Performance of RDMA-capable Storage Protocols on Wide-Area Network [*]

Weikuan Yu[†] Nageswara S.V. Rao[†] Pete Wyckoff[‡] Jeffrey S. Vetter[†]

Oak Ridge National Laboratory[†]
Computer Science & Mathematics
Oak Ridge, TN 37831
{*wyu,raons,vetter*}*@ornl.gov*

Ohio Supercomputer Center[‡]
1224 Kinnear Road
Columbus, OH 43212
pw@osc.edu

Abstract

Because of its high throughput, low CPU utilization, and direct data placement, RDMA (Remote Direct Memory Access) has been adopted for transport in a number of storage protocols, such as NFS and iSCSI. In this presentation, we provide a performance evaluation of RDMA-based NFS and iSCSI on Wide-Area Network (WAN). We show that these protocols, though benefit from RDMA on Local Area Network (LAN) and on WAN of short distance, are faced with a number of challenges to achieve good performance on long distance WAN. This is because of (a) the low performance of RDMA reads on WAN, (b) the small 4KB chunks used in NFS over RDMA, and (c) the lack of RDMA capability in handling discontiguous data. Our experimental results document the performance behavior of these RDMA-based storage protocols on WAN.

Keywords: *NFS, iSCSI, RDMA, InfiniBand, 10GigE*

1. Introduction

Many geographically distributed high-performance computing systems and data centers are producing and/or supplying large volumes of data sets. Such data often must be transported to storage, visualization and analysis systems that are remotely located. The traditional approach of storing the data sets on local files, and utilizing TCP/IP tools such as GridFTP [18] and bbcp [2] for data transfer on the wide-area network (WAN), requires additional very sophisticated per-connection optimizations that are onerous to the users. On the other hand, RDMA has been extensivly exploited as a technology in the local area networks for various storage protocols. It is used to leverage the benefits from the latest networking technologies such as InfiniBand (IB) [8] and iWARP [15]. For example, networked storage subsystems have introduced new data movement layers to utilize RDMA in the existing storage protocols, such as NFS over RDMA (NFSoRDMA) [4] and iSCSI over RDMA (iSER) [9].

Very recently, motivated by the potential to extend IB performance to the wide-area, there have been hardware implementations of IB over Wide-Area network devices (IBoWAN), in particular Longbow XR from Obsidian Research Corporation [11] and NX5010ae from Network Equipment Technologies [10]. They extend the reach of IB connections to several thousand miles, and open up the possibility of connecting supercomputers, clusters and storage systems located thousands of miles apart. Initial results indicated that IB-based RDMA technologies can sustain multiple Gbps network transfer rates over thousands of miles, for example, 7Gbps over 8600 mile connections [13, 19, 14]. However, performance studies of these IBoWAN devices for long-range storage protocols are very limited. Thus, it is important to examine and understand the performance and implications of RDMA technologies to storage access on WAN. In this presentation, we examine the models of Non-RDMA and RDMA-based transport protocols for NFS and iSCSI. We show that RDMA-based storage protocols are faced with a number of challenges to expose the bandwidth potential of RDMA on WAN, because of the low performance of RDMA reads, the small 4KB chunks used in NFSoRDMA, and the lack of RDMA capability in handling discontiguous data. Our initial results document such performance behavior of

[*] This research is sponsored by the Office of Advanced Scientific Computing Research; U.S. Department of Energy. The work was performed at the Oak Ridge National Laboratory, which is managed by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725.

RDMA-based storage on WAN.

The remainder of the paper is structured as follows. In next section, we compare RDMA and Non-RDMA based transport for storage protocols. Then we show initial results of NFS and iSCSI using RDMA. In Section 3, we review related work on NFS and iSCSI. Finally, we discuss possible strategies to improve the performance of RDMA-based storage protocols, and conclude the paper.

2. Distance Scalability of RDMA-based Storage Protocols

In this section, we first examine RDMA- and Non-RDMA based transport for NFS and iSCSI. Then we show the performance of RDMA at the network level. At the end, we present the performance of NFS and iSCSI on top of RDMA.

2.1. Non-RDMA and RDMA Based Transport

While storage protocols are quite different in their internal code paths, their data transport can be generalized into Non-RDMA (a.k.a Send/Receive)- and RDMA-based. As shown in Fig. 1, in the Non-RDMA case, data transport is directly implemented via a two-way mechanism: the sending of a request from the client to the server and the receiving of a reply back from the server to the client. Bulk data, if any, are transmitted inline with the request and the reply. In contrast, in the RDMA-based case, requests and replies become pure control messages, serving the purposes of requesting the beginning and notifying the completion of the data transfer, respectively. The actual data transmission is executed through separated RDMA read or write operations. In the case of NFS over RDMA, when the list of memory segments are too long to be contained in a single request, the server needs additional RDMA operations to pull a long list of address/length vectors from, and/or return an updated list of the same to, the client. We measure the performance of these RDMA-based storage protocols, and examine the implications of RDMA to them on WAN.

2.2. Experimental Environment

Hardware – We have used UltraScience Net (USN) at Oak Ridge National Laboratory for performance measurement of storage protocols on WAN. USN is a wide-area experimental network testbed that supports the development of next-generation computational science applications. It spans more than four thousand miles from Oak Ridge (Tennessee) to Atlanta, Chicago, Seattle and Sunnyvale (California) using dual OC192 backbone connections (c.f. [14] for more details). Two NX5010ae devices from Network Equipment Technologies Inc. are configured for IB-over-SONET operation, and are connected to CDCIs on WAN side and to InfiniBand switches on the edges. Two Intel WoodCrest nodes were used in this study, each connected to one of the IB switches. The Woodcrest nodes contained two Xeon 5150 dual-core processors and 8 GB of FB-DIMM memory. The processor clock rate was 2.66 GHz. These computer nodes are equipped with both InfiniHost-III and connect-X DDR HCAs from Mellanox.

Software – The OFED version 1.3.1 was used as the InfiniBand software stack. The NFSoRDMA implementation in Linux version 2.6.25-rc3 was used. Linux version 2.6.23.14 was used in the iSCSI experiments, which provides an integral iSCSI over RDMA (iSER) initiator. The iSER target was from the release of Ohio Supercomputer Center.

RDMA Performance on WAN – We measured the basic network-level bandwidths of RDMA operations on WAN. Fig. 2 shows that both RDMA reads and RDMA writes can reach the peak of 934 MBytes/sec, but at very different message sizes. In particular, the small message bandwidths are very different. For 4KB messages at 0.2 mile, RDMA writes achieve 925 MBytes/sec, RDMA reads do only 223 MBytes/sec, using InfiniHost-III HCAs. At 1400 miles, 4KB RDMA writes achieve only 14.7 MBytes/sec, and 4KB RDMA reads do only 0.59 MBytes/sec. Such behaviors are largely due to the combined reasons of physical distance, the InfiniBand flow control issues, and the different nature of RDMA writes and reads [19]. The latest connect-X HCAs can significantly improves the performance RDMA read because it supports a much larger number (16) of concurrent RDMA read operations, compared to only four in the InfiniHost-III HCAs. Fig. 2 also includes the performance of connect-X at 0.2 mile. Due to circuit availability, we do not have the performance of connect-X HCAs at 1400 miles.,

2.3. NFS over RDMA

We measured the performance of NFS using the IOzone benchmark [1]. All tests were conducted using direct I/O, 16MB request size, and 128MB file size per threads. To avoid disk bottleneck, the NFS server exports its memory to the client. Two different RPC transport protocols were used, RDMA and TCP. The InfiniBand (IB) stack supports TCP through its IPoIB im-

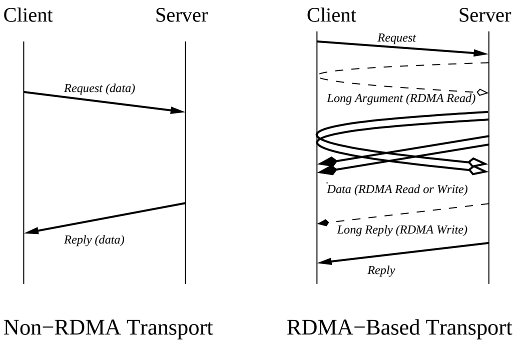


Fig. 1. Non-RDMA and RDMA-Based Transport Models

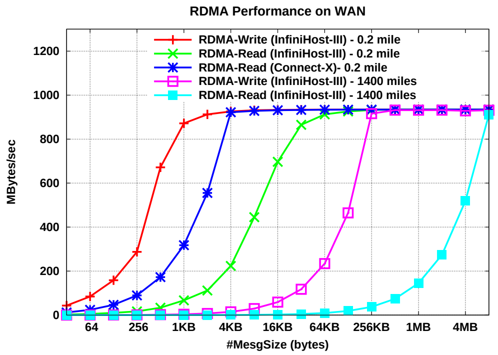
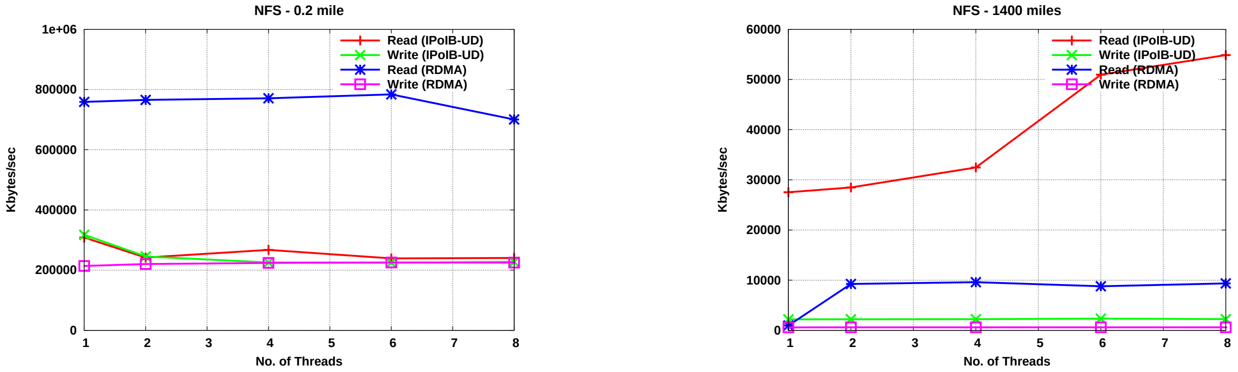


Fig. 2. The Performance of RDMA on WAN (using NX5010e InfiniBand Extension Devices)



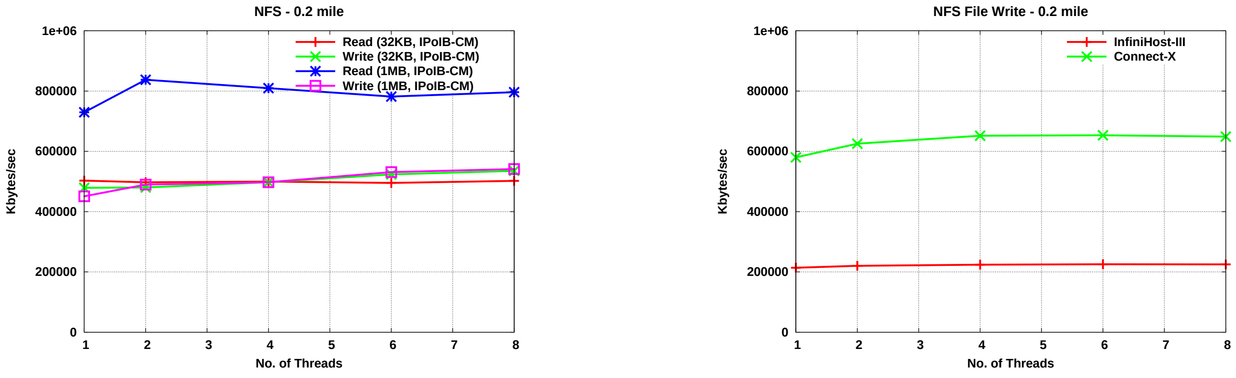
(a) NFS at 0.2 mile

(b) NFS at 1400 miles

Fig. 3. The Performance of NFS on WAN

plementations, either under a connected mode (CM) or an unreliable datagram (UD) mode. Fig. 3 shows that the performance of NFS using InfiniHost HCAs, At 0.2 mile, NFS reads and writes both can achieve 220 MBytes/sec using IPoIB-UD. RDMA increases the performance of NFS file reads significantly, but not file writes. As shown in Fig. 1, for NFS reads, the server sends data via RDMA writes, while the data of NFS writes are pulled by the server via RDMA reads. Though the NFS block size is 32KB by default, its RDMA-based RPC transport protocol (as released for Linux) divides a single block into page-sided chunks, 4KB each. Compared to NFS file reads that use RDMA write operations, the low performance of RDMA read operations at 4KB significantly limits that of NFS file writes. Fig. 3(b) shows the performance measurement of NFS at 1400 miles. Because the bandwidths of 4KB RDMA writes and reads are much lower at long distance, both NFS reads and writes are not able to achieve good performance. NFS with IPoIB-CM has very low performance (data not shown). NFS with IPoIB-UD has the best performance at 1400 miles, suggesting that IB unreliable datagram can benefit storage protocols on WAN besides MPI [19].

NFS by default uses a block size of 32KB. Larger block ksize can improve its performance. As of Linux version 2.6.25, the NFS/RDMA implementation in Linux has a hard-coded maximum block size of 32KB. The performance of IPoIB-UD imposes a limit to achieve better performance for NFS. So we only measured the performance of NFS on top of IPoIB-CM with varying block sizes. Fig. 4(a) shows the performance of NFS with different block sizes. With 1MB block size, the bandwidth of NFS reads is increased from 560MBytes/sec to 818MBytes/sec, but that of writes is only marginally increased. Note that, compared to IPoIB-UD, IPoIB-CM improves the performance of NFS significantly at 0.2 mile. This is because IPoIB-CM supports a connected IP mode in which it uses RDMA for data transfer. It has a much larger MTU



(a) NFS block size

(b) Connect-X vs. InfiniHost HCAs

Fig. 4. Other Factors of NFS Performance

(65520 bytes), compared to the 2044-byte MTU that is used in IPoIB-UD. As mentioned earlier, the connect-X HCA allows more concurrent RDMA read operations, which can be beneficial to NFS file writes. We also measured the performance of NFS file writes, comparing InfiniHost-III and connect-X HCAs. As shown in Fig. 4(b), connect-X indeed improves the performance of NFS writes to 650MBytes/sec at 0.2 mile.

2.4. iSCSI over RDMA

We measured the performance of iSCSI using a program that does block reads and writes through the bsg SCSI mid layer. Only one client and one server are used. Data in the iSCSI tests are contiguous with a size ranging from 4KB to 512KB. Fig. 5 shows the performance of ISCSI on top of RDMA, IPoIB-CM, and IPoIB-UD. At 0.2 mile, iSCSI over RDMA provides the best performance. With RDMA, iSCSI writes achieve lower bandwidth than iSCSI reads because of the use of RDMA reads in iSCSI data writes. Compared to IPoIB-UD, IPoIB-CM provides better performance for iSCSI for the reason as discussed for NFS, Furthermore, at 1400 miles, iSCSI with RDMA performs lower compared to iSCSI with IPoIB until the message size increases to 384KB or higher. In contrast to NFS, there is not much performance difference between iSCSI writes and reads at 1400 miles. This suggests that other factors have masked out the differences between reads and writes. This awaits more investigation.

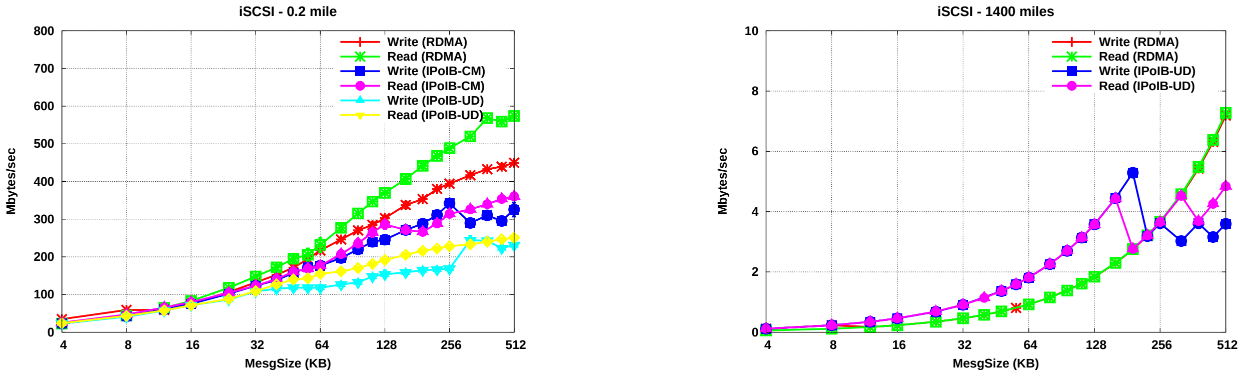
3. Related Work

High speed networks with direct access protocols such as RDMA first led to the development of a specification [16] to enable fast data transfer over RDMA-capable networks. A number of groups studied the benefits of leveraging RDMA for NFS performance improvements. Callaghan et. al. [3] then provided an initial implementation NFS over RDMA (NFSoRDMA) on Solaris. A team at the Ohio State University further completed the design and development of NFSoRDMA on Open Solaris for performance enhancement, compliant to IETF specification [4]. Talpey et. al. [17] recently announced the availability of initial Linux NFSoRDMA implementations.

Leveraging RDMA for iSCSI data transfer in storage also received significant interests both academia and industry [9, 5, 6]. Chadalapaka [5] articulated the architecture of iSER, and explored how iSER could be beneficial. The storage research team at Ohio Supercomputing Center released their open-source iSER initiators and targets. They also studied the performance of iSER and its integration into upper file system or object-based storage environment [6, 7]. In a comparative study, Radkov et. al. [12] investigated the performance difference between the file-based NFS protocol and the block-based iSCSI protocol. They found that aggressive meta-data caching can benefit the NFS protocol. Our work complements these efforts to provide a performance evaluation of NFSoRDMA and iSER on WAN. This presentation documents the effectiveness of existing approaches to utilizing RDMA for wide-area storage protocols.

4. Conclusions

We have shown the performance of RDMA-based storage protocols, NFS over RDMA and iSCSI over RDMA, on WAN. Because NFS presents fine grained 4KB data chunks to the RDMA-based transport protocol, and because RDMA is not able to handle discontiguous data, the bandwidth potential of RDMA is not fully utilized by NFS on WAN. In addition, RDMA reads



(a) iSCSI at 0.2 mile

(b) iSCSI at 1400 miles

Fig. 5. The Performance of iSCSI on WAN

limit the performance of both NFS and iSCSI. Our results document such performance behaviors of RDMA-based storage on WAN.

In the future, we intend to optimize the RDMA-based storage protocols. We plan to explore a couple of different ways, either by increasing the contiguity of data chunks before they are presented to RDMA, or by making use of the gather-send/receive-scatter mechanism on InfiniBand.

Acknowledgment

Weikuan Yu is grateful to the equipment support from Network Equipment Technologies (NET) Inc., and to the numerous technical discussions with Andrew DiSilvestre, Rich Erikson, and Brad Chalker from NET. Weikuan Yu is also very thankful to Susan Hicks and Dr. Philip Roth from Oak Ridge National Laboratory for their technical help.

References

[1] IOzone Filesystem Benchmark. In *http://www.iozone.org*.

[2] bbcp. http://www.slac.stanford.edu/∼abh/bbcp/.

[3] B. Callaghan, T. Lingutla-Raj, A. Chiu, P. Staubach, and O. Asad. NFS over RDMA. In *Proceedings of the ACM SIGCOMM workshop on Network-I/O convergence*, pages 196–208. ACM Press, 2003.

[4] B. Callaghan and T. Talpey. RDMA Transport for ONC RPC. http://www.ietf.org/internet-drafts/draft-ietf-nfsv4-rpcrdma-02.txt.

[5] M. Chadalapaka, H. Shah, U. Elzur, P. Thaler, and M. Ko. A study of iscsi extensions for rdma (iser). In *Proceedings of the ACM SIGCOMM workshop on Network-I/O convergence (NICELI)*, pages 209–219, 2003.

[6] D. Dalessandro, A. Devulapalli, and P. Wyckoff. iSER Storage Target for Object-Based Storage Devices. 0:107–113, 2007.

[7] D. Dalessandro, A. Devulapalli, and P. Wyckoff. Non-Contiguous I/O Support for Object-Based Storage. September 2008.

[8] Infiniband Trade Association. http://www.infinibandta.org.

[9] M. Ko, M. Chadalapaka, , U. Elzur, H. Shah, P. Thaler, and J. Hufferd. iSCSI Extensions for RDMA Specification. http://www.ietf.org/internet-drafts/draft-ietf-ips-iser-05.txt.

[10] Network Equipment Technologies, http://www.net.com.

[11] Obsidian Resarch Corporation, http://www.obsidianresearch.com/.

[12] P. Radkov. A Performance Comparision of NFS and iSCSI for IP-Networked Storage. In *FAST*, 2004.

[13] N. S. V. Rao, W. R. W. S. E. Hicks, S. W. Poole, F. A. Denap, S. M. Carter, and Q. Wu. Ultrascience net: High-performance network research test-bed. In *International Symposium on on Computer and Sensor Network Systems*, 2008.

[14] N. S. V. Rao, W. Yu, W. R. Wing, S. W. Poole, and J. Vetter. Wide-area performance profiling of 10gige and infiniband technologies. November 2008.

[15] A. Romanow and S. Bailey. An Overview of RDMA over IP. In *Proceedings of International Workshop on Protocols for Long-Distance Networks (PFLDnet2003)*, 2003.

[16] R. Srinivasan. RPC: Remote Procedure Call Protocol Specification Version 2. http://www.ietf.org/rfc/rfc1831.txt.

[17] T. Talpey et. al. NFS/RDMA ONC Transport. http://sourceforge.net/projects/nfs-rdma.

[18] GT 4.0 GridFTP. http://www.globus.org.

[19] W. Yu, N. Rao, and J. Vetter. Experimental analysis of infiniband transport services on WAN. In *International Conference on Networking, Architecture and Storage*. 2008.