Online Remote Data Backup for iSCSI-based Storage Systems

Dan Zhou, Li Ou, Xubin (Ben) He
Department of Electrical and Computer Engineering
Tennessee Technological University
Cookeville, TN 38505, USA
{dzhou21, lou21, Hexb}@tntech.edu

Stephen L. Scott
Computer Science & Mathematics Division
Oak Ridge National Laboratory
Oak Ridge, TN 37831, USA
scottsl@ornl.gov

Abstract

Data reliability is critical to many data sensitive applications, especially to the emerging storage over the network. In this paper, we have proposed a method to perform remote online data backup to improve reliability for iSCSI based networked storage systems. The basic idea is to apply the traditional RAID technology to iSCSI environment. Our first technique is to improve the reliability by mirroring data among several iSCSI targets, and second is to improve the reliability and performance by striping data and rotating parity over several iSCSI targets. Extensive measurement results using Iozone have shown that both techniques can provide comparable performance while improving reliability.

Key Words: iSCSI, RAID, Reliability, Online Backup

1. Introduction

With the increasing demand to deploy storage over the network [2,4], iSCSI [8] is a newly emerging technology with the goal of implementing the storage area networks (SAN) [9] technology over Internet infrastructure. Since iSCSI was proposed, much research has been sparked to protocol development [5], performance evaluation and analysis [1,3], and implementation [7], but there is not much research on iSCSI reliability. While it was observed that even though a server may be very robust and highly reliable, it is still vulnerable to an unexpected disaster such as a terrorist attack or a natural disaster. RAID (Redundant Array of Independent Disks) [6] is a known, mature technique to improve reliability of disk I/O through redundancy.

This paper introduces a method to perform remote online data backup and disaster recovery for iSCSI based storage systems by striping data among several iSCSI targets in a similar way to RAID. Since it's a distributed RAID across several iSCSI targets, we name it iSCSI RAID, or *iRAID* for short. The difference between *iRAID* and traditional RAID is that in traditional RAID, disk is the unit, while in *iRAID* each iSCSI target is a unit. Similar to traditional RAID, we may have different layouts/RAID levels. In this paper we only focus on two layouts: mirroring (*M-iRAID*) and rotated parity (*P-iRAID*). Both M-iRAID and P-iRAID improve the reliability and provide remote online data backup. It's remote because all iSCSI targets may be

physically on different locations, and it is online since they are connected through the Internet. M-iRAID achieves data backup by mirroring multiple iSCSI targets. If the primary target fails, data is still available from the backup iSCSI target. P-iRAID improves reliability by calculating and rotating a parity block for the iSCSI targets. In case one iSCSI target fails, data can be reconstructed from other n-1 iSCSI targets.

To quantitatively evaluate the performance potential of *iRAID* in real world network environment, we have implemented the prototype of *iRAID* and measured system performance. Extensive measurement results show that *M-iRAID* and *P-iRAID* demonstrate comparable performance while improving reliability.

The rest of the paper is organized as follows. Next section presents the design and implementation of *iRAID* including *M-iRAID* and *P-iRAID*, followed by our performance evaluation. We conclude our paper in Section 4.

2. Design of iSCSI RAID (iRAID)

We introduce iSCSI RAID, or *iRAID,* to solve the reliability problems of iSCSI storage systems. The basic idea of *iRAID* is to organize the iSCSI storage targets similar to RAID by using mirroring and rotated parity techniques. In *iRAID*, each iSCSI storage target is a basic storage unit in the array, and it serves as a storage node as shown in Figure 1. All the nodes in the array are connected to each other through a high-speed switch to form a local area network. *iRAID* provides a direct and immediate solution to boost iSCSI performance and improve reliability. Parallelism in *iRAID* leads to performance gain while using the RAID parity technique improves the reliability. This paper focuses on two *iRAID* configurations: mirroring *iRAID (*M-iRAID*)* and rotated parity *iRAID (P-iRAID).*

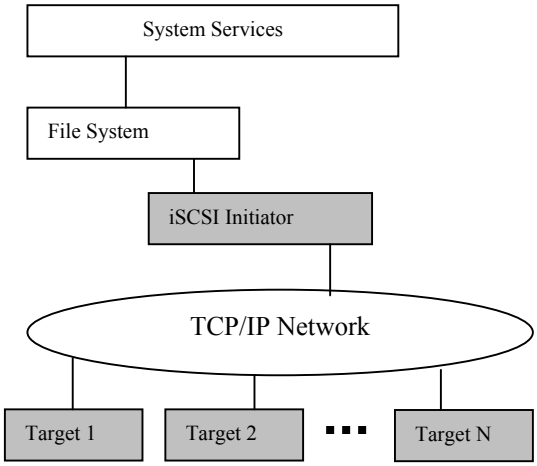

Figure 1: iRAID architecture. Data are striped among iSCSI targets.

2.1 M-iRAID

In the mirroring *iRAID* (M-*iRAID*), all data are striped and mirrored uniformly among all the *iRAID* nodes, which is illustrated in Figure 2. It borrows the concept from RAID level 1.

Figure 2 shows the data organization of each *iRAID* node for a *M-iRAID* system consisting of n iSCSI targets, where D_{ij} indicates that data block i on iSCSI target j.

2.2 P-iRAID

The *M-iRAID* increases the reliability of iSCSI through mirroring and provides remote data backup but also increases the cost since each data block is mirrored and stored in a backup iSCSI target. To improve the reliability at a lower cost, we introduce parity *iRAID* (*P-iRAID*) where in addition to data being striped and distributed among the *iSCSI* targets, a parity code for each data stripe is calculated and stored in an *iRAID* node. The parity block is rotated among the n iSCSI targets as shown in Figure 3, where the shadowed blocks are parity blocks, and others are data blocks. Each bit in a parity block is the XOR operation on the corresponding bits of the rest of the data blocks in each stripe. For example, $P_1 = D_{11} \otimes D_{12} \cdots \otimes D_{1,n-1}$.

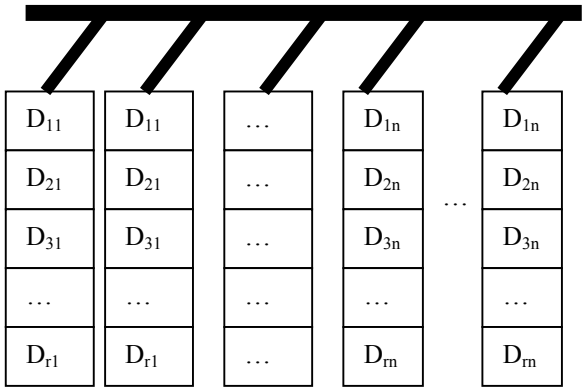

Figure 2: Data organization of M-iRAID.

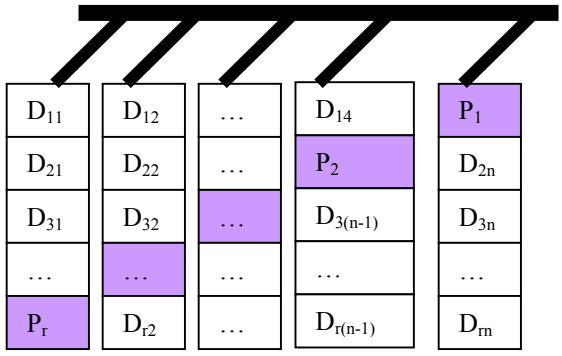

Figure 3: Data organization of P-iRAID.

3. Performance Evaluations

3.1 Experimental Setup

For the purpose of performance evaluation, we have implemented *iRAID* prototype (for both *M-iRAID* and *P-iRAID*) based on Linux software RAID and Intel iSCSI code [1]. Our experimental settings are shown in Figure 4. Six PCs are involved in our experiments – named *STAR1* through *STAR6*. *STAR1* serves as the iSCSI initiator, and *STAR2-5* are four iSCSI targets, which are organized as our *iRAID*. The data block size is set to 64KB, which is the default chunk size of Linux software RAID. All these machines are interconnected through a DELL PowerConnect 5012, 10-ports managed Gigabit Ethernet switch to form an isolated LAN. Each machine is running Linux kernel 2.4.18 with a 3COM 3C996B-T server network interface card (NIC) and an Adaptec 39160 high performance SCSI adaptor. *STAR6* is used to

[1] URL: *http://sourceforge.net/projects/intel-iscsi*

monitor the network traffic over the switch. The configurations of these machines are described in Table 1 and the characteristics of the disks are summarized in Table 2.

We use the popular file system benchmark tool, Iozone[2], to measure the performance. The benchmark tests file I/O performance for a wide range of operations. We will focus on performance of sequential read/write, random read/write because those are generally the primary concerns for any storage systems. The average throughput listed here is the arithmetic average of above four I/O operations. We run Iozone for different request size and data sets under various scenarios as follows:

Iozone –Ra –S *dataset size* –r *request size* –P –i0 –i1 –I 2 –f /mnt/iRAID/test

Where dataset size and request size are configurable. We reboot all machines after each round of test.

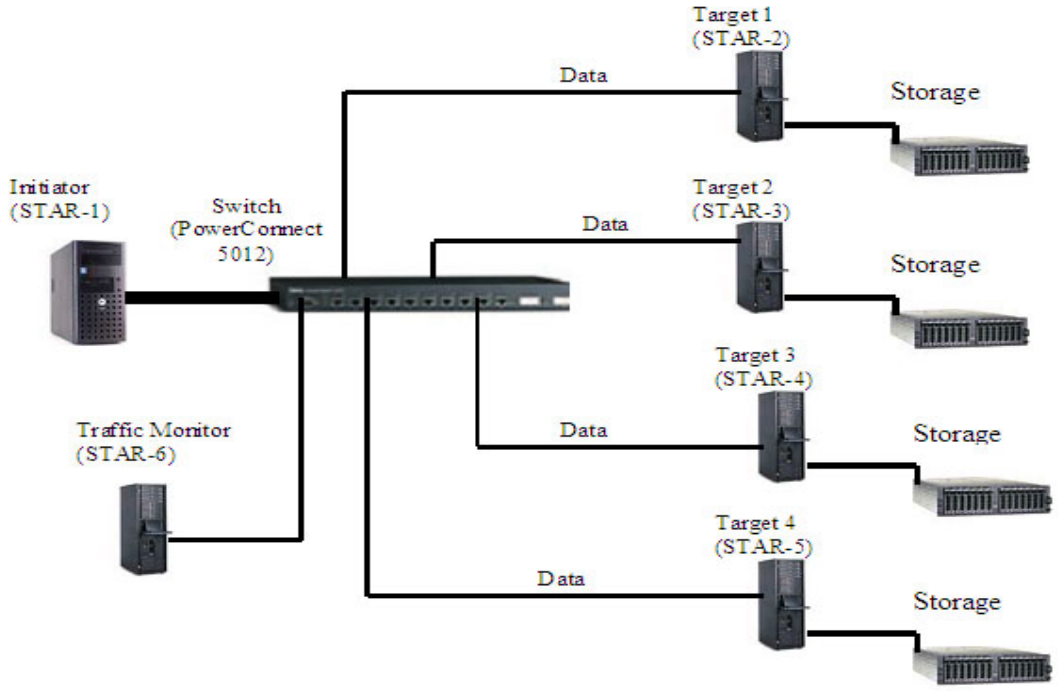

Figure 4:Experimental setup with 4 iSCSI targets and 1 initiator.

Table 1: Machines configurations

Machines	Processor	RAM	IDE disk	SCSI Controller	SCSI disk
STAR-1	PIII 1.4GHZ/512K Cache	1024MB	N/A	Adaptec 39160, Dell PERC RAID controller	4x Seagate ST318406LC
STAR2…5	P4 2.4GHZ/512K Cache	256MB	WDC WB400BB	Adaptec 39160	IBM Ultrastar 73LZX
STAR6	P4 2.4GHZ/512K Cache	256MB	WDC WB400BB	N/A	N/A

Table 2: Disk parameters

Disk Model	Interface	Capacity	Data buffer	RPM	Latency (ms)	Transfer rate (MB/s)	Average Seek time (ms)
ST318406LC	Ultra 160 SCSI	18GB	4MB	10000	2.99	63.2	5.6
Ultrastar 73LZX	Ultra 160 SCSI	18GB	4MB	10000	3	29.2-57.0	4.9
WB400BB	Ultra ATA	40GB	2MB	7200	4.2	33.3	9.9

3.2 Numerical Results

Our first experiment is to use Iozone to measure the I/O throughput for iSCSI without redundancy and *M-iRAID* with one backup iSCSI target under Gigabit Ethernet. The data set is 1G bytes and I/O request sizes range from 4KB to 64KB. Figure 5 shows the average throughputs for single iSCSI target without redundancy and two-targets M-iRAID. M-iRAID even shows better performance than single target iSCSI. The reason is that by mirroring the iSCSI target, the read performance is improved since the read requests may be satisfied by primary target and secondary target simultaneously.

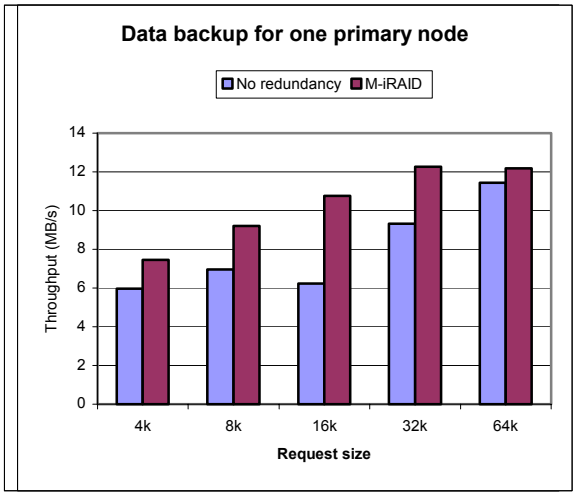

Figure 5: Using one target as backup

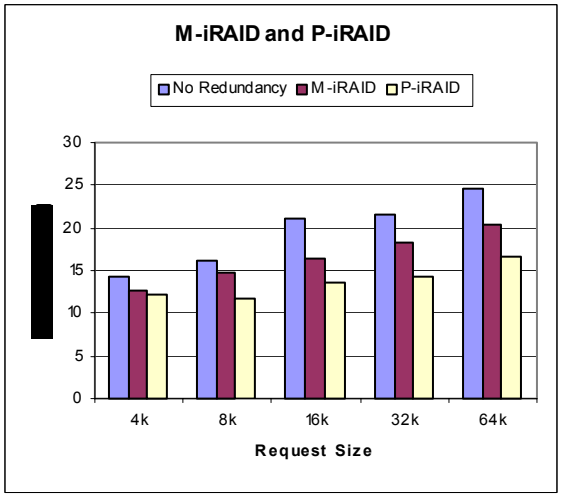

Figure 6: M-iRAID and P-iRAID

Our next experiment is using 4 iSCSI targets to construct a 4-node M-iRAID, where 2 targets (backup targets) are used to back up another 2 targets (primary targets). Similarly we construct a 4-node P-iRAID, where a parity block is calculated and rotated among the 4 iSCSI targets. Figure 6 shows the performance. It is obvious that M-iRAID performance is comparable to non-redundant configuration, while P-iRAID shows lower performance because of the small-write problem.

4. Conclusions

In this paper, we have proposed a method to perform remote online data backup to improve reliability for iSCSI based storage system. We have introduced *M-iRAID* to improve the reliability by mirroring data among several iSCSI targets, and introduced *P-iRAID* to improve the reliability and performance by striping data and rotating parity over several iSCSI targets. We have carried out prototype implementations of *M-iRAID* and *P-iRAID* under the Linux operating system. Extensive measurement results using Iozone have shown that *M-iRAID* and *P-iRAID* can demonstrate comparable performance while improving reliability, which is critical for disaster recovery.

Acknowledgments

This work is partially supported by Research Office under Faculty Research Award and Center for Manufacturing Research at Tennessee Technological University and the Mathematics, Information and Computational Sciences Office, Office of Advanced Scientific Computing Research, Office of Science, U. S. Department of Energy, under contract No. DE-AC05-00OR22725 with UT-Battelle, LLC. The authors would also like to thank the anonymous reviewers for a number of insightful and helpful suggestions.

References

[1] S. Aiken, D. Grunwald, A. Pleszkun, and J. Willeke, "A Performance Analysis of the iSCSI Protocol," *20th IEEE Conference on Mass storage Systems and Technologies*, 2003.

[2] E. Gabber, et al., "StarFish: highly-available block storage," *Proceedings of the FREENIX track of the 2003 USENIX Annual Technical Conference*, San Antonio, TX, June 2003, pp. 151-163.

[3] Y. Lu and D. Du, "Performance Study of iSCSI-based Storage Systems," *IEEE Communications*, Vol. 41, No. 8, 2003.

[4] R. V. Meter, G. G. Finn, S. Hotz. "VISA: Netstation's Virtual Internet SCSI Adapter." In *Proceedings of the 8th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS VIII)*, pp. 71-80, October 4-7,1998.

[5] K. Meth and J. Satran, "Features of the iSCSI Protocol," *IEEE Communications*, Vol. 41,No. 8, August 2003.

[6] D.A. Patterson, et al., "A Case for Redundant Arrays of Inexpensive Disks (RAID)," *ACM International Conference on Management of Data (SIGMOD)*, pp. 109-116, 1988.

[7] P. Sarkar, S. Uttamchandani, and K. Voruganti, "Storage Over IP: When Does Hardware Support Help?" *USENIX Conference on File And Storage Technologies*, 2003.

[8] J. Satran, et al. "iSCSI draft standard," *URL: http://www.ietf.org/internet-drafts/draft-ietf-ips-iscsi-20.txt*, Jan. 2003.

[9] P. Wang, et al.,"IP SAN-from iSCSI to IP-Addressable Ethernet Disks," *20th IEEE Conference on Mass storage Systems and Technologies*, 2003.