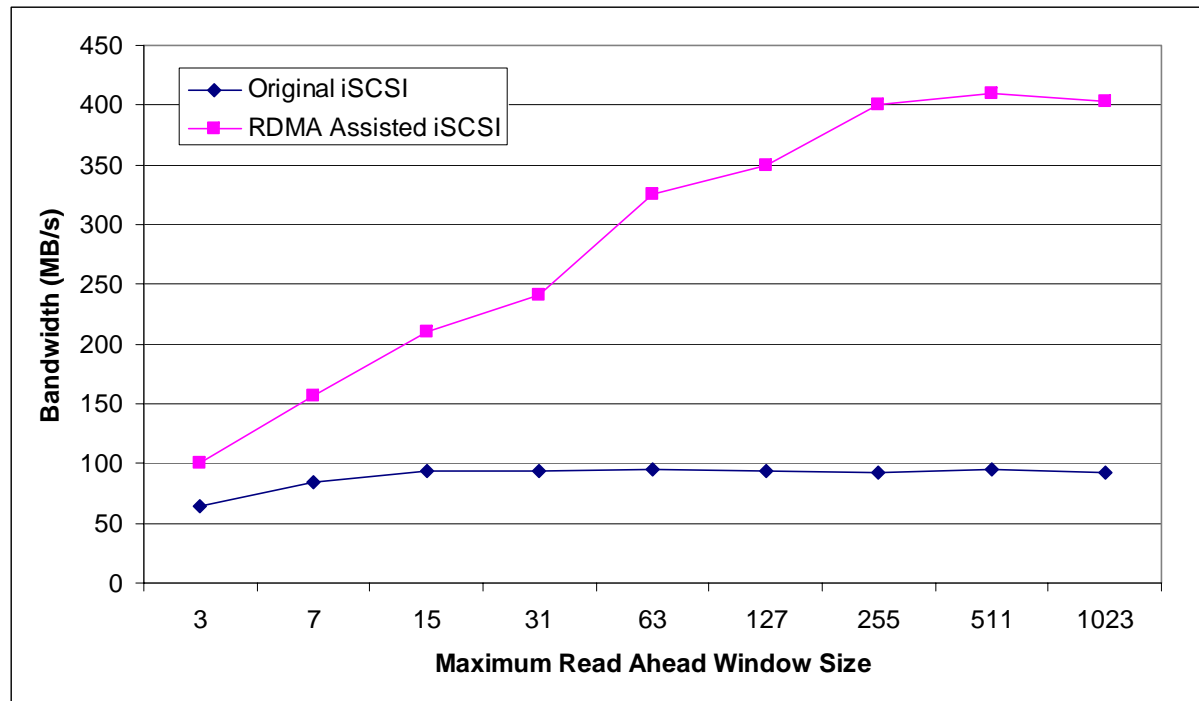
- SuperMicro SUPER P4DL6 nodes (2.4 GHz Xeon, 400MHz FSB, 512K L2 cache)
- Mellanox InfiniHost MT23108 4X HCAs (A1 silicon), PCI-X 66bit 133MHz
- Mellanox InfiniScale MT43132 switch

File Read Bandwidth



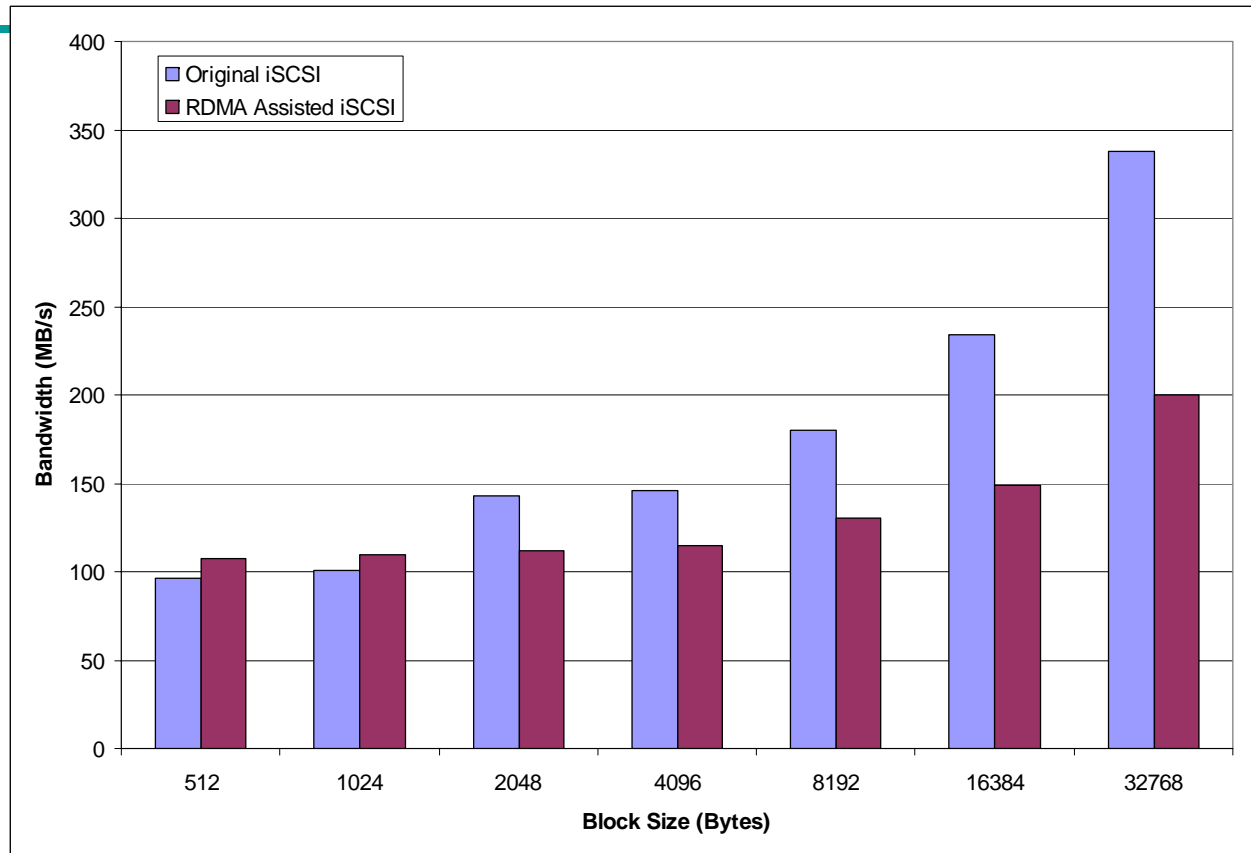
- Buffered I/O used
- RDMA can improve peak bandwidth from 97MB/s to over 400MB/s
- 16KB block size performs best

Impact of Read Ahead



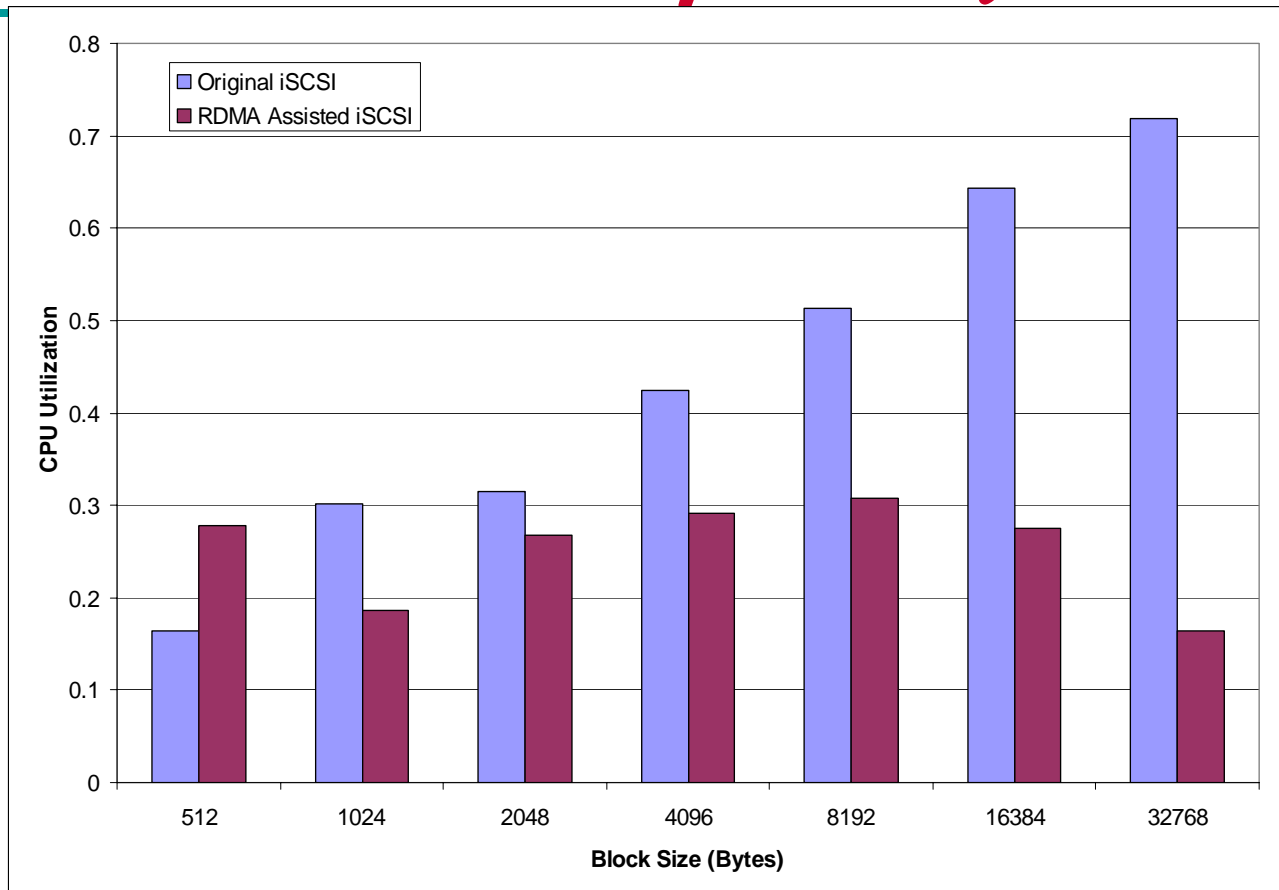
- Prefetching has greater impact on the performance of RDMA assisted iSCSI

File Read Latency



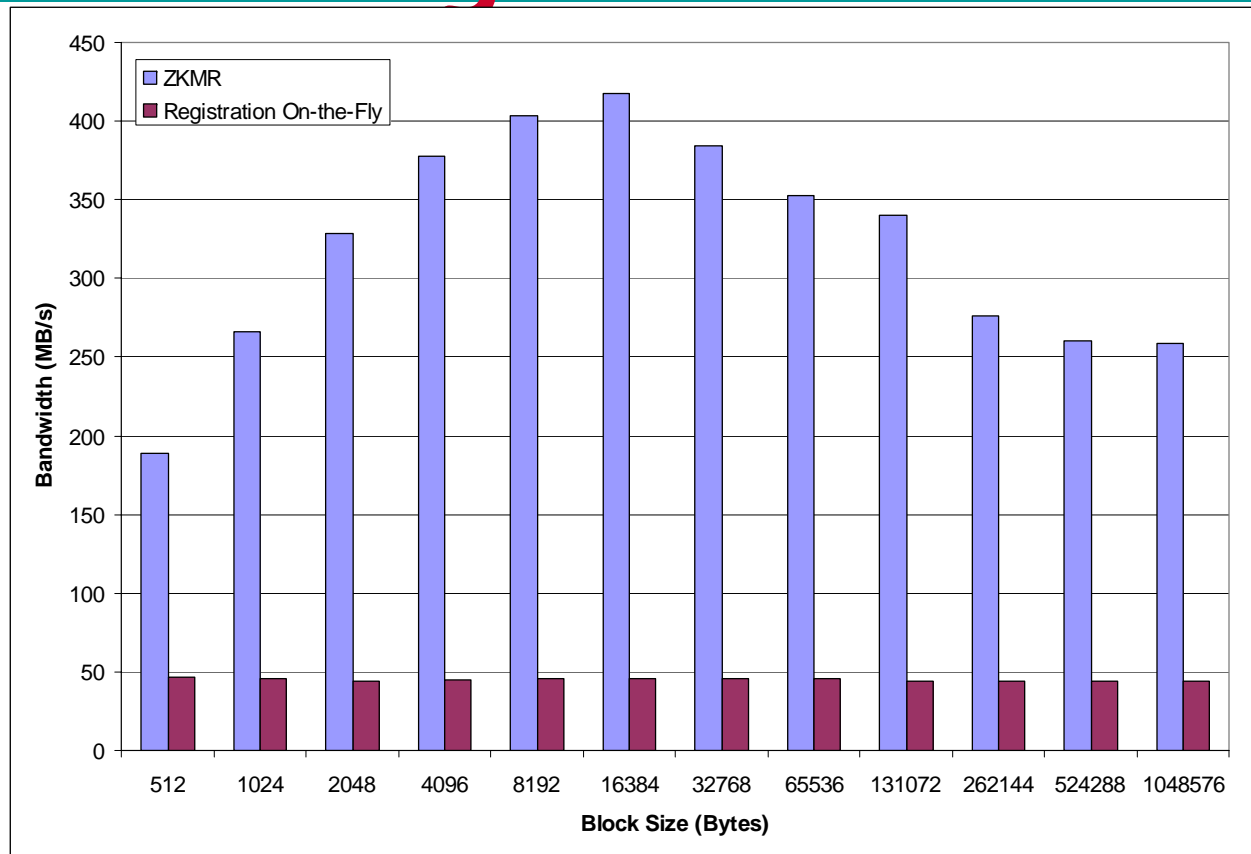
- RAW I/O used instead of buffered I/O
- RDMA improves performance for large block sizes

Host CPU Utilization (File Read Latency Test)



- RDMA reduces CPU utilization for large block sizes

Impact of Buffer Registration



- Registration cost significantly degrades performance
- RDMA is beneficial only when registration cost can be avoided/reduced



Conclusion



- RDMA assisted iSCSI over InfiniBand
 - Evaluating the use of RDMA in storage protocols
 - RDMA can significantly improve storage communication performance
 - Provide useful insight for other protocols such as iSER and SRP
 - A practical storage solution for InfiniBand clusters