LBNL-52456 (2003)

The Global Unified Parallel File System (GUPFS) Project: FY 2003 Activities and Results

Gregory F. Butler (Project Lead),[1] William P. Baird,[1] Rei C. Lee,[1]
Craig E. Tull,[1] Michael L. Welcome,[2] Cary L. Whitney[1]

[1]NERSC Center Division
[2]Computational Research Division
Ernest Orlando Lawrence Berkeley National Laboratory
Berkeley, CA 94720

April 23, 2004

This work was supported by the Director, Office of Science, Office of Advanced Scientific Computer Research, Mathematical, Information, and Computational Sciences Division, of the U.S. Department of Energy under Contract No. DE-AC03-76SF00098.

Table of Contents

The Global Unified Parallel File System (GUPFS) Project: FY 2003 Activities and Results

1 Executive Summary

The Global Unified Parallel File System (GUPFS) project is a multiple-phase project at the National Energy Research Scientific Computing (NERSC) Center whose goal is to provide a scalable, high-performance, high-bandwidth, shared file system for all of the NERSC production computing and support systems. The primary purpose of the GUPFS project is to make the scientific users more productive as they conduct advanced scientific research at NERSC by simplifying the scientists' data management tasks and maximizing storage and data availability. This is to be accomplished through the use of a shared file system providing a unified file namespace, operating on consolidated shared storage that is accessible by all the NERSC production computing and support systems.

In order to successfully deploy a scalable high-performance shared file system with consolidated disk storage, three major emerging technologies must be brought together: (1) shared/cluster file systems software, (2) cost-effective, high-performance storage area network (SAN) fabrics, and (3) high-performance storage devices. Although they are evolving rapidly, these emerging technologies individually are not targeted towards the needs of scientific high-performance computing (HPC). The GUPFS project is in the process of assessing these emerging technologies to determine the best combination of solutions for a center-wide shared file system, to encourage the development of these technologies in directions needed for HPC, particularly at NERSC, and to then put them into service.

With the development of an evaluation methodology and benchmark suites, and with the updating of the GUPFS testbed system, the project did a substantial number of investigations and evaluations during FY 2003. The investigations and evaluations involved many vendors and products. From our evaluation of these products, we have found that most vendors and many of the products are more focused on the commercial market. Most vendors lack the understanding of, or do not have the resources to pay enough attention to, the needs of high-performance computing environments such as NERSC. (For a comprehensive discussion of our findings and conclusions, see Section 7 and Appendix F.)

Despite the limitations of existing products on the market, we have made good progress in several significant areas, and our successes span the three technology areas that the project is focused upon:

- Bridging between multiple fabric technologies is not only doable but also exhibits reasonably good performance. Successful bridging will allow us to create a hybrid fabric with multiple fabric technologies to support fabric connectivity for different existing and future system architectures, and at different price points for storage connections to various NERSC systems.
- iSCSI is a good option for storage traffic. A major strength of iSCSI is that it can be used on any fabric or interconnect that supports IP network traffic.
- InfiniBand is another good option for storage traffic. Although the technology is still relatively immature, InfiniBand offers performance (high bandwidth and low latency) beyond that of either Ethernet or Fibre Channel, with even higher bandwidths planned.

- After the first round of evaluations, three of the candidate file systems have been identified as being worth further testing: ADIC's StorNext File System, the Lustre File System , and IBM's GPFS File System. Three additional shared file systems have been identified as possible candidates and may be tested next year (IBM's TotalStorage SANFS, Panasas's ActiveScale Driect Flow, and the IBRIX file system).
- Sustained file system performance was achieved in excess of 1 gigabyte per second (GB/s) for several shared file systems.
    - A sustained read performance of 1140 MB/s was achieved with the ADIC StorNext file system.
    - A sustained read rate of 1484 MB/s was achieved with IBM's GPFS 1.3 file system

We have also uncovered several areas that are problematic and that we are continuing to investigate:

- Traditional storage devices do not scale in performance and are unsuited for use in a GUPFS production environment. These traditional storage devices, such as the Dot Hill device or the Silicon Gear device, are usually offered as direct-attach storage devices, connected directly to the hosts. These storage devices often have a limited number of interface ports, with limited aggregate bandwidth, and so they are not suitable for GUPFS.
- High-performance storage devices, with the ability to export logical unit numbers (LUNs) to all interfaces and to simultaneously drive all interfaces at their maximum capacity, are needed for block-based, shared file systems. These storage devices must be able to scale as the number of host connections increases. We have identified several promising storage devices with many of these characteristics, but we need to continue working with the vendors to improve their performance and functionality.
- The software iSCSI option performs very well, but at the expense of increased system overhead. The hardware assist iSCSI HBA cards are not cost effective at this time and have lackluster performance. Due to the recent emergence of the iSCSI standard, no high-performance, scalable iSCSI-connected storage devices currently are available. Better, scalable, and more extensive bridging capabilities between iSCSI and FC are needed.
- Fibre Channel switch interoperability between vendors is still a problem. The lack of support for the interoperability standards by some vendors makes it very difficult to create a SAN fabric using switches from multiple vendors.
- We face another challenge with the inter-switch link (ISL) — the interface between two Fibre Channel switches. ISL trunking[*] is often only supported between switches of the same brand. Without trunking, 2Gb/s ISL will only be able to support very limited bandwidth for inter-switch traffic. The ISL bandwidth problem may be alleviated once the 4 Gb/s ISL or 10Gb/s FC ISL becomes available.
- Although much progress has been made in the file system arena, more work needs to be done. The file system technology remains the component with the highest risk. Most file systems need better support in the areas of stability; parallel I/O and meta-data performance and functionality; scalability; DMAPI support for HPSS integration; and multiple-cluster, cross-platform/OS support.

[*] Trunking is a technique to group multiple links to achieve a better aggregate bandwidth.

Future investigations will address these areas as the focus of the GUPFS project shifts from primarily technology investigations to acquisition and deployment planning. We will perform fewer but more concentrated evaluations. Investigations and evaluations will be continued in all three critical component technology areas (shared file systems, storage, and SAN fabrics), but the main focus will be on multi-cluster, multi-fabric shared file system testing to prepare for Requests for Information (RFIs) and Requests for Proposals (RFPs), as well as the planned deployment of the GUPFS solution at NERSC. Activities will include networking, investigating HPSS and HSM integration, multi-cluster and cross platform installation and operations, and initial acquisition and deployment planning.

2 Introduction

The Global Unified Parallel File System (GUPFS) project is a multiple-phase project at the National Energy Research Scientific Computing (NERSC) Center whose goal is to provide a scalable, high-performance, high-bandwidth, shared file system for all the NERSC production computing and support systems. The primary purpose of the GUPFS project is to make the scientific users more productive as they conduct advanced scientific research at NERSC by simplifying the scientists' data management tasks and maximizing storage and data availability. This is to be accomplished through the use of a shared file system providing a unified file namespace, operating on consolidated shared storage that is accessed by all the NERSC production computing and support systems.

This report relates the activities and progress of the GUPFS project during its second year, FY 2003. It also presents the results of the evaluations conducted during FY 2003, as well as plans for near-term and longer-term investigations.

2.1 Evaluation Goals Overview

In order to successfully introduce a scalable, high-performance, parallel, shared file system and to consolidate storage, it is necessary to determine which file system, SAN fabric, and storage technologies will provide the best performance and reliability in the NERSC production environment. To this end, the GUPFS project conducted evaluations of these component technologies to determine which of the existing and emerging alternative technologies are most suitable to the NERSC environment.

In FY 2002 we developed our evaluation strategy and methodologies. Our goal has been to test and evaluate both the individual components and the collective environment in order to determine the correctness of the file system operations and data transfers; fault tolerances of all components, the ability of components to survive extensive stress testing; interoperability of components; performance of the file system and hardware components; and high-level functionality factors provided by the file system.

Also in FY 2002, we began developing the methodologies and benchmark mechanisms for conducting evaluations of each of the three technology components (file system, SAN fabric, and storage). Particular emphasis was placed on the techniques for evaluating the shared file systems, specifically benchmarks for parallel I/O and metadata operations. A discussion of these evaluation goals, methodologies, and mechanisms can be found in the GUPFS project's FY 2002 report, *The Global Unified Parallel File System (GUPFS) Project: FY 2002 Activities and Results* [1].

During FY 2003, we turned our emphasis to using the developed methodologies and benchmarks for evaluating the shared file systems, with specific attention on parallel I/O and metadata operations. Using the selected evaluation methodology, existing and newly developed benchmarks, and an upgraded testbed system, we evaluated several file system technologies with Fibre Channel attached storage. These evaluations, which are discussed in more detail in Section 3, were conducted after determining the baseline storage and SAN fabric performance. Determining this baseline performance is an important first step in determining which elements of the file system performance are due to the file system and which are due to hardware performance limitations.

Through this evaluation process, the methodologies and tools developed in FY 2002 were refined and revised. The evaluation methodologies and benchmark codes are discussed in Section 3.1, “Evaluation Methodology and Baseline Configuration.”

By conducting technology evaluations, we expect that the best and most appropriate combinations of file systems, SAN fabric technologies, and storage technologies can be identified to ensure a successful GUPFS deployment. We will communicate any deficiencies identified during the evaluations to the appropriate vendors so that they can be fixed.

2.2 FY 2003 Activity and Analysis Overview

During the second year of the GUPFS project, we continued to focus on identifying, testing, and evaluating existing and emerging shared/cluster file system, SAN fabric, and storage technologies; and developing appropriate benchmarking methodologies and codes for a parallel environment.

During FY 2003, we conducted a number of significant activities:

- Demonstrated sustained file system performance in excess of 1 gigabyte per second (GB/s) for several shared file systems.
- Wrote and published a technical report, *The Global Unified Parallel File System (GUPFS) Project: FY 2002 Activities and Results*, LBNL-52456. Reconfigured the GUPFS testbed to better test file system, SAN fabric, and storage technologies in a scientific parallel computational cluster environment.
- Identified and tracked existing and emerging file system, SAN fabric, and storage technologies and technology trends.
- Continued to develop and revise testing methodologies and benchmarks for shared/cluster file systems in a parallel environment.
- Obtained baseline storage performance characteristics and conducted testing of several file system and fabric technologies.
- Further expanded the testbed to extend scalability testing capabilities and to investigate new component technologies.
- Presented the GUPFS project purpose, goals, and plan to DOE and other organizations
- Developed new and continued existing relationships with file system, SAN fabric, and storage vendors, as well as other laboratories and possible collaborators

The following observations in the GUPFS target technology areas are based on our evaluation of many technologies and products examined during FY 2003:

General

- Many vendors are involved and have various products available. However, most vendors are more focused on the commercial market.
- There is a need for standardized, centralized monitoring and management for fabric and storage.

File System Technologies

- Progress has been made, but more work is needed. The file system technology remains the component with the highest risk. Areas where more work is needed include:
    - Stability
    - Parallel I/O and metadata performance and functionality
    - Scalability
- Not enough file systems can or will support multiple platforms and operating systems
    - At present, ADIC StorNext File System (SNFS) has the greatest variety
    - The IBM SAN File System (also known as StorageTank) supported several platforms/systems when initially released, and there are plans to support more
    - IBM General Parallel File System (GPFS) supports only AIX and Linux
- File system vendors should open source their client software to assist in wide-scale adoption
    - StorageTank Linux client and Lustre are already open source
    - Open sourcing is under consideration by others

Fabric Technologies

- Better and more extensive bridging capabilities are needed
- Better inter-switch links and higher aggregate bandwidth are needed
- Policy-driven quality of service (QoS) capabilities for all fabrics are needed

Storage Technologies

- Storage devices with multiple types of interfaces and with a large number of ports are desired
- Devices with higher aggregate and individual port bandwidth are needed
- Storage that supports a very large number of initiators still needs improvement

The following sections of this report present the technology evaluations that led to the observations listed above, as well as our findings and our future plans.

3 FY 2003 Technology Evaluations

During the past year, we have evaluated a number of products and technologies that we believe are key technologies to the GUPFS Project. These evaluations were focused on the three component technology areas critical to the deployment of a center-wide shared file system: shared file systems, storage, and SAN fabrics. We will present testing results in subsequent sections for some of the following products and technologies evaluated:

- File Systems: Sistina GFS [2] 5.1 and 5.2 Beta; ADIC StorNext (CVFS) File System [3] 2.0 and 2.1; Lustre 0.6 [4] (1.0 Beta 1); and GPFS [5] 1.3 for Linux
- Fabric Technologies
  - Fibre Channel Switches: Brocade SilkWorm 3800 [6],and Qlogic SANbox2-16 [7] and SANbox2-64
  - iSCSI [8]: Cisco SN 5428 [9], Intel iSCSI host bus adapters (HBAs) [10], iSCSI over IB
  - InfiniBand: InfiniCon [11] and Topspin [12], with IB to Fibre Channel (FC) and Gigabit Ethernet bridges
  - Inter-connect: Myrinet, Gigabit Ethernet
- Fibre Channel Storage Devices
  - 1 Gb/s FC: Dot Hill, Silicon Gear, Chaparral
  - 2 Gb/s FC: Yotta Yotta NetStorager GSX 2400 [13], EMC CX 600 [14], 3PARdata [15]

The remainder of this section will cover the evaluation methodology and baseline configuration used to conduct the evaluations of the three technologies. In Section 4, we will present the storage technology evaluations and results. In Sections 5, we will present the fabric technology evaluations and results, and in Section 6 we will present the file system evaluations and results. Section 7 is our summary findings, future directions, and conclusion.

3.1 Evaluation Methodology and Baseline Configuration

The GUPFS project uses a testbed system to conduct investigations and evaluations of the component technologies needed for a center-wide shared file system, and to explore these components' interactions. The evaluations were conducted in accordance with the benchmarking methodology developed during the previous year. This methodology is detailed in the GUPFS Project FY 2002 Report [1].

The GUPFS testbed system presented a microcosm of a parallel scientific cluster — dedicated computational nodes, special-function service nodes, and a high-speed interconnect for message passing. It used an internal jumbo frame Gigabit Ethernet as the primary high-speed message passing interconnect. An internal 10/100 Mb/s Fast Ethernet LAN was employed for system management and NFS distribution of the user home file systems. The testbed supplied Fibre Channel as the base SAN fabric, as well as Fibre Channel storage, and a variety of alternative fabrics and bridges between these fabrics.

The testbed was configured as a Linux parallel scientific cluster, with a management node, a core set of 16 dedicated dual Pentium-4 compute nodes, a set of six special-purpose dual Pentium-4 nodes, and a reserve of five auxiliary dual Pentium-3 compute nodes. The Fibre Channel SAN fabric was based on two 16-port 2 Gb/s FC switches. The Gigabit Ethernet fabric was used as the message passing interconnect for the parallel benchmarks.

The base configuration of the testbed during the evaluations included the following major components:

System nodes

- Twenty-two dual Xeon Pentium-4 nodes
- Six dual Pentium-3 nodes

Fabric

- Ethernet
  - One 32-port Extreme 7i Gigabit Ethernet switch
  - One 16-port Extreme 5i Gigabit Ethernet switch
  - Two 10/100 Ethernet switches for system management
- Fibre Channel
  - Two 16-port 2 Gb/s Fibre Channel Switches (Brocade SilkWorm 3800 and Qlogic SANbox2-16)
  - One 16-port 1 Gb/s Fibre Channel Switch (Brocade SilkWorm 2800)
  - One Cisco SN5428 iSCSI Router fabric bridge to Ethernet
- Myrinet
  - One Myrinet 2000 8-port switch with eight host interface cards
- InfiniBand
  - One InfiniCon ISIS InfinIO 7000 1x InfiniBand switch, with eight 1x HCA host adapters, and fabric bridge modules for Fibre Channel and Gigabit Ethernet

Storage

- An EMC CLARiiON CX600 disk subsystem
- A Dot Hill 7124 RAID disk subsystem
- A Silicon Gear Mercury II RAID subsystem
- A Chaparral A8526 RAID subsystem with attached storage

The systems nodes had the following hardware configuration:

- All Pentium-4 nodes had the same base configuration. This configuration consisted of the same motherboards, dual 2.2 GHz Pentium IV Prestonia Xeon CPUs, 2 GB of DDR memory, 10/100 and Gigabit Ethernet interfaces, 36 GB SCSI disks, and Qlogic 2340 2 Gb/s Fibre Channel HBAs.
- All Pentium-3 nodes had the same base configuration. This consisted of identical motherboards, dual Pentium III 1 GHz CPUs, 1 GB of memory, 10/100 and Gigabit Ethernet interfaces, and 18 GB SCSI disks. One of the Pentium-3 nodes was configured as a management node and had additional 10/100 and Gigabit Ethernet interfaces. The remaining five Pentium-3 nodes were configured as auxiliary compute nodes with Qlogic 2310 2 Gb/s Fibre Channel HBAs

In addition to the testbed nodes being attached to the Fibre Channel SAN fabric, the various storage devices, both permanent and under evaluation, were attached to the same switched FC fabric, as were a number of fabric bridges. These fabric bridges included the Cisco SN5428 Storage Router bridging between Gigabit Ethernet iSCSI storage traffic from hosts and FC fabric attached storage devices, and the InfiniCon InfinIO fabric bridge between InfiniBand attached hosts and FC fabric attached storage.

Each of the RAID controllers had two or more Fibre Channel ports for connecting to the switch, and these FC ports could be used simultaneously. All four storage devices supported various RAID configurations and utilized similar 10,000 RPM 73 GB disk drives. The DotHill contained 20 such drives, the Silicon Gear 12 drives, the Chaparral 10 drives, and the EMC 30 drives. Total unformatted SAN-attached storage capacity was approximately 5.3 terabytes (TB), with a nominal maximum of 4.2 TB of formatted RAID 5 storage. The DotHill and Silicon Gear were both limited to 1 Gb/s FC interfaces, while the Chaparral and the EMC supported 2 Gb/s interfaces.

The software configuration of the testbed was that of a Linux-based parallel scientific system. All testbed nodes ran Linux, based on the RedHat 7.1, 7.2, 7.3, 8.0, or 9.0 distribution, depending on the requirements of the file system tested. The testbed utilized MPICH as the MPI implementation for parallel jobs, with Portland Group C, C++, and various types of FORTRAN compilers providing the compilation and execution environment for the parallel jobs.

A detailed discussion of the testbed hardware and software configurations and details of the hardware configurations of the Pentium-3 and Pentium-4 nodes are presented in Appendix A. A detailed discussion of the updates made to the testbed configuration for the next year’s activities is presented in Appendix B.

The hardware and software configuration described above was the baseline configuration used to conduct the technology evaluations and obtain the results presented in the next three sections.

4 Storage Technology Evaluations and Results

During FY 2003, more storage vendors started to introduce storage systems supporting 2 Gb/s FC ports and an increased number of ports for SAN connectivity. Some of these storage systems include EMC's CX 600, Yotta Yotta's NetStorager GSX 2400, 3PARdata's S400 and S800, and DataDirect Networks' S2A 8500. All of these systems can support eight or more 2 Gb/s FC ports, with a possible 16 Gb/s or more peak aggregate performance.

In FY 2003, we tested several of these new storage systems, including the newly acquired CX 600 and an S400 evaluation system from 3PARdata. We also continued our evaluation of the Yotta Yotta NetStorager GSX 2400 system.

In this section, we present a high-level summary of the storage technology evaluations we conducted during FY 2003. These include evaluations of the EMC CX 600, Yotta Yotta NetStorager, and 3PARdata 2 Gb/s FC storage devices. For the purpose of comparison, we also include an evaluation of 1 Gb/s Fibre Channel storage devices.

4.1 1 Gb/s FC Storage Performance

In order to provide a baseline for understanding the performance and scalability characteristics of the 2 Gb/s Fibre Channel storage devices, we conducted an evaluation of the 1 Gb/s FC storage devices that were already connected to the testbed. At the beginning of FY 2003, the GUPFS testbed had a Fibre Channel SAN with the following 1 Gb/s storage devices:

- A dual-controller DotHill 7124 RAID subsystem, with an expansion cabinet
- A dual-controller Silicon Gear Mercury II RAID subsystem
- A single-controller Chaparral A8526 RAID subsystem with attached storage

Each of the RAID controllers had dual Fibre Channel ports for connecting to the switch, both of which could be used simultaneously. All three storage devices supported RAID 0, 1, 3, and 5 configurations and used similar 10,000 RPM 73 GB U160 SCSI or FC-AL disk drives. The DotHill contained 20 drives, the Silicon Gear 12 drives, and the Chaparral 10 drives. Total unformatted SAN-attached storage capacity was approximately 3.1 TB, with a nominal maximum of 2.7 TB of formatted RAID 5 storage.

These storage devices were originally selected when the project was initially targeted toward early evaluations of the Sistina Global File System (GFS) system, which at that time utilized the novel and promising Device Memory Export Protocol (DMEP) SCSI extension in its distributed lock management. However, Sistina has since moved away from DMEP and has adopted a new IP-based Global Universal Lock Manager (GULM) implementation. Without the DMEP support, these storage devices are no different from any other storage devices and therefore can only be used for file system and fabric technology evaluations. We believe these storage devices are representative of traditional 1 Gb/s RAID disk storage systems. The MPTIO benchmark was run to obtain the baseline performance of these 1Gb/s storage devices.

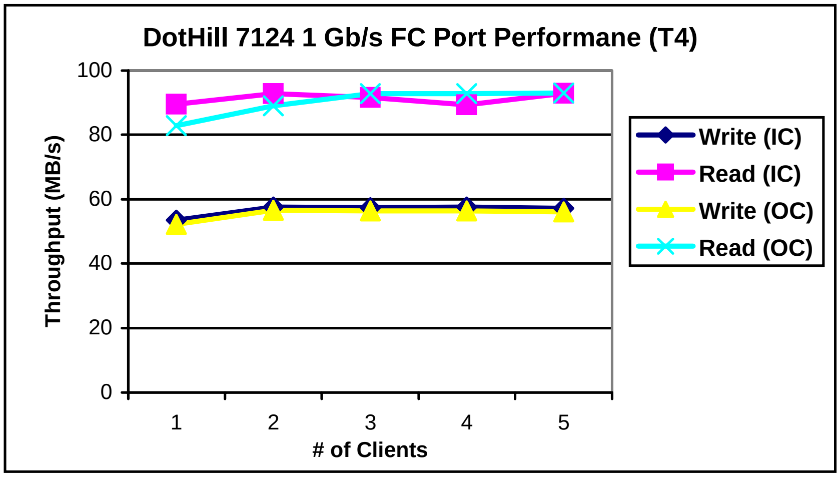


Figure 1. DotHill 1 Gb/s FC performance: in-cache (IC) vs. out-of-cache (OC).

Figure 1 shows the MPTIO results of the DotHill storage device, specifically DotHill's 1 Gb/s FC port performance. The results indicate very little performance scalability on the DotHill storage. The in-cache (IC) performance was about the same as the out-of-cache (OC) performance, and the best performance was about 57 MB/s for writes and 93 MB/s for reads. The MPTIO tests were also run on the Silicon Gear and Chaparral storage devices, and the results also indicated that these 1 Gb/s storage devices had no performance scalability when the number of clients was increased. The DotHill results are representative of the performance and scalability of these storage devices.

The lack of scalability makes these 1 Gb/s devices of very little use for file system evaluation; therefore, we have determined that they are unsuitable for future GUPFS deployment. These 1 Gb/s FC storage devices are now mainly being used for test development and/or the functionality evaluation of file system and fabric technologies.

4.2 EMC CX 600 Performance

The EMC CX 600 was added to the GUPFS testbed in FY 2003. Figure 2 shows performance of a single EMC CX 600 2 Gb/s FC port, using the PIORAW test with four processes accessing files of different file sizes (FS) and using different I/O sizes.

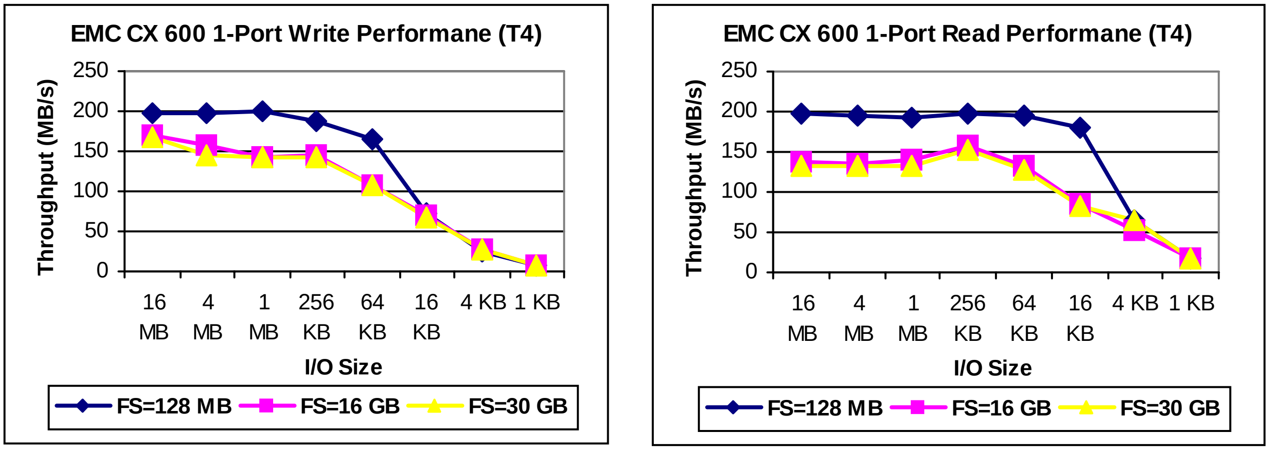


Figure 2. CX 600 2 Gb/s performance: Small File (FS = 128 MB) vs. Large File (FS = 16 GB or 30 GB).

Three file sizes were selected to study how I/O performance may be affected by the cache in the CX 600 controllers. The FS = 128 MB results show CX 600 performance for in-cache reads and writes. For I/O sizes larger than 256 KB, the small file (FS = 128 MB), in-cache read/write performance is about 200 MB/s. The results indicate that the CX 600 controller is capable of sustaining I/O throughput that is very close to what a 2 GB/s FC port can sustain. However, the read/write performance for the large file tests (FS = 16 GB and 30GB) was only about 150 MB/s, for I/O sizes larger than 256 KB. This seems to indicate that, for writes, the controller was not able to flush the data to the backend disks fast enough to match the writes to its front-end FC ports, and for reads, the controller read-ahead rate was not fast enough to match the front-end reads. This may indicate that the CX 600 storage has a very limited backend disk performance.

Figure 3 shows the best in-cache (IC) and out-of-cache (OC) I/O performance scalability on the EMC CX 600 system, scaling from one client to four clients using the PIORAW benchmark, with 1 MB I/O size. The file size for the IC tests was 128 MB and for the OC tests was 4,231 MB, using four individual RAID-5 volumes. Each RAID-5 volume was created using five 73-GB disks.

The I/O scalability of the EMC CX 600 was very disappointing. Except for the in-cache (IC) reads, the CX 600 box showed very poor scalability. Even for the in-cache reads, the performance was limited at about 600 MB/s. For out-of-cache I/Os, the performance was much lower, at about 200 MB/s. All these results make the EMC CX 600 and storage with similar architecture less attractive as a possible choice for the underlying storage of shared file systems.

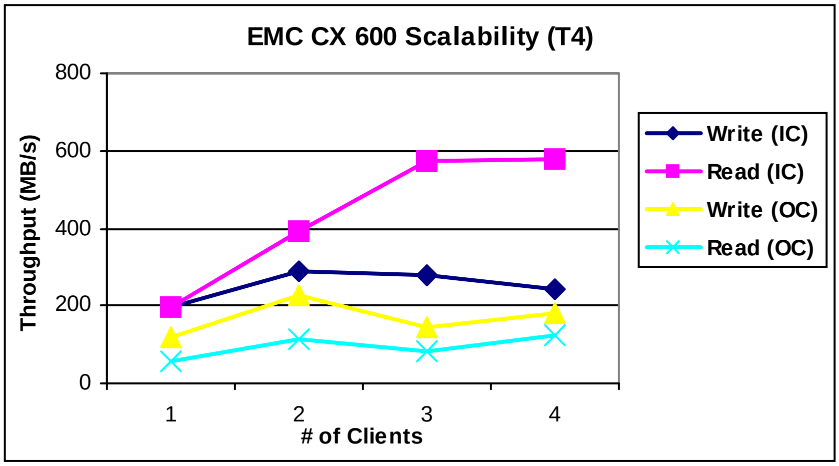


Figure 3. EMC CX 600 Scalability: in-cache (IC) vs. out-of-cache (OC).

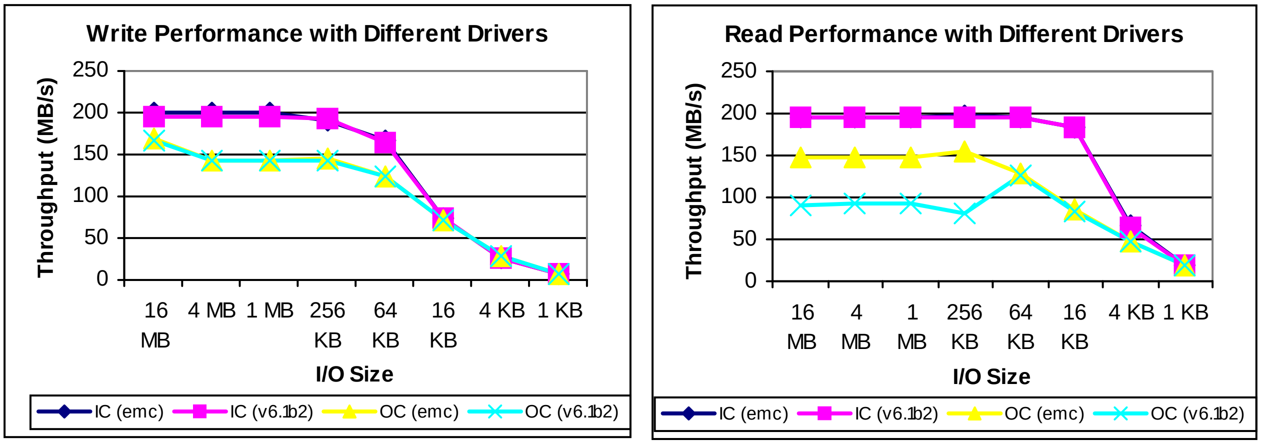


Figure 4. EMC CX 600 performance with different Qlogic drivers.

We have also noticed some performance differences for out-of-cache reads when different Qlogic drivers were used with the CX 600 system. The v6.1b2 release of the Qlogic driver has been the default driver used in our testbed environment for I/O benchmarks. However, EMC also shipped a modified Qlogic driver with the CX 600 system. Figure 4 shows in-cache (IC) and out-of-cache (OC) performance over a single 2 Gb/s FC port using different Qlogic drivers.

The results show no performance difference for writes when different drivers were used. However, for OC reads with larger I/O sizes, the read performance was lower (92 MB/s vs. 143 MB/s) with the default Qlogic v6.1b2 driver. This demonstrates the importance of determining the baseline storage performance and picking good device drivers before starting any file system and fabric evaluations, since the I/O performance of the same storage device may be different with different device drivers.

4.3 Yotta Yotta NetStorager Performance

Since May 2002, we have been beta testing YottaYotta's NetStorager system. The evaluation has been very successful. The NetStorager has improved substantially and is now delivering very good scalable performance. In 2003, a Yotta Yotta NetStorager GSX 2400 system was added to the GUPFS testbed to facilitate the evaluation of file system and fabric technologies.

Figure 5 shows the best in-cache (IC) and out-of-cache (OC) I/O performance scalability on the Yotta Yotta's GSX 2400 NetStorager system, scaling from one client to sixteen clients using the MPTIO benchmark, with 1 MB I/O size. The file size for the IC tests was 2,048 MB and for the OC tests was 32,768 MB, using a single shared 1-TB RAID-0 volume over 32 disks.

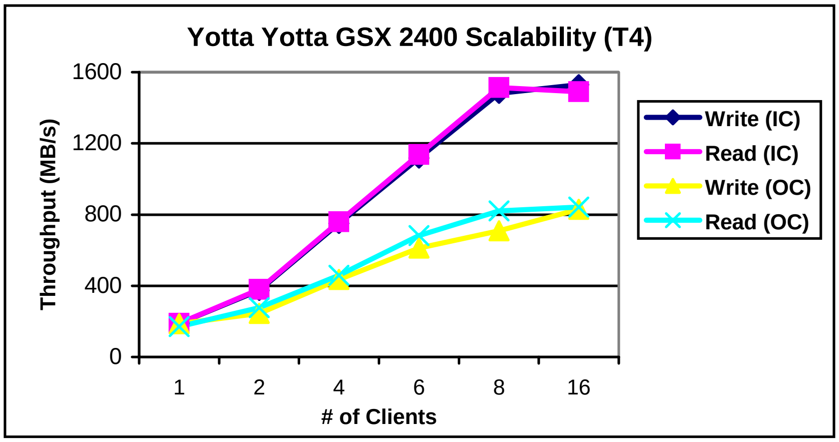


Figure 5. Yotta Yotta GSX 2400 performance.

The overall performance of the GSX 2400 was very good when it was used as the underlying storage for the evaluation of file system and fabric technologies. Being able to support shared access to the same LUNs from multiple ports makes GSX 2400 a viable storage technology for GUPFS. However, the lack of support for any RAID implementation except RAID-0 (striping) makes GSX 2400 less attractive. Another weakness of GSX 2400 is its slow GUI and command line interface (CLI) — it often takes several seconds for each step to complete. For operations (e.g., defining a new LUN) that require multiple steps, its slowness can be very annoying.

Storage devices like the GSX 2400 (which can provide acceptable scalability while exporting the same LUN through multiple interfaces and can also approach the aggregate performance of their interfaces) are beginning to appear. These represent a different class of storage devices from the traditional 1 and 2 Gb/s devices and the middle tier CX600. We believe this new class of storage devices is what will be required for the deployment of a block-based shared file system solution for GUPFS. Other types of shared file system solutions, although not requiring such storage devices, would benefit from their high performance and their ability to effectively service multiple hosts. More details about the GSX 2400 performance are presented in Appendix C.

4.4 3PAR S400 Performance

A single 3PAR system can be configured as a cluster of two to eight controller nodes. Each node can scale from four to twenty-four 2-Gb/s full-bandwidth Fibre Channel ports. Controller nodes connect to drive chassis and to hosts via Fibre Channel, and to each other via a high-bandwidth, low-latency backplane. Each 3PAR system features a full-mesh, passive system backplane that provides a dedicated 1 GB/s link between each controller node. This low-latency backplane enables tight and rapid coordination among the controller nodes, allowing them to form a single, cache-coherent system. As a result, all volumes can be exported by any or all controller nodes simultaneously and coherently. In the event of a node failure, its work is transferred to another node in the cluster.

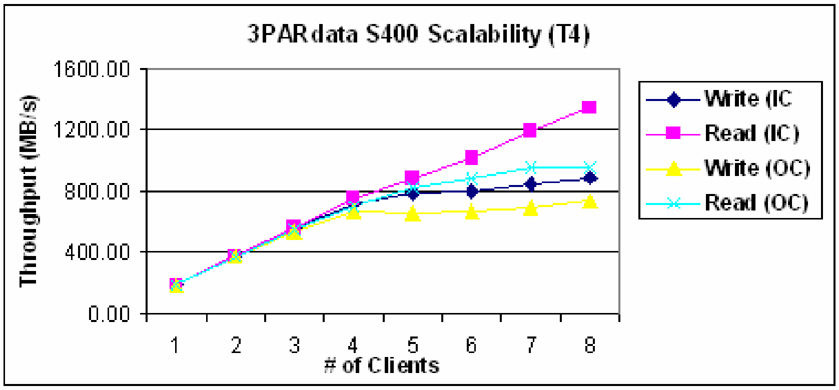


Figure 6. 3PARdata S400 performance.

Figure 6 shows the best in-cache (IC) and out-of-cache (OC) I/O scalability on the 3PAR S400 storage system, scaling from one to eight clients using the MPTIO benchmark, with 1 MB I/O size. The file size used for the IC tests was 2,100 MB and for the OC tests was 50,400 MB, using a single shared 1-TB RAID-5 volume.

Similar to the Yotta Yotta GSX 2400 storage, the 3PARdata S400 storage also shows very good I/O scalability. Both Yotta Yotta GSX 2400 and 3PARdata S400 are far better storage systems than the traditional 1 GB/s storage or EMC CX 600 when used as the underlying storage for the shared file systems like GUPFS.

If we compare the 3PARdata results with the YottaYotta results in the previous section, both storage systems show very good scalability for in-cache (IC) reads and writes. For in-cache I/Os, with the exception of 3PARdata for in-cache writes, both systems are capable of sustain an I/O throughput higher than 1 GB/s (or 8 Gb/s or 1000 MB/s). For out-of-cache (OC) reads and writes, the performance on both storage systems seemed to be limited by their bandwidth to their backend disks, which was about 800 MB/s. The reason that 3PARdata was slower for writes may be because the LUN tested was defined as a RAID-5 volume while, in the Yotta Yotta case, the LUN was a RAID-0 volume. A more complete report on the evaluation of 3PARdata S400 can be found in Appendix D.

4.5 Storage Performance and Scalability

Storage can be a performance bottleneck in any file system. While a storage device may be able to sustain very good single-port performance, having good single-port performance is not sufficient for a shared disk file system like GUPFS. For GUPFS, the underlying storage devices must also demonstrate a very good scalability when the number of clients increases (to thousands or tens of thousands). A shared file system will not scale if the underlying storage does not scale.

Figure 7 shows how storage devices scale when the number of clients increases. The figure gives the results of four storage devices: Yotta Yotta GSX 2400 (YottaYotta), 3PARdata S400 (3PARdata), Silicon Gear Mercury II (SilconGear), and DotHill SANnet (DotHill). Both Silicon Gear and Dot Hill have only two 1 Gb/s front-end ports. Yotta Yotta has eight 2 Gb/s ports, and 3PARdata has sixteen 2 Gb/s ports (although only eight were used during the test). On each storage system, with the exception of Yotta Yotta, which only supports RAID-0 volumes, we created a single RAID-5 volume for shared access by multiple clients.

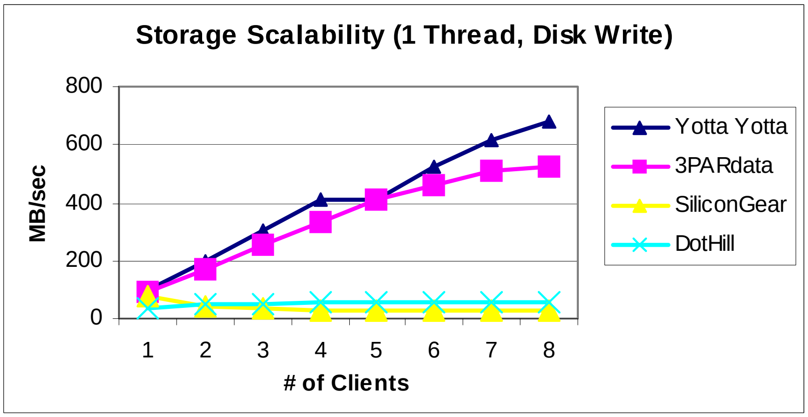


Figure 7. Storage scalability.

The figure indicates that both the Silicon Gear and Dot Hill devices did not scale when the number of clients increased. Silicon Gear performance actually dropped when the number of clients increased. On the other hand, both the Yotta Yotta storage and the 3PARdata storage did scale very well when the number of clients increased.

Figure 8 shows the aggregate performance of the four storage devices: DotHill, Silicon Gear, Yotta Yotta, and 3PARdata, using the MPTIO benchmark with different test conditions. The results indicate that the Yotta Yotta and 3PARdata storage systems will be able to sustain higher performance than Silicon Gear or Dot Hill in a shared file system. These performance results indicate that Yotta Yotta and 3PARdata, and storage systems with similar performance characteristics would potentially be the choices of the underlying storage for shared file systems.

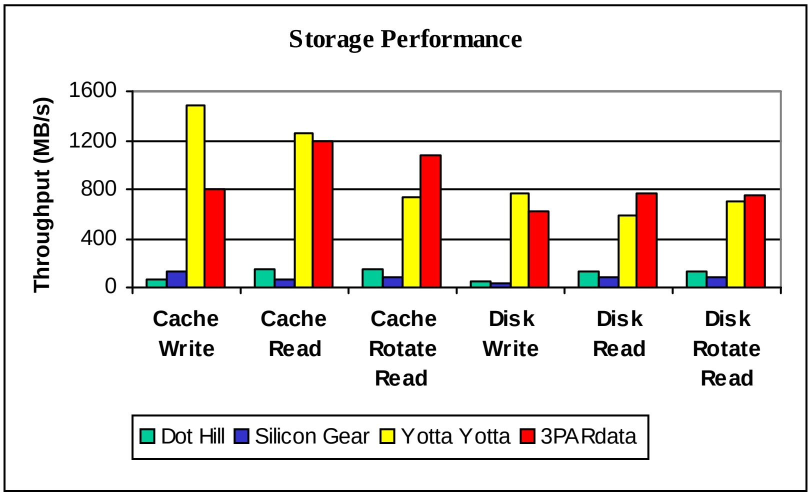


Figure 8. Storage aggregate performance.

4.6 Summary

The results of the storage evaluations we conducted indicated that the performance of traditional 1 Gb/s Fibre Channel storage devices is limited and not scalable, and that such storage systems are unsuitable for use in the future GUPFS deployment. Our testing of the EMC CX 600 showed that this 2 Gb/s storage device, which is representative of the evolution of the traditional 1 Gb/s storage devices with some added middle-tier functionality, has poor scalability as the number of clients increases, except in the case of in-cache writes. We believe this is due to an architectural limitation caused by an imbalance between the bandwidth available to host connections and the device's back-end disks, which its caching algorithms were unable to overcome. Given the limited performance and poor scalability of the EMC CX 600, we have determined that it is unsuitable for use in the GUPFS production environment, as are the related traditional middle-tier 2 Gb/s FC storage devices.

Both the Yotta Yotta GSX 2400 and 3PARdata S400 show good I/O performance and scalability. They represent a new class of storage — they can provide acceptable scalability in performance, they are capable of exporting the same storage to many hosts through multiple interfaces, and their performance approaches the aggregate performance of their interfaces. We believe this new class of storage devices is what will be required for the successful deployment of a GUPFS solution, including block-based shared file system solutions.

With the appearance of the scalable storage devices, storage technologies are generally on track to be ready for the GUPFS deployment.

5 SAN Fabric Technology Evaluations and Results

Storage area networks (SANs), by providing a high-performance network fabric oriented toward storage device transfer protocols, allow direct physical data transfers between hosts and storage devices. Currently, most SANs are implemented using Fibre Channel (FC) protocol-based fabric. Emerging alternative SAN protocols, such as iSCSI (Internet Small Computer System Interface), FCIP (Fibre Channel over IP), and SRP (SCSI RDMA [Remote Direct Memory Access] Protocol), are enabling the use of alternative fabric technologies, such as Gigabit Ethernet and the InfiniBand, as SAN fabrics.

One of the major GUPFS goals in FY 2003 was to evaluate the performance characteristics of SAN fabric technologies and the interoperability of these technologies in a hybrid, multiple-fabric environment [16,17].The ability of the file systems to operate in such a mixed environment is very important to the ultimate success of the GUPFS project.

A picture of the current GUPFS fabric configuration is shown in Figure 9.

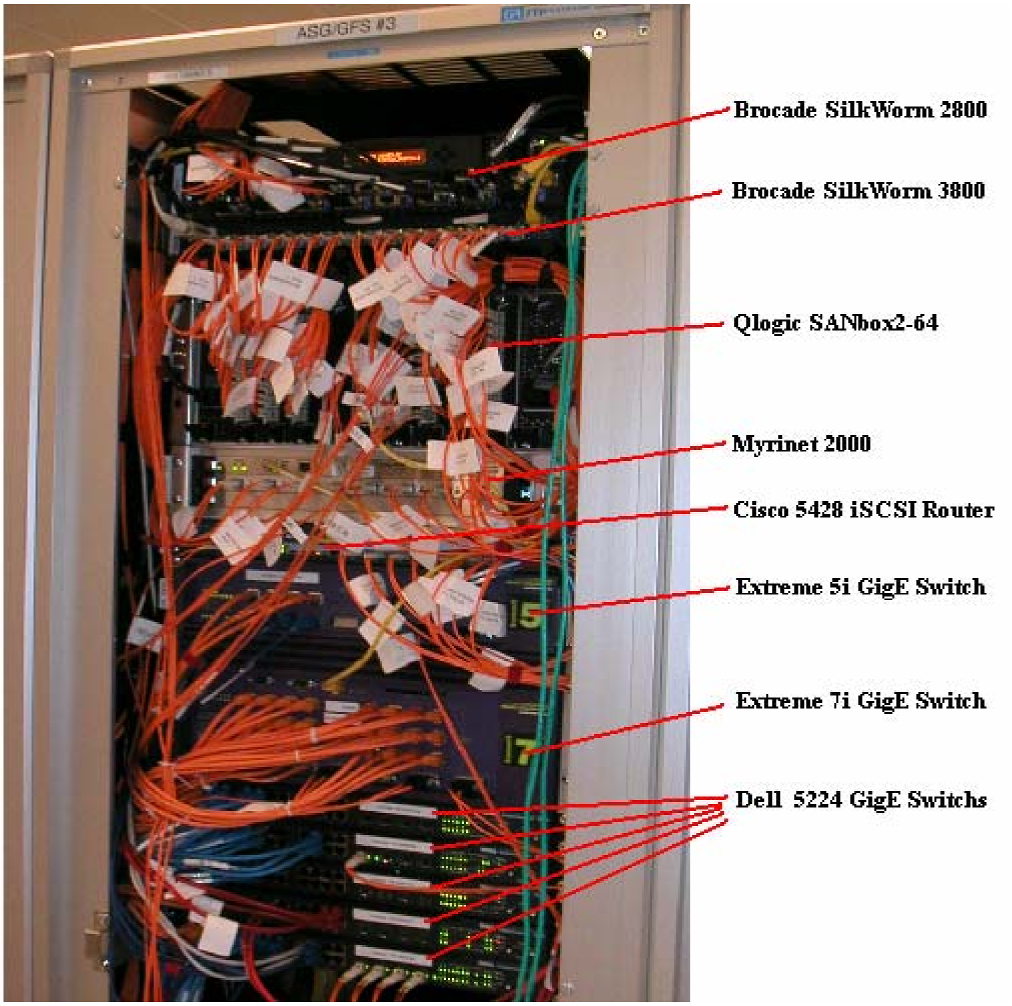


Figure 9. The GUPFS fabric configuration.

5.1 2 Gb Fibre Channel Switch

The 2 Gb/s Fibre Channel (FC) technology has matured and become the standard product offering by most FC storage and switch vendors. Many FC vendors have introduced new switch products with higher port connectivity. The increase in port connectivity allows more clients and storage to be connected to the same switch, which allows fewer cascading switches to be used in the deployment and substantially simplifies fabric management. It is important to evaluate the improved FC technology to determine how this may affect the fabric topology and the selection of the file systems.

SAN scalability is another important area of investigation when looking at the deployment of new switches in an existing fabric. SAN scalability measures how a fabric design can grow without requiring a substantial re-layout of existing fabric topology. An effective SAN architecture needs to be able to accommodate additional servers, switches, and storage with minimal impact to the existing SAN operation.

The initial GUPFS testbed had a Brocade SilkWorm 2800, which has a 16-port 1 Gb/s switch. In early FY 2003, we added two additional switches to the testbed:

- A Brocade SilkWorm 3800 16-port 2 Gb/s Fibre Channel switch
- A Qlogic SANbox2-16 16-port 2 Gb/s Fibre Channel switch

Both switches are capable of sustaining 32 Gb/s (full duplex) nonblocking switch throughput. During FY 2003, we tested the new 2 Gb/s switches to determine the impact of 2 Gb/s FC on I/O performance in the SAN fabric. The tests were run in three configurations: Direct Attach, SAN fabric with the Brocade 3800 switch, and SAN fabric with the Qlogic SANbox2-16 switch. In the Direct Attach configuration, an EMC CX 600 2 GB/s FC port was directly connected to a Linux host. The results obtained from the Direct Attach configuration are used as the baseline for comparison. In the SAN fabric setups, the EMC FC port and the Linux host were connected to a switch. The two charts in Figure 10 show the PIORAW performance for in-cache (IC) reads and writes in the three test configurations.

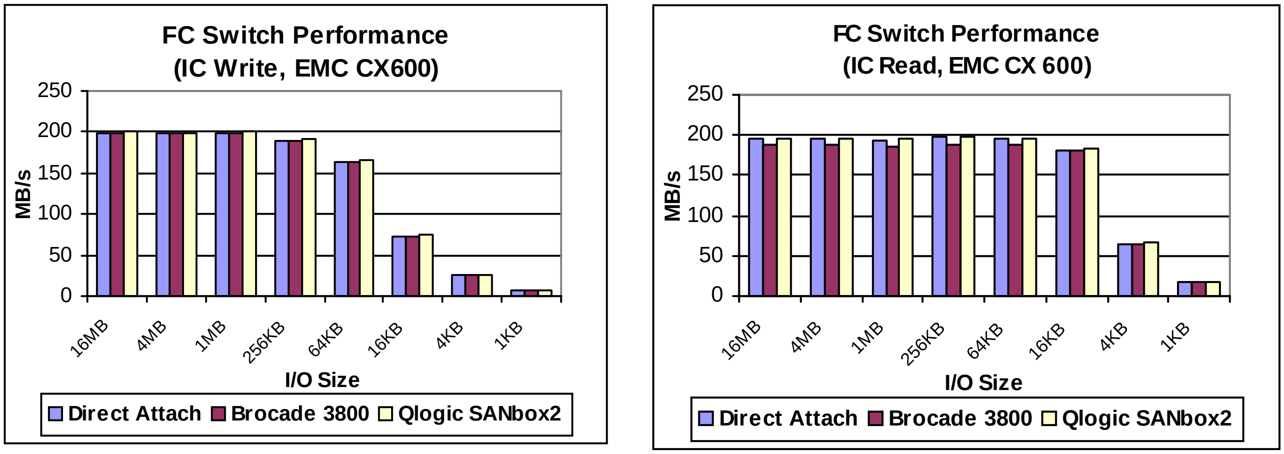


Figure 10. Fibre Channel switch performance.

The results indicate that both the Brocade SilkWorm 3800 switch and Qlogic SANbox2-16 switch were capable of sustaining a 200 MB/s transfer rate. For in-cache writes, there was no noticeable performance difference between the tests in the direct-attach setup and the tests in a switched configuration. However, the read performance on the Brocade SilkWorm 3800 was about 3% to 4% slower than the performance in the direct-attach configuration or with the SANbox2-16 switch for I/O sizes larger than 16 KB. These results indicate that, from a performance point of view, the Qlogic SANbox2-16 switch is a better switch than the Brocade SilkWorm 3800 switch,.

We also tried to put all three switches into a single fabric to investigate how switches from different vendors may work together in a hybrid environment; however, our attempt was a total failure. The switches could not interoperate with one another. The original Brocade 2800 that we obtained in FY 2002 only had a base license and did not have any optional licenses such as zoning support. FC zoning allows port isolation, which allows hosts to only access storage that is in the same zone. When we put the SilkWorm 2800 switch together with the SilkWorm 3800 switch with zoning defined, the SilkWorm 2800 switch failed to function properly. We also could not link the Brocade SilkWorm 3800 switch to the Qlogic SANbox2-16 switch. Interoperability between a Brocade switch and a non-Brocade switch would not work unless we changed the interoperablity mode on the SilkWorm 3800 switch. However, if we did that, according to Brocade Support, it would void the support contract and the SilkWorm 3800 switch would become unsupported by Brocade. Our experience seems to indicate that switch interoperability is still a big issue and is a long way from becoming a reality.

The interoperability issue also existed between the Cisco SN5428 iSCSI switch and the Brocade switch. This was because the embedded FC switch inside the Cisco SN5428 switch is actually a Qlogic switch. Because of this, there was no interoperability issue between the Cisco SN5428 switch and the Qlogic SNAbox2-16 switch.

Inter-Switch Link (ISL) trunking is another area in which we encountered interoperability problems during our evaluation. ISL trunking allows load balancing between multiple ISL links, which improves switch-to-switch performance and reduces traffic congestion. Many FC switch vendors provide ISL trunking up to four paths combined into one logical path up to 8 Gb/s for inter-switch connections. Unlike Ethernet or IP trunking, which are commonly available on the IP network switches, most FC switch vendors support ISL trunking only for their own switches, and often ISL trunking is not supported when switches are from different vendors. The lack of ISL trunking support for heterogeneous switches may become an issue for the GUPFS project if a hybrid fabric is required.

5.2 iSCSI Evaluation

Additional SAN fabric technologies are beginning to appear. Using Ethernet as a SAN fabric is becoming possible because of the iSCSI standard for doing SCSI storage traffic over IP networks. This is very attractive as it allows lower-cost SAN connectivity than can be achieved with Fibre Channel, although with lower performance. It will allow large numbers of inexpensive systems to be connected to the SAN and use the cluster file system through commodity components.

In FY 2003, we evaluated the iSCSI technology in the following areas:

- Software iSCSI performance over traditional Gigabit Ethernet (GigE) interfaces
- Hardware iSCSI performance over Intel iSCSI HBA (Intel PRO/1000 T IP HBA)
- Interoperability among IP routers, the storage router, and FC switches

Figure 11 shows the difference between the software iSCSI performance and the native FC performance, with the FC performance as the baseline. The FC performance was obtained using a Qlogic QLA2300 HBA. Both tests used the same storage on the Yotta Yotta GSX 2400 NetStorager accessed through its 2 Gb/s FC ports.

The results indicate that the iSCSI performance was only about 60% to 80% of the baseline Gigabit Ethernet performance over the GigE interface. The main reason for the performance degradation of iSCSI was probably the overhead of the TCP/IP software stack, on top of which the iSCSI protocol is implemented. Using software iSCSI for storage access appears to cause the usual performance degradation caused by TCP/IP traffic. However, since TCP/IP is available across almost every fabric and interconnect, iSCSI can be a nearly universal mechanism for accessing storage with shared file systems. In addition, because software iSCSI does not require additional expensive interfaces, it is a cost-effective way for accessing storage with shared file systems for inexpensive client systems that can tolerate the lower performance it offers.

We also measured the differences in CPU utilization between doing transfers over FC and transfers using software iSCSI over a regular GigE interface in order to study the CPU overhead associated with different storage transfer protocols.

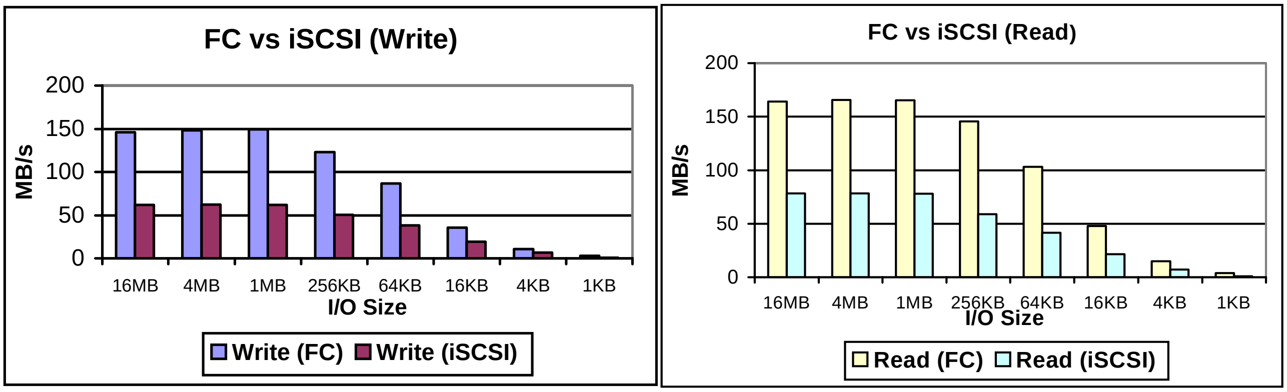


Figure 11. FC and iSCSI performance comparison.

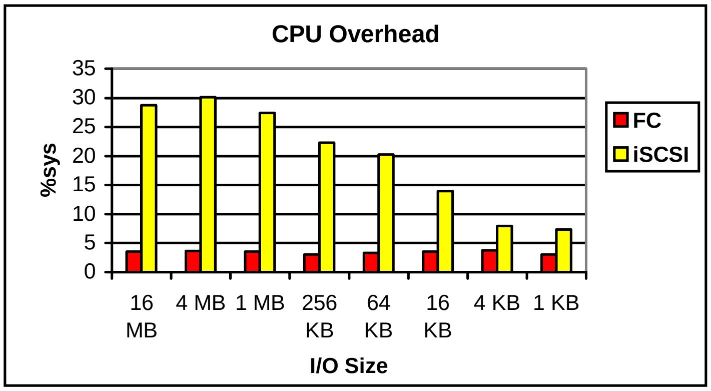


Figure 12. CPU utilization comparison.

Figure 12 shows a comparison of the CPU overhead of a single-thread read when done with software iSCSI over a regular GigE interface and when done using the native FC interface. The CPU measurements were reported by the Linux *vmstat* command. The figure indicates that, for single-thread sequential read, software iSCSI used more CPU cycles while achieving lower throughput numbers, whereas FC used less CPU (about 1/8 for larger I/O sizes and 1/2 for smaller I/O sizes) and had higher throughput numbers.

We are interested in studying the capabilities, performance, and overhead of hardware-assisted iSCSI solutions for use in the deployed GUPFS solution. Because iSCSI HBAs appeared to offer the potential for the greater performance and lower CPU overhead by offloading both the iSCSI protocol processing and Ethernet processing, we chose to study these devices rather TOE cards.

To study the benefit of iSCSI HBAs, we evaluated the Intel iSCSI HBA (Intel PRO/1000 T IP HBA). Figure 13 gives a comparison between the performances of software-based iSCSI using a standard GigE interface and the hardware-assisted Intel iSCSI HBA. The tests were run with the Cisco SN5428 iSCSI storage router bridging the Gigabit Ethernet fabric to a 2 Gb/s FC port on the Chaparral controller. The figure shows that the performance of software iSCSI (using the Cisco's iSCSI software driver and a regular GigE interface) was actually higher than Intel's iSCSI HBA. This seems to contradict the general expectation that higher performance would be achieved with a hardware-assisted iSCSI HBA card than with the software iSCSI implementation going through the regular TCP/IP stack.

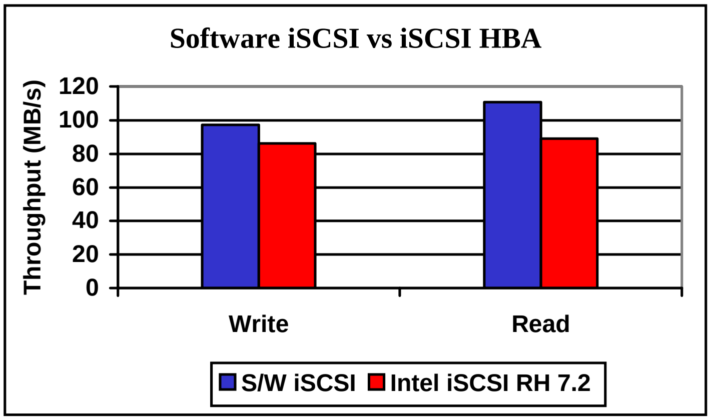


Figure 13. Comparison between Software iSCSI and iSCSI HBA.

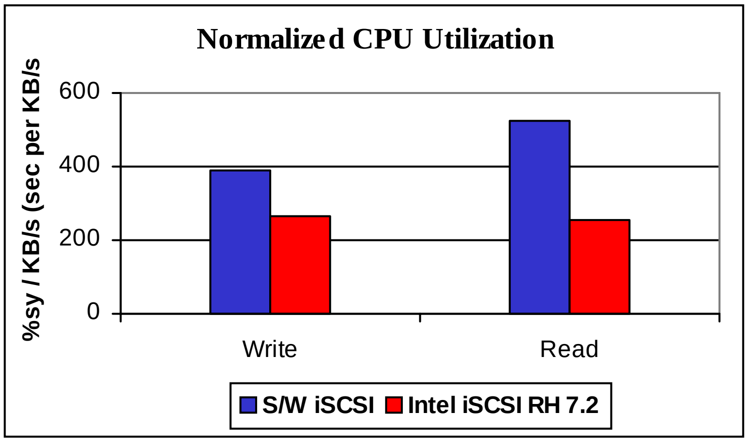


Figure 14. Normalized CPU overhead.

To gain further insight into the differences between these two technologies, we measured the CPU overhead associated with each. We did see lower CPU utilization when the iSCSI HBA was used. However, due to the lower I/O performance accomplished by the iSCSI HBA, it was difficult to do a fair comparison without normalizing the CPU utilization. Figure 14 shows the normalized CPU utilization comparison between these two technologies.

The normalized CPU utilization is defined as the %sys reported by the *vmstat* command divided by the throughput number. The normalized CPU utilization is used to compare the CPU overhead for transferring the same amount of data in each second. As the figure indicates, although the software iSCSI was able to achieve a better throughput number, it also used more CPU cycles to transfer the same amount of data. However, it was a surprise to us that the saving on the CPU utilization with the iSCSI HBA was only about 50% over the software iSCSI. This seems to indicate that, even with offloading TCP/IP processing to the iSCSI HBA, the CPU overhead of iSCSI is still significantly higher than the CPU overhead of a Fibre Channel HBA. In fairness, we must note that both the iSCSI HBA hardware and driver software were very new and clearly not yet optimized. This was also true of the purely software iSCSI implementation tested. Nonetheless, we are disappointed in the iSCSI HBA performance and CPU overhead.

Another consideration when evaluating the utility of iSCSI HBAs and their role in a deployed GUPFS solution is their cost. Although these early iSCSI HBAs are priced somewhat cheaper (25% to 50%) than the FC HBAs, they are still much more expensive than regular Gigabit Ethernet interfaces, which are under 10% of the cost of an FC HBA. Given their limited performance, substantial system overhead, and high cost, it is our conclusion that iSCSI HBAs are not cost effective and software iSCSI solutions using standard Ethernet interfaces should be used, unless there are compelling CPU overhead considerations for specific systems.

Another important consideration for deploying iSCSI technology is the availability of iSCSI target devices (storage devices with iSCSI protocol interfaces). Currently, very few storage vendors have supported or plan to support native iSCSI target interfaces on their storage systems. An alternative is to use the iSCSI routers or bridges that provide protocol conversion between the iSCSI protocol and the FC protocol. The Cisco SN 5428 iSCSI router and Cisco MDS 9509 switch are two examples of such devices. However, none of these technologies currently supports a large number of iSCSI ports for shared access to the same storage pool. The low iSCSI port count per storage system or iSCSI router/bridge will impact the maximal aggregate bandwidth of an IP-based SAN. In order to achieve high aggregate bandwidth as envisioned by the GUPFS project, a much more complicated fabric infrastructure may be needed, with several tiers of switches and routers, and the iSCSI routers/bridges on the edge of the fabric fanning out for host connectivity.

Despite some performance limitations and the limited availability of iSCSI storage devices and fabric bridges, iSCSI shows real promise for enabling systems to cost effectively access a block-based shared file system without requiring expensive fabric interfaces be added to those systems..

5.3 InfiniBand Switches

In addition to using Ethernet as a SAN, the InfiniBand interconnect shows promise for use in a SAN as a transport for storage traffic. InfiniBand offers performance (high bandwidth and low latency) beyond that of either Ethernet or Fibre Channel, with even higher bandwidths planned. With the possibility of commodity-level pricing and the likelihood of some future NERSC systems using InfiniBand as an interconnect , this is a technology that needs to be studied, even in its very early stages. As with Ethernet fabrics, an important part of the InfiniBand technology that needs to be examined is fabric bridges between InfiniBand and Fibre Channel and Ethernet SANs. Another important area of exploration is storage transfer protocols (e.g., SRP) and methodologies. We investigated both of these.

In 2003, we evaluated the InfinIO 7000 InfiniBand switch from InfiniCon Systems and the Topspin 90 switch from Topspin. Both the 1x and 4x InfiniBands were tested. When 4x InfiniBand results become available, we will present those results instead of the older 1x results. Figure 15 shows a picture of the InfinIO 7000 and Topspin 90 switches in the GUPFS testbed environment.

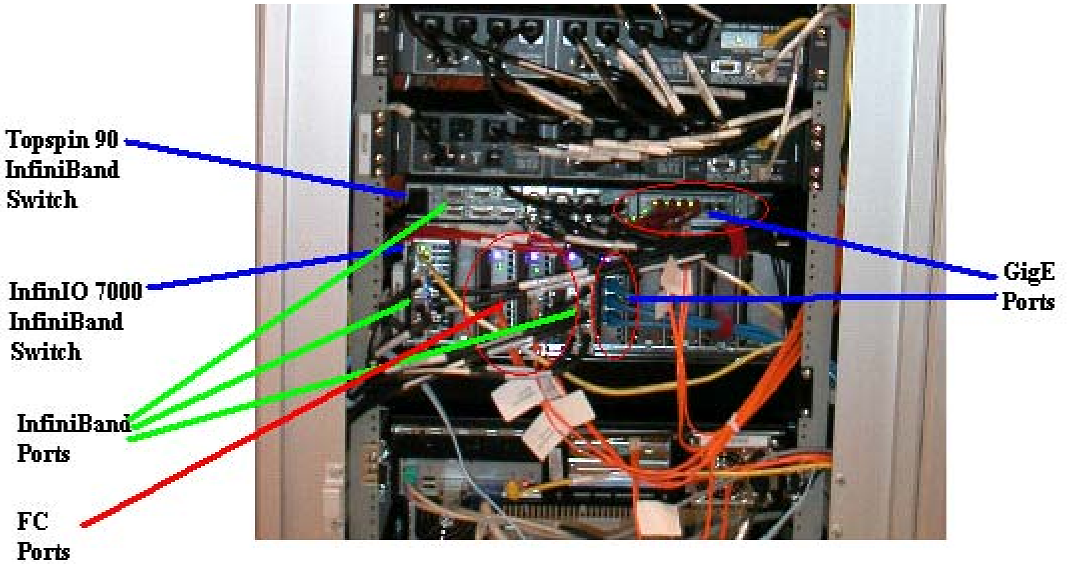


Figure 15. The InfinIO 7000 shared I/O switch and the Topspin 90 switch.

5.3.1 InfiniCon InfinIO 7000 InfiniBand Switch

The InfinIO 7000 shared I/O system is a multi-protocol networking system for shared I/O and InfiniBand switching. The server nodes attach to the InfinIO 7000 switch via a high-speed, 10 Gb/s (4x) InfiniBand connection to access a pool of virtual I/O resources, including Fibre Channel SANs, Ethernet SAN or network attached storage (NAS), and native InfiniBand fabrics. This shared I/O architecture eliminates the need for separate individual Ethernet NICs, FC HBAs, and cabling on the server nodes. This saving of operational and capital costs for infrastructure could be significant.

Each InfinIO chassis supports dual 4x InfiniBand switch modules and up to eight I/O "personality modules." The I/O personality module can be a three-port 1 Gb Virtual Ethernet Exchange (VEx) card, a two-port 2 Gb Virtual Fibre Channel Exchange (VFx) card, or a six-port InfiniBand Expansion (IBx) card. Chassis slots can be populated with any mix of personality modules, which can be hot-swapped to accommodate configuration changes.

Figure 16 shows what a single 4x HCA was able to achieve with a single Linux host, using two Infin7000 Fibre Channel (VFx) line cards. Each VFx card has two 2 Gb/s FC ports. With two VFx line cards, the aggregate FC bandwidth is 8 Gb/s. A single 4x HCA is able to sustain a 10 Gb/s data transfer rate, so theoretically a single 4x HCA would be able to saturate the 8 Gb/s aggregate FC bandwidth with the two VFx cards. However, our results indicate that the best performance we were able to achieve was about 470 MB/s. The underlying storage used was the Yotta Yotta GSX 2400 storage system with four 2 Gb/s FC ports. The Yotta Yotta GSX 2400 was demonstrated to be able to sustain more than 750 MB/s for in-cache read and write with four 2 Gb/s FC ports, so the underlying storage was not likely a bottleneck. It seems that, in this test configuration, the Linux host seemed to be the bottleneck. However, it is not clear, without more investigation, whether the bottleneck was the CPU, the PCI-X bus, benchmark software, driver software, Vfx module, or something else.

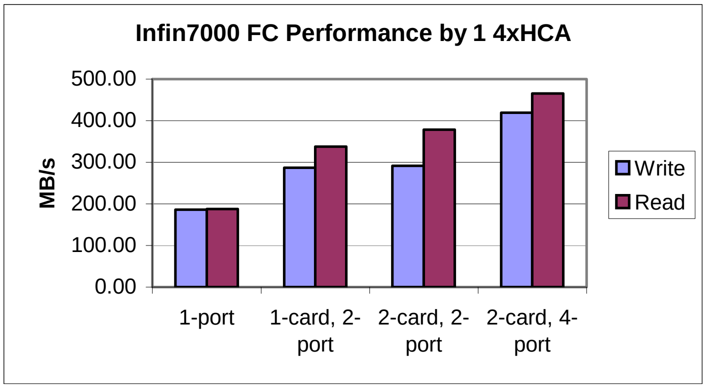


Figure 16. Aggregate Infin7000 FC performance with a single 4xHCA.

Nonetheless, these SRP results are very encouraging. They show that a single 4x HCA can sustain twice the I/O performance of a single 2 Gb/s FC HBA, and the cards cost about the same. From a price-performance point of view, it seems to indicate that InfiniBand with the SRP protocol may be more cost-effective than the FC solution. However, since InfiniBand technology is still immature, both InfiniBand and the SRP technologies are still evolving. Currently, it may be a little bit risky for a full adoption of the InfiniBand technology for GUPFS, but it is likely that in the future InfiniBand will factor into the deployed GUPFS solution.

5.3.2 Topspin TS90 InfiniBand Switch

The Topspin 90 contains 12 InfiniBand ports that are used to create a single 10 Gb/s fabric for inter-process communications, storage, and networking. The Topspin 90 switch can be expanded from the 12-port base configuration by inserting an optional Ethernet Gateway-Router Module or a Fibre Channel Gateway Module into the expansion slot. The Fibre Channel Gateway Module supports two 2 Gb/s Fibre Channel (FC) ports, while the Ethernet Gateway-Router Module supports four 1 Gb/s GigE ports.

Topspin also manufactures InfiniBand 4xHCA adapters that support a full 10 Gb/s of peak bandwidth. With the 4xHCA adapter, virtual NICs or HBAs can be created in every server for networking or storage access.

Figure 17 shows a comparison between the native 2 Gb/s FC HBA performance and the SRP performance of a single 4xHCA and the Topspin 90 switch with a Fibre Channel Gateway Module. Both tests used the same FC-connected storage device. The results indicate that a Linux host with the SRP driver was able to achieve a performance similar to a 2 Gb/s FC HBA. These results match the SRP performance we saw earlier with the InfiniCon 4xHCA (see Section 5.3.1, above). It is encouraging that there are more InfiniBand vendors that support the InfiniBand to FC bridging capability. (Voltaire is another InfiniBand vendor that support this bridging capability.) The fact that more than one vendor is supporting this InfiniBand to FC bridging capability indicates that there is a perceived market for it and the competition may make the technology more robust with improved performance. A more complete evaluation of the Topspin 90 switch performance can be found in Appendix E.

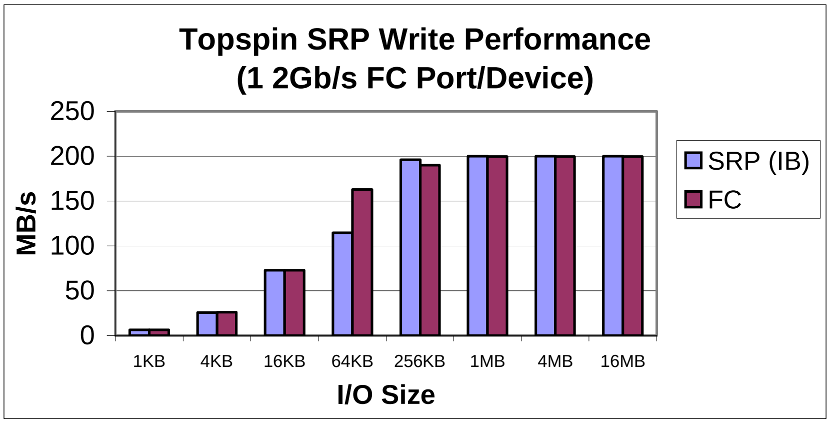


Figure 17. The SRP performance with the Topspin 90 switch.

5.4 Fabric Performance Comparison

A single-process read performance comparison was made using several fabric technologies: Fibre Channel (FC), SRP over InfiniBand (SRP), iSCSI over GigE (ISCSI_GE), and iSCSI over IPoIB (ISCSI_IB).

The FC performance was obtained using a Qlogic QLA2300 HBA, which provides the baseline. The SRP performance was obtained using a 1x HCA from InfiniCon. An InfiniCon InfinIO 7000 switch was used to provide InfiniBand to FC bridging. The iSCSI over GigE (ISCSI_GE) performance was obtained using the software iSCSI driver over a GigE network interface card. A Cisco SN5428 iSCSI router was used to provide iSCSI to FC bridging. The iSCSI over IPoIB (ISCSI_IB) performance was obtained using the software iSCSI driver and the IPoIB (IP over InfiniBand) protocol, over a 1x HCA from InfiniCon. An InfiniCon InfinIO 7000 switch was used to provide InfiniBand to Ethernet IP bridging and a Cisco SN5428 iSCSI router was used to provide iSCSI Ethernet (IP) to FC bridging.

Figure 18 shows the results of single-thread reads of different I/O sizes using different fabric technologies. The best performance was achieved by the 2 Gb/s FC interface, followed by the SRP protocol over InfiniBand. Since the iSCSI traffic was passing through a single Gigabit Ethernet interface, the best iSCSI performance would be less than 100 MB/s. With the additional TCP/IP software stack overhead of IpoIB on top of the immature InfiniBand drivers, iSCSI over IPoIB delivered the lowest performance for single-thread reads. It is clear from these results that FC is the leader and delivers the best performance for storage access. However, when InfiniBand is used as the cluster interconnect, SRP may be the preferred mechanism for storage access as it eliminates the cost of additional FC HBAs for each host and still delivers a good I/O performance.

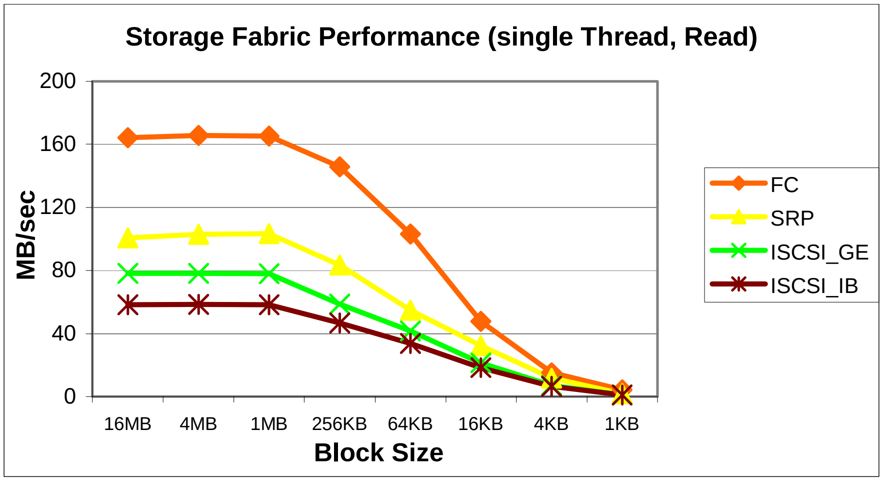


Figure 18. Storage fabric performance.

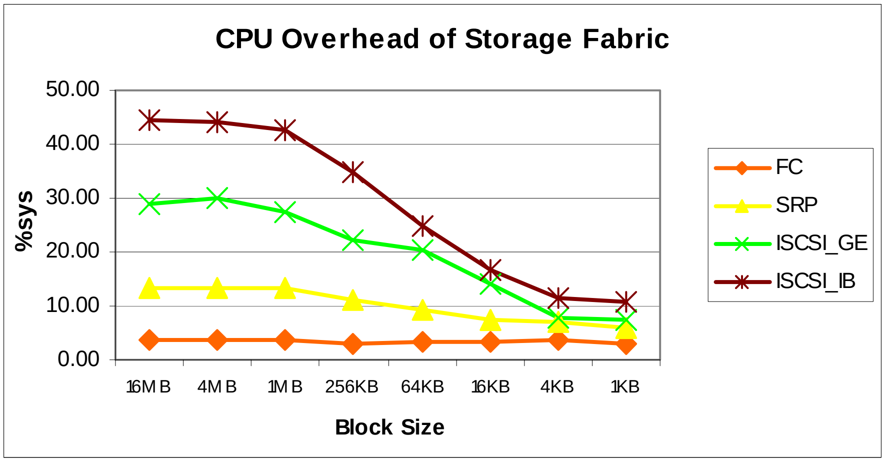


Figure 19. CPU overhead of storage fabric.

Figure 19 shows the CPU overhead of different protocols for single-thread reads. FC, which delivered the best performance, also generated the least CPU overhead. The iSCSI protocol allows the standard SCSI packets to be enveloped in IP packets and transported over standard Ethernet infrastructure, allowing SANs to be deployed on any networks supporting IP. This option is very attractive as it allows lower-cost SAN connectivity than can be achieved with Fibre Channel, although with lower performance. It will allow large numbers of inexpensive systems to be connected to the SAN and use the shared file system through commodity-priced components. While attractive from a hardware cost perspective, this option does incur a performance impact on each host because of increased traffic through the host's IP stack (Figure 19).

Note that the SRP results and the iSCSI over IB results shown here were obtained using the InfiniCon 1x HCA, which was the only InfiniBand HCA available when the tests were conducted. Improved results may be possible with the newer 4x HCA, but due to the lack of resources, the tests were not repeated. However, separate SRP tests with the 4xHCA seemed to demonstrate better I/O performance, as shown in Section 5.3.1.

5.5 Summary

The results of our 2 Gb/s Fibre Channel SAN fabric testing show that individually the 2 Gb/s FC switches perform well and are able to achieve the entire 200 MB/s per port performance provided by 2 Gb/s FC. The two 2 Gb/s Fibre Channel switches evaluated performed nearly identically except for the minor 3% to 4% lower read performance of the Brocade 3800 compared to the Qlogic SANbox2-16. The performance of storage transfers through 2 Gb/s FC fabrics based on each of these switches was nearly identical to that of the same 2 Gb/s FC storage devices when directly connected to the hosts. Overall, the performance of 2 Gb/s FC fabrics was good.

While the performance of 2 Gb/s FC fabrics was good, we did find interoperability problems between different vendors' switches. Two areas of interoperability deficiency were identified. One problem was that Brocade FC switches were unable to coexist with other vendors' switches without resorting to unacceptable configuration and support options. We recommend the use of FC switches by vendors other than Brocade for GUPFS deployment. A second problem area was in the use of trunking Inter-Switch Links (ISL) to achieve higher inter-switch bandwidth, which will be needed for larger fabrics. Those vendors that provide FC ISL support, normally do so only between their own switches. This may become a problem in large heterogeneous fabrics.

Standardization efforts are underway in the storage industry that are targeting improved Fibre Channel interoperability. We need to continue monitoring and testing FC fabric interoperability as GUPFS moves towards deployment.

Our evaluation of iSCSI has lead us to believe that it is a viable technology that allows block storage transfers across any fabric or system interconnect that supports IP protocol traffic. We tested both hardware and software iSCSI implementations. The results of our tests lead us to believe that software iSCSI works well and is inexpensive. However, software iSCSI has a lower performance level than FC, and this comes with the cost of substantially higher system overhead. The results of our tests of the hardware iSCSI HBA solutions lead us to conclude that at this time software iSCSI solutions perform better and that the iSCSI HBAs are currently not cost effective based on their lesser performance and still substantial system overhead.

Our iSCSI evaluations demonstrated that multiple fabrics can be bridged together for block storage transfers. We successfully bridged iSCSI traffic from Gigabit Ethernet–attached hosts to FC storage devices and bridged iSCSI traffic from InfiniBand-attached hosts to a Gigabit Ethernet fabric and then to FC storage devices. This illustrates the flexibility of iSCSI and also how lower-cost fabrics, such as Gigabit Ethernet, can be used to cost effectively access storage for block-based shared file systems running on low-cost systems. In order to prepare for a multiple fabric GUPFS deployment, we need to continue tracking iSCSI developments, especially in the areas of additional iSCSI fabric bridges and iSCSI attached storage devices, and conduct evaluations of these as they become available .

The evaluations that we conducted of InfiniBand fabrics from two different vendors showed that InfiniBand is a successful and effective interconnect for SAN storage traffic. We were successful at accessing FC-attached storage through the InfiniBand fabric using both the native IB SRP storage transfer protocol and iSCSI with IP over IB. This was accomplished through the use of FC and Gigabit Ethernet gateway modules in the IB switches, which allowed all three fabrics to be bridged together into a single fabric. The IB SRP protocol showed very good performance that easily matched and in some configurations exceeded that of the 2 Gb/s Fibre Channel.

InfiniBand is very interesting in that it allows both high-performance message passing and storage transfers to be conducted using a single interconnect and host adapter. We expect to see InfiniBand in future NERSC systems. Therefore, in order to determine how to best integrate IB into a GUPFS deployment, we need to continue tracking its development and evaluating its SAN performance as new switches, fabric bridge gateways, and higher speed fabrics appear.

6 File System Evaluations and Results

One of the major focuses of the GUPFS project in FY 2003 was to investigate and evaluate a large number of shared file systems. The Sistina Global File System (GFS) file system was first evaluated in FY 2002. As new versions of GFS became available, they were evaluated to track GFS's progress. In general, our plan was to conduct periodic evaluations of each of the shared file systems that showed promise in order to track their continued progress, and to give feedback to the vendors about the performance of new versions and additional features that were required.

During FY 2003, we conducted evaluations of the following shared file systems:

- Sistina GFS 5.1 and 5.2 Beta
- ADIC StorNext File System 2.0 and 2.1
- Lustre File System 0.6 (Release 1.0 Beta 1)
- IBM GPFS File System on Linux 1.3

This list of file systems to be evaluated is not exhaustive. There are a number of other file systems that the GUPFS project is interested in evaluating within the next year. These include the IBM TotalStorage SAN File System (also known as StorageTank), the Ibrix file system, and the Panasas file system. These and other file systems will be evaluated based on their availability and adequate time resources.

6.1 Sistina GFS File System Evaluation

The GUPFS project first evaluated GFS 5.1 in the beginning of FY 2003, followed by the GFS 5.2 Beta evaluation in June. Some of the performance results are presented here.

Figure 20 shows GFS 5.1 parallel in-cache I/O scalability using MPTIO with 1 GB file size, as compared to the underlying Yotta Yotta (YY) GSX 2400 storage performance. The underlying storage used to create the file system was a single shared 28-way RAID-0 LUN exported on eight 2 Gb/s FC ports. The MPTIO parallel I/O test was conducted with parallel I/O processes performing streaming I/Os, with each process reading/writing a separate file, and with the same amount of data (which is equal to 1 GB divided by the number of nodes). The results presented by the 'GFS (Read)' curve show cached read performance.

With a scalable storage system like Yotta Yotta's GSX 2400 as the underlying storage, GFS 5.1 was able to demonstrate good scalability, as demonstrated in Figure 20.

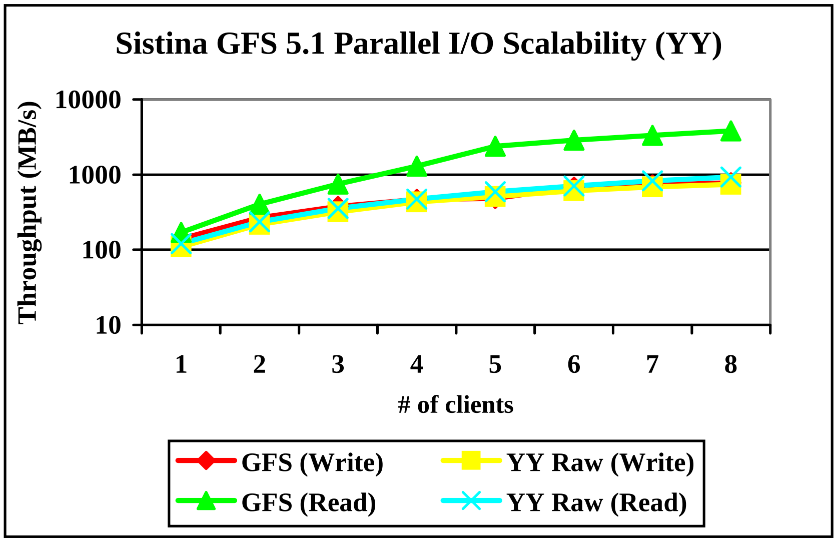


Figure 20. GFS 5.1 parallel I/O scalability(with YY storage).

Figure 21 shows GFS 5.1 shared I/O scalability. The MPTIO shared I/O test was conducted with parallel I/O processes reading/writing the same file simultaneously. Parallel access to a single file is common in the GUPFS target environment, especially for scientists that organize their data via HDF or netCDF. The MPTIO shared I/O test measures a file system performance for shared and concurrent access to a single file. The result indicates that GFS 5.1 had no scalability for parallel writes to the same file. Later tests conducted with the GFS 5.2 Beta release also showed no scalability for the shared I/O tests.

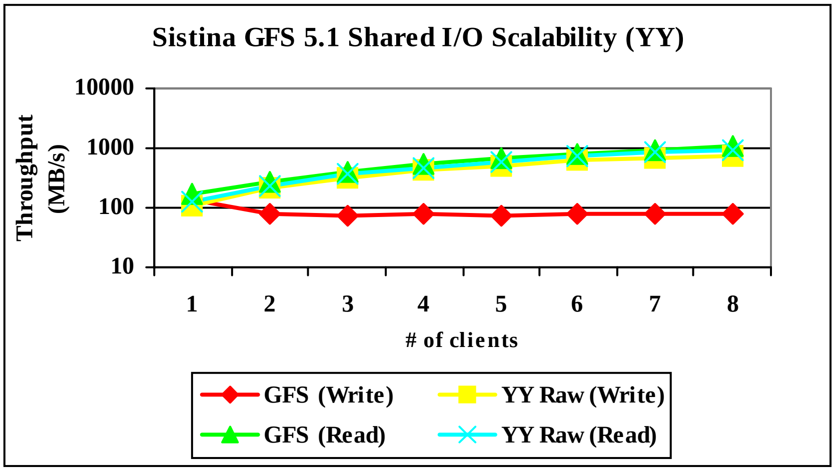


Figure 21. GFS 5.1 shared I/O scalability (with YY storage).

The Sistina GFS has faded from our radar screen as a candidate for further investigation, mostly due to the complete lack of scalability in simultaneous concurrent writes to the same file. Other factors contributing to our abandoning GFS were Sistina's lack of interest in supporting platforms other than Linux, and the lack of other functionalities (e.g., DMAPI and shared I/O support) that are important to the expected GUPFS production environment. Although some strides were made in scalability, with GFS being supported on systems with between 128 and 256 nodes, Sistina's commitment to support much larger numbers of nodes, as in the envisioned GUPFS target environment, is questionable.

6.2 ADIC StorNext File System Evaluation

The ADIC StorNext File System (SNFS), previously known as the CentraVision file system (CVFS), was one of the first two file systems evaluated in FY 2003 (the other file system was GFS). Version 2.02 of the StorNext system was first installed on the GUPFS testbed. ADIC provided on-site training to facilitate our evaluation. Through the evaluation, ADIC established a very good working relationship with the GUPFS project and NERSC, and has been working with NERSC to address its requirements for a center-wide shared file system.

The StorNext File System is unusual in the shared file system arena in that it already provides support for multiple operating systems — Linux, SGI IRIX, Microsoft Windows, Solaris, and AIX. The multiple platform support provided by StorNext will allow the GUPFS project to test a shared file system in a truly heterogeneous hardware platform and OS environment, potentially involving Sun Solaris SPARC systems, an IBM AIX system in the form of the dev2 system, and an SGI IRIX system in the form of the NERSC visualization system.

Figure 22 shows test results of an early eight-client MPTIO benchmark running on StorNext 2.0. The tests were run using the DotHill storage system, with two RAID-5 LUNs exported on the two 1 Gb/s FC ports. In this test configuration, the DotHill storage became a bottleneck, limiting file system performance.

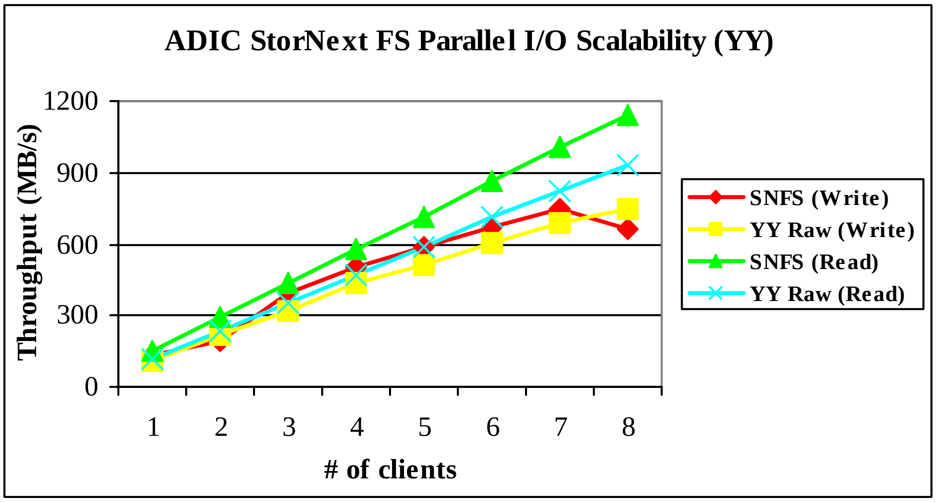


Figure 22. StorNext 2.0 MPTIO performance results (with DotHill).

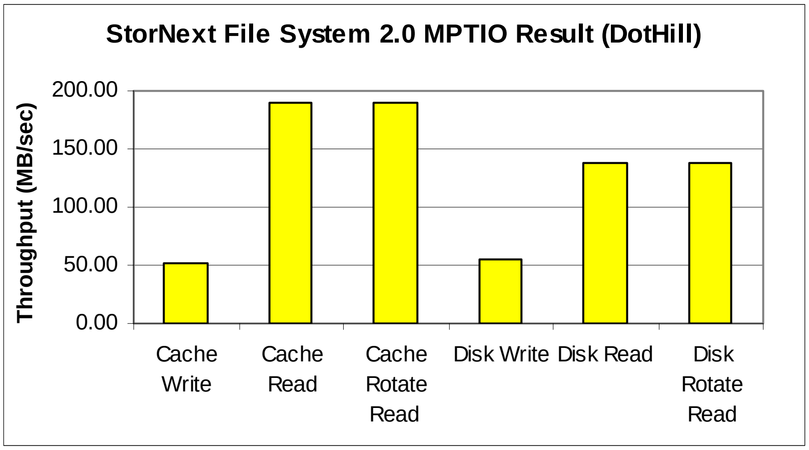


Figure 23. StorNext FS 2.0 parallel I/O scalability (with YY storage).

Figure 23 shows a comparison between the performance achieved using StorNext and the performance of the underlying Yotta Yotta storage, using a single shared 28-way RAID-0 LUN exported on eight 2 Gb/s FC ports. The benchmark used was PIORAW. With a scalable storage subsystem as the underlying storage, StorNext was able to achieve more than 1140 MB/s (>1 GB/s) for reads with eight clients.

The ADIC's StorNext File System will continue to be a candidate for further investigation. It is the file system that has the broadest platform coverage with cross-platform support. Installing and setting up a StorNext file system was a very simple task and the ease of management makes it a very attractive solution for the shared file system. A relevant SNFS development in FY 2003 was Cray's selection of the ADIC StorNext File System as the file system for the Cray X1, with improved scalability. Unfortunately, ADIC appears to be phasing out DMAPI support in StorNext, in favor of ADIC's own proprietary interface. The lack of support for a larger number of nodes may be another challenging area of concern. Currently, the StorNext File System has only been tested and supported for up to 256 nodes. It is not clear at this point whether there would be any technology limitations that may limit the maximal number of nodes supported by the StorNext File System. This is an issue for further investigation.

6.3 Lustre File System Evaluation

With the award of the ASCI PathForward SGSFS file system development contract to Hewlett-Packard (HP) and Cluster File System, Inc., there has been rapid progress on the Lustre file system. As of the middle of the first quarter of FY 2003, versions of Lustre appeared to be stable and functional enough to begin testing. However, the first Lustre file system version (Release 0.6, also known as Release 1.0 Beta 1) failed to complete any of our standard MPTIO tests, with the exception of the 'Cache Write' and 'Cache Read' tests (as shown in Figure 24).

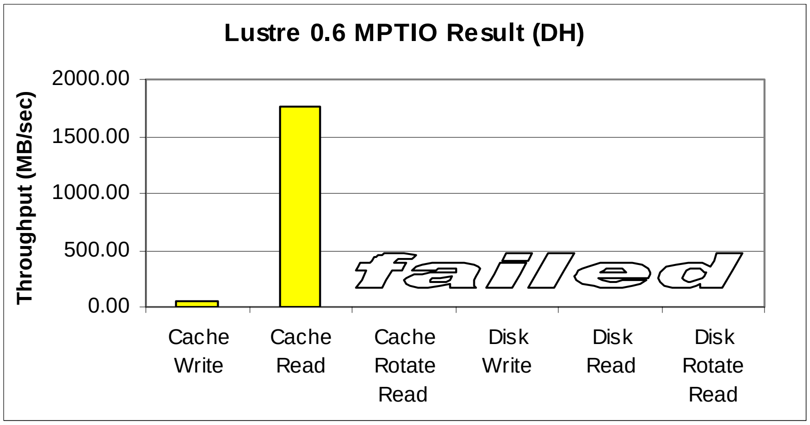


Figure 24. Lustre 0.6 (1.0 Beta 1.0) MPTIO performance (with DotHill).

Due to its popularity in many of the HPC laboratories and the heavy investment made by other DOE National Laboratories and the Open Source community, Lustre will continue to be a candidate for further investigation. Although its lack of interest in supporting platforms other than Linux and certain missing functionalities (e.g., DMAPI) that are important to GUPFS have made Lustre much less attractive to GUPFS at this time, we expect Lustre to improve substantially as it is developed further and its functionality and stability mature. We will continue to track and investigate Lustre as it progresses.

6.4 GPFS File System Evaluation

IBM’s General Parallel File System (GPFS) file system is now available on two systems at NERSC: the production Seaborg IBM SP system and the Intel-processor-based LBNL Alvarez Linux cluster. Initially, GPFS was not a candidate file system for center-wide deployment under the GUPFS project because of licensing issues and limited platform and OS support. However, during FY 2003, changes in the IBM licensing and support policies enabled GPFS 1.3 for Linux to be installed on the GUPFS testbed for evaluation.

GPFS 1.3 for Linux can operate in two modes: SAN mode and NSD server mode. In SAN mode, hosts that are attached to a SAN and can directly access the storage used for the file system perform data transfers directly using the SAN. In NSD server mode, hosts that cannot directly access the storage for the file system communicate with NSD servers via an IP network, and perform data transfers over the network through the NSD servers that have direct access to the storage devices.

Figure 25 illustrates the I/O scalability of GPFS for Linux 1.3 in SAN mode, using the 3PARdata storage (eight 2 Gb/s FC ports accessing one Raid-5 LUN). The chart on the left shows GPFS 1.3 I/O scalability while the chart on the right shows the baseline performance of the 3PARdata S400 storage.

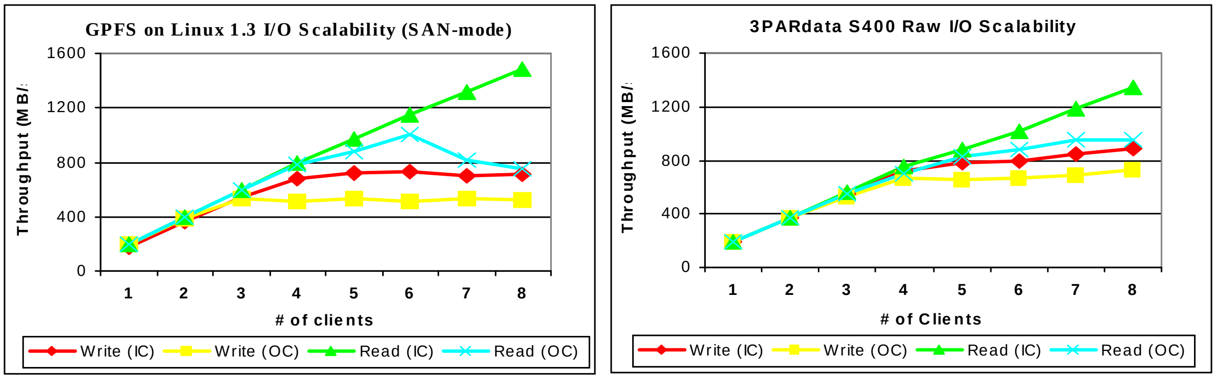


Figure 25. I/O scalability: GPFS on Linux 1.3 (SAN-mode) vs. 3PARdata raw performance.

GPFS 1.3 running in SAN mode, as shown in Figure 25, closely tracked the performance of the underlying storage. Due to read-ahead, GPFS 1.3 read performance generally was better than the raw I/O performance, with the exception of the out-of-cache (OC) read for clients seven and eight. The reason for the performance drop at the end of the Read (OC) curve is not clear. Thus, an area for further investigation may be the study of GPFS read scalability in a larger cluster. For writes, GPFS 1.3 generally was slower than the underlying raw storage. After four nodes, GPFS write performance seemed to increase very little when the number of nodes increased. These preliminary results seem to indicate that GPFS 1.3 has very limited write scalability. We suspect this to be caused by tuning issues and intend to investigate it further.

Figure 26 shows GPFS 1.3 running in NSD server mode, using the Yotta Yotta GSX 2400 as the underlying storage. This GPFS experiment was created with eight NSD servers. Each server was connected to a 2 GB/s Yotta Yotta FC port via a Qlogic SANbox2 switch. The LUN exported on each NSD server was a RAID-0 volume of four 73-GB disks. Gigabit Ethernet was used as the network fabric for transferring data between the NSD servers and the client hosts.

Unlike the test in which GPFS ran in SAN mode (Figure 25), the GPFS/NSD server mode test did not show any meaningful scalability. Both the in-cache (IC) write and the out-of-cache (OC) write performances stayed at about same 200 MB/s after four nodes, while the in-cache-read and out-of-cache read performances, although showing some scalability, were very poor. Once again, we suspect tuning issues are causing this.

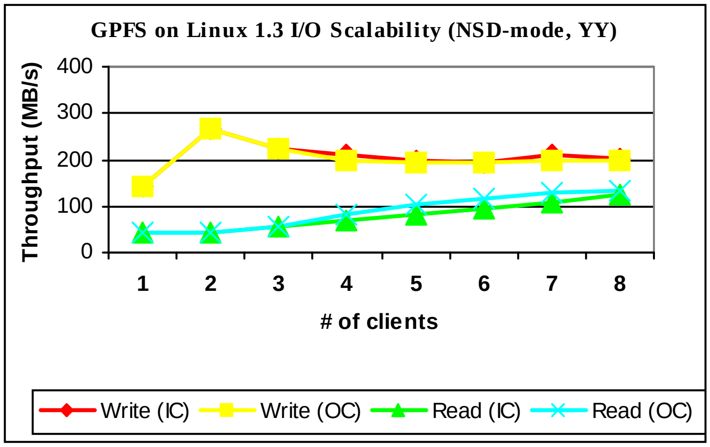


Figure 26. GPFS 1.3 I/O scalability (NSD-mode, with YY storage).

Due to the equipment resource conflict, we were not able to compare GPFS performance running in the SAN-mode and the NSD-mode, using the same underlying storage devices. However, from the previous storage evaluations we know the 3PARdata S400 and the Yotta Yotta GSX 2400 had comparable performances. This allowed us to fairly compare the GPFS results of these two modes. From these results, it seems that GPFS running in the SAN-mode outperforms GPFS running in the NSD-mode, at least in clusters with a small number of nodes. However, it is not clear how they compare in larger clusters with hundreds or thousands of nodes. We are interested in conducting further evaluations using NERSC's Alvarez cluster or the PDSF cluster to study GPFS's scalability in larger clusters.

IBM GPFS will continue to be a candidate for further investigation. GPFS is currently only available on AIX and Linux, and cross-platform and multi-cluster support is as yet unavailable. The earlier lack of IBM's interest in supporting GPFS on other platforms than AIX and restrictive licensing policies had been the biggest obstacles to GPFS being a candidate for GUPFS. However, with the release of GPFS for Linux and greatly improved licensing, these obstacles have been removed. In addition, it is one of the few file systems that support DMAPI (and the only file system that has a parallel DMAPI implementation), which has been the mechanism proposed by the GUPFS team to support HPSS integration and to provide a hierarchical storage management (HSM) functionality similar to that provided by the Data Migration Facility (DMF) in Cray's T3E. GPFS has also been one of the very few cluster/global file systems that have been available in production environments of large clusters for many years. All of these considerations have made GPFS a possible candidate for GUPFS.

So far, all file system tests were conducted with the default file system settings. There was no effort to tune underlying storage and/or file system configuration parameters to optimize I/O performance. Thus, these results are preliminary and their purpose was mainly to help the GUPFS team better understand the file systems for possible inclusion in the final candidate list for deployment.

6.5 A Performance Comparison between File Systems

Figure 27 shows the eight-client MPTIO results for three file systems: Sistina GFS, ADIC StorNext File System, and Lustre, under different test scenarios using the same DotHill storage as the underlying device.

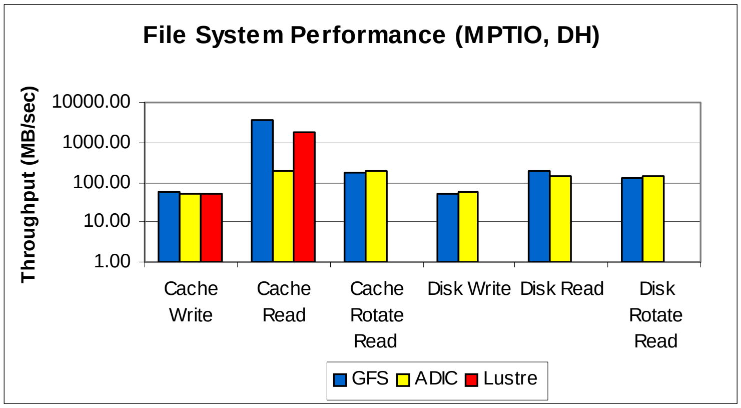


Figure 27. Shared file system comparison.

The test configuration was bottlenecked at the storage system. As a result, the file system performance was limited by what the underlying storage was able to sustain. However, even with a slow storage system, the performance comparison shows some differences among file system implementations.

These results indicate that there was not much difference in parallel I/O performance between GFS 5.1 and ADIC's StorNext File System 2.1, except in the case of cache read. The StorNext File System was found to be the slowest of the three file systems tested in the cache read test. This was because StorNext was doing direct I/O even when operating on files that can fit in the OS cache, while the other two file systems were able to cache the file blocks on the clients. The result also shows the importance of having a scalable storage system for file system evaluation.

Figure 28 shows the METEBENCH parallel file creation test results for Sistina GFS and the ADIC StorNext File System, using the same Yotta Yotta storage as the underlying device. The file creation tests were run in two modes: parallel file creation in separate directories for each process, and parallel file creation by all processes in the same single directory. Lustre did not participate due to its instability at the time of the tests.

ADIC StorNext File System uses metadata servers to process all metadata operations. These servers process the metadata requests, based on the order of the requests received. The ADIC StorNext File System instance under test used a single metadata server, which was a dual 2.2 GHz Pentium-4 node running a Linux 2.4.18-10smp kernel. The metadata server was connected to the file system nodes by a switched Gigabit Ethernet network. Figure 28 seems to indicate that, with the ADIC StorNext File System, there was very little performance difference in file creation between files created in the same directory and those created in separate directories. The file creation rate by the single metadata server was about 500 files/s, and the file creation rate seemed to stay at the same level, with a gradual degradation when more nodes were added.

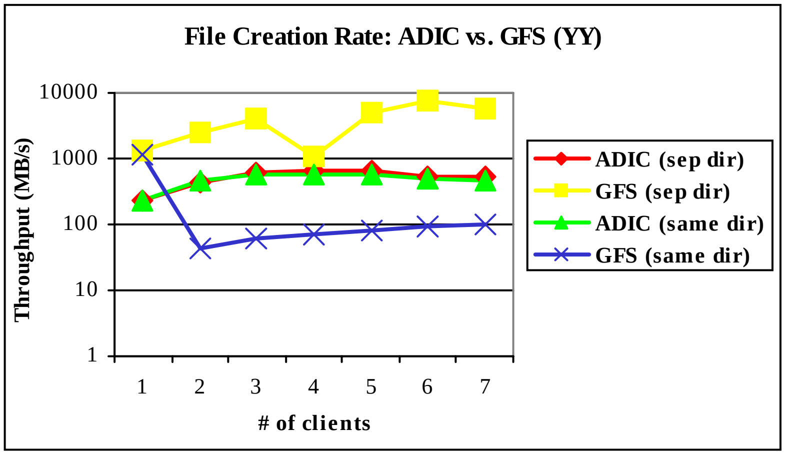


Figure 28. File system file creation rate comparison.

Sistina GFS does not use a metadata server. All the metadata requests are processed by all of the participating file system nodes in the cluster. When the files were created in separate directories, GFS seemed to scale very well, except in the case of the 4-node test. In this instance, the creation rate was about 1000 to 1200 files/s per node and the saturated at about 7500 files/s with 6 or 7 nodes. On the other hand, when files were created in the same directory, the file creation operations need to be synchronized among the nodes. As a result, we saw much lower file creation rate with GFS when files were created in the same directory. The file creation rate with a single node was about 1200 files/s since there was not synchronization overhead. The rate then dropped to 45 files/s with 2 nodes, and then gradually climbed back up at an increment rate of 10 files/s for each additional node. The file creation rate of GFS for 7-node file creation in the same directory was only about 100 files/s.

6.6 GPFS 1.3 Scalability Test on Larger Clusters

The GUPFS project has been very interested in expanding the file system evaluation to larger clusters for the purpose of scalability testing. We conducted such an investigation using the LBNL Alvarez Linux cluster in conjunction with the GUPFS testbed. File systems can be tested on Alvarez using its high-performance Myrinet 2000 interconnect for data transfers. Data transfers can be done over the Myrinet interconnect using both IP and, when possible, GM[†] protocols.

The GUPFS project conducted such a file system scalability test of GPFS 1.3 for Linux on Alvarez. This test was conducted with GPFS 1.3 because its NSD Server mode allowed file system clients to perform file system data transfers over the system's Myrinet interconnect. Because Alvarez's storage is low performance and not scalable, the eight Myrinet attached GUPFS testbed systems and high performance, scalable GUPFS testbed storage devices were used as the NSD Servers and storage for the GPFS file system. The eight GUFPS testbed nodes were integrated with the Alvarez cluster through the expedient of merging the GUPFS and Alvarez Myrinet networks by moving the GUPFS Myrinet line card into the Alvarez switch, thus providing file system services to the Alvarez compute nodes.

We reconfigured Alvarez’s Myrinet fabric to allow the eight GUPFS NSD Server nodes to be merged into the Alvarez cluster. Alvarez had an existing GPFS instance that was created with two storage nodes using two RAID-5 LUNs on the IBM storage. A second GPFS instance was created with the 8 GUPFS nodes as the storage nodes. The following information summarizes the test configuration:

- Linux Cluster: NERSC Alvarez (87 compute nodes) + 8 GUPFS nodes
- Alvarez Nodes: Dual 866 MHz P3, 1 GB system memory
- GUPFS Nodes: Dual 2.2 GHz Xeon P4, 2 GB memory, QLA2340 FC HBA
- File System: GPFS on Linux 1.3
- Test were run on two GPFS instances:
  - ALV: with two storage nodes (on two IBM Raid-5 LUNs, 100 MB/s max per LUN)
  - GUS: with eight storage nodes (on eight Yotta Yotta Raid-0 LUNs, eight x 2 Gb FC ports)

[†] GM stands for Glenn’s Messages, a low-level communication layer for Myrinet.

- Interconnect: Myrinet 2000 (Rev C) with LAPI
- Benchmark: PIORAW

Figure 29 shows the file I/O scalability from 1 client to 64 clients, using the two GPFS instances on Alvarez. The 'ALV' results are the PIORAW results on the existing GPFS instance using the two IBM storage nodes, and the 'GUS' results are the PIORAW results on the GPFS instance that used the eight GUPFS nodes as the storage nodes. This figure shows several interesting points worth mentioning:

- First, the underlying storage affects how the file system scales. When the tests were run on the ALV instance, the results showed no scalability at all. The bottleneck seemed to be at the two storage nodes, possibly because of slower CPU speed, less system memory, a slower storage device, or all of these factors. When the tests were run on the GUS instance, writes scaled very well until 16 clients and at which point, it may have reached the underlying storage capacity and the storage nodes may again have become the bottleneck.
- Secondly, for the GUS instance, reads did not seem to scale beyond 8 nodes, and the read performance was terribly low. We did not have the time to investigate the poor read performance before we separating the eight GUPFS nodes from the Alvarez cluster for other evaluations. However, we have been able to reproduce the same read performance problem on a smaller cluster and we have been working with IBM engineers to investigate this issue with a more recent GPFS. We suspect that this is a configuration tuning issue.
- Thirdly, for the GUS instance, writes seemed to scale well until 16 nodes where writes saturated at about 720 MB/s and stayed at the same level even when the number of nodes increased in the test. This result is encouraging since it seems to indicate that, when the number of nodes increases, the node count will not cause any performance degradation in GPFS for writes. We are also encouraged by the aggregate data rate, which shows acceptable values for high-volume IP-based traffic.

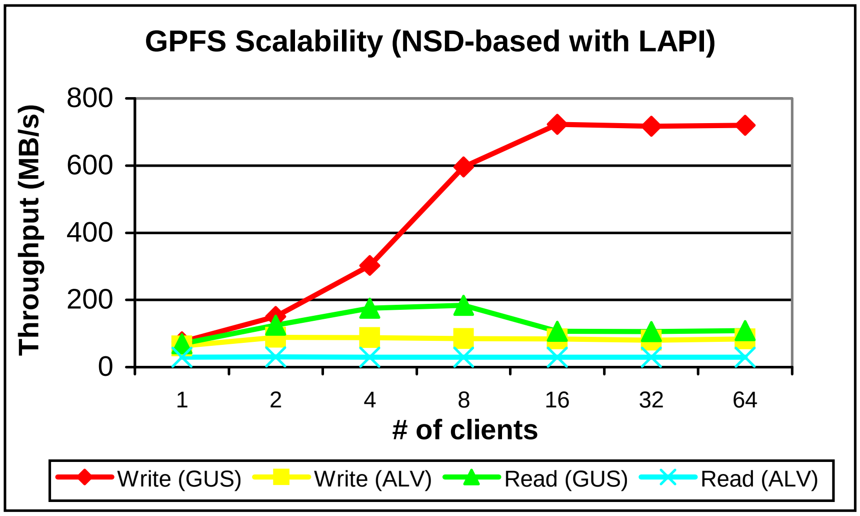


Figure 29. GPFS on Linux 1.3 performance.

6.7 Summary

GUPFS has evaluated four file systems: Sistina GFS 5.1 and 5.2 Beta, ADIC StorNext File System 2.0 and 2.1, Lustre File System 0.6 (Release 1.0 Beta 1), and IBM GPFS 1.3 File System for Linux. In addition, we plan to evaluate several others in 2004, including the IBM TotalStorage SAN File System (also known as StorageTank), the Ibrix file system, and the Panasas file system, (contingent upon availability and time resources). Of the four systems examined, three continue to be possible candidates for deployment. Sistina has been eliminated.

The Sistina GFS has proven to be a disappointment, mostly because of its lack of scalability in simultaneous concurrent writes to the same file, the company's lack of interest in supporting platforms other than Linux, and the lack of other functionalities (e.g., DMAPI and shared I/O support) that are important to the future GUPFS production environment. Therefore, Sistina GFS is no longer a candidate for further investigation

The ADIC StorNext File System already provides support for multiple operating systems — Linux, SGI IRIX, Microsoft Windows, Solaris, and AIX. This multiple platform support will allow future tests of a shared file system in a truly heterogeneous hardware platform and OS environment. This unusual aspect of the ADIC StorNext File System makes it a plus and a definite candidate for future investigation.

Although more stable preliminary versions of Lustre appeared in the beginning of FY 2003, this early Lustre file system version failed to complete any of our standard MPTIO tests, with the exception of the Cache Write and Cache Read tests. In addition, Lustre is missing certain functionalities (DMAPI), and it only supports the Linux platform and there do not appear to be any plans to expand to other systems. However, Lustre is very popular in the HPC, DOE, and Open Source laboratory communities, so it will continue to be a candidate for further investigation. We expect that the large support base will encourage Lustre to improve.

IBM's General Parallel File System (GPFS) file system is now available on two systems at NERSC: the production Seaborg IBM SP system and the Intel-processor-based LBNL Alvarez Linux cluster. GPFS is currently only available on AIX and Linux, and cross-platform and multi-cluster support is as yet unavailable. The earlier obstacles of restrictive licensing and availability only for AIX have been resolved, making GPFS a viable candidate for GUPFS. Additionally, it is one of the few file systems that support DMAPI, the mechanism proposed by the GUPFS team to support HPSS integration and to provide the necessary hierarchical storage management (HSM) functionality. GPFS has also been one of the very few cluster/global file systems that have been available in production environments of large clusters for many years. Therefore, IBM GPFS will continue to be a candidate for further investigation.

7 Findings, Future Activities, and Conclusion

This section presents the findings we have reached, describes the anticipated activities of the GUPFS project, both in the near term and over the long term as GUPFS moves toward deployment, and presents a brief conclusion. During FY 2004, the focus of the GUPFS project will shift from primarily technology investigations to acquisition and deployment planning, with fewer and more focused evaluations.

7.1 Summary Findings

With the development of an evaluation methodology and benchmark suites, and with the updating of the GUPFS testbed system, the GUPFS project did a substantial number of investigations and evaluations during FY 2003. The investigations and evaluations have involved many vendors and products. From our evaluation of these products, it seems that most vendors and many of the products are more focused on the commercial market. Most vendors lack the understanding of, or do not have the resources to pay enough attention to, the needs of high-performance computing environments like NERSC. The following summary findings are based on the FY 2003 test evaluations:

1. Traditional storage devices do not scale in performance and are unsuitable for use in the GUPFS production environment.

   These traditional storage devices, such as the Dot Hill device or the Silicon Gear device, are usually offered as direct-attach storage devices, connected directly to the hosts. Architecturally, they are not designed for shared access in a SAN environment. Some of these storage devices do support multiple interface ports on each controller, but these ports are mainly for LUN fail-over in a high-availability setup. These traditional storage devices often have a limited number of interface ports (with limited aggregate bandwidth), which makes them not suitable for GUPFS.

2. High-performance storage devices, with the ability to export LUNs to all interfaces and to simultaneously drive all interfaces at their maximum capacity, are needed for block-based shared file systems. These storage devices must be able to scale as the number of host connections increases. Such storage devices are beginning to appear.

3. Fibre Channel switch interoperability between vendors is still a problem. The lack of support for the interoperability standards by some vendors makes it very difficult to create a SAN fabric using switches from multiple vendors. We face another challenge with the inter-switch links (ISL) — the interface between two Fibre Channel switches. Often, ISL trunking is only supported between switches of the same brand. Without trunking, a 2 Gb/s ISL will only be able to support very limited bandwidth for inter-switch traffic. However, the ISL bandwidth problem may be alleviated once the 4 Gb/s ISL or 10 Gb/s FC ISL becomes available.

4. Bridging between multiple fabric technologies is not only doable, but the performance is reasonably good. Successful bridging will allow us to create a hybrid fabric with multiple fabric technologies to support fabric connectivity for different existing and future system architectures, and at different price points for storage connections to various NERSC systems.

5. iSCSI is a good option for storage traffic. A major strength of iSCSI is that it can be used on any fabric or interconnect that supports IP network traffic. The software iSCSI option performs very well, although it is at the expense of increased system overhead. The iSCSI protocol hardware assist iSCSI HBA cards are not cost effective at this time and have lackluster performance. Due to the recent emergence of the iSCSI standard, no high-performance, scalable iSCSI-connected storage devices are currently available. Better, scalable, and more extensive bridging capabilities between iSCSI and FC are needed. iSCSI is a good option for storage traffic.

6. InfiniBand is another good option for storage traffic. Although the technology is still relatively immature, InfiniBand offers performance (high bandwidth and low latency) beyond that of either Ethernet or Fibre Channel, with even higher bandwidths planned.

7. Although much progress has been made in the file system arena, more work needs to be done. The file system technology remains the component with the highest risk. Most file systems need better support in the following areas:

    - Stability
    - Parallel I/O and meta-data performance and functionality
    - Scalability
    - DMAPI support for HPSS integration
    - Multiple-cluster, cross-platform/OS support

7.2 Near-Term Investigations

Near-term investigations and evaluations in FY 2004 will be continued in all three critical component technology areas: shared file systems, storage, and SAN fabrics, with the main focus being on multi-cluster, multi-fabric file system testing to prepare for Requests for Information (RFIs) and Requests for Proposals (RFPs), as well the planned deployment of the GUPFS solution at NERSC in FY 2006.

We expect near-term activities to include more focused evaluations of specific candidate file systems and important new technologies whose availability is anticipated in the GUPFS deployment timeframe. Additional near-term activities are expected to include investigations of issues important to deployment, such as networking, HPSS and HSM integration, multi-cluster and cross platform installation and operations, and initial acquisition and deployment planning.

Multi-Cluster File System Testing

Deploying a center-wide file system at NERSC involves a number of issues beyond those already investigated by the GUPFS project. These include operating the file system on multiple, otherwise independent systems and clusters, with different hardware architectures and different operating systems, while accessing shared storage using multiple fabrics and integrating with the center's networking infrastructure.

During FY 2004, the GUPFS project will investigate the following deployment-related topics to determine the suitability and feasibility of select file systems, and to begin determining the best operational methodologies for each method of deployment:

- Multiple clusters
- Different platforms and operating systems
- Different Interconnects (FC, iSCSI) and integration with network infrastructure

A central issue for deployment of a center-wide shared file system is the ability to access the shared file system and shared storage from multiple systems and clusters that are otherwise independent and only loosely coupled through LAN connections. A shared file system requires coordinated access to storage among systems that had previously been largely independent. This has many implications for network access, interconnect bridging, software compatibility, and security. In order to investigate these issues and to begin developing operating methodologies for NERSC , the GUPFS project plans to perform multiple system and cluster testing during FY 2004, including investigations using the GUPFS testbed and various Linux clusters. We will run shared file systems using GUPFS testbed storage and will involve the GUPFS testbed and the Alvarez and PDSF clusters (previous tests in conjunction with Alvarez involved making the GUPFS testbed part of the Alvarez cluster). These investigations will be run on several file systems — potentially the GPFS, Lustre, StorNext, and Panasas — and will explore many of the issues mentioned earlier, but in a context restricted to IA 32 architectures running multiple versions of Linux.

Another central issue for the deployment of a center-wide shared file system is the ability of client systems with different architectures to simultaneously access the same file system and storage. This includes clients with different hardware architectures, clients running different operating systems, or both, as would be expected in the heterogeneous NERSC center-wide environment. Issues to be examined include the ability of the shared file system to accommodate different hardware architectures, such as a mixture of 32- and 64-bit architectures, Intel IA32 architecture, AMD Opteron, and IBM Power 3 systems, and the various operating systems that are run on these architectures. In order to investigate these factors and to determine any potential limitations, we intend to conduct heterogeneous client tests during FY 2004 using the GUPFS testbed in conjunction with systems of dissimilar architecture and/or operating systems. These investigations will be done using file systems supporting simultaneous access by dissimilar client architectures. Such file systems will include IBM's GPFS (Linux and AIX) and SANFS (Linux and AIX), and ADIC's StorNext (Linux, AIX, and Solaris). For these investigations, we expect to use the GUPFS testbed, an IBM Power 3 system, such as Dev2, and a Sun Sparc system. It may also be possible to investigate the interoperability of the Intel IA32 Linux systems with the 64-bit AMD Opteron Linux systems.

Because of cost and technological reasons, we anticipate that the deployed center-wide file system will have to be accessed over multiple fabrics. It is likely that these fabrics will include Fibre Channel, Ethernet, and InfiniBand. Each of these fabrics has its own installation and operational requirements and issues that need to be explored in order to develop workable networking integration and operational methodologies. The difficulties of integrating these fabrics into the production environment are substantially increased when they are bridged together into a single fabric, as will likely be the case. In addition, these fabrics will have to be integrated with the existing center networks.

Among the operational and networking issues that need to be addressed are security and ensuring the integrity of the file system data and metadata. The GUPFS project plans to begin examining these concerns in conjunction with the NERSC networking staff during FY 2004 as part of the multiple system and multiple cluster testing, and the dissimilar hardware architecture and OS testing. These tests will require connecting the storage fabrics, system interconnects, and Ethernet fabric of the GUPFS testbed with systems external to the testbed via the NERSC networks. Such testing will provide both insight and experience in networking and security issues associated with a center-wide shared file system that can be used to develop operational methodologies, assist in the selection of the correct components, and aid in deployment planning.

The GUPFS project is interested in investigating, or re-examining, newer versions of a variety of technologies. The criteria for this examination are their actual availability, the amount of time they are available for examination, and their future relevance. Arranged by technology component, these items are:

File Systems

- IBM GPFS File System
- ADIC StorNext File System
- IBM TotalStorage SAN File System
- Panasas ActiveScale Storage Cluster
- Lustre File System
- IBRIX File System

SAN Fabric Technologies

- 4 Gb/s and 10 Gb/s Fibre Channel
- InfiniBand
- ISCSI HBA and Switch
- Myrinet Gigabit Ethernet Blade

Storage Systems

- DataDirect Networks

7.3 Longer-Term Investigations and Activities

While the GUPFS project will continue conducting some investigations in the near-term; it is also planning and preparing for other investigations and activities that need to be conducted over the longer term. Many of these activities relate to preparing for the second phase (FY 2005–2006) of the GUPFS project, while others involve continuation of the evaluation process and beginning to address additional functionality that is required for a successful deployment.

Planned longer-term investigations and activities include:

- Additional technology evaluations from all component technology areas. Which additional technology evaluations will be conducted depends on the outcome of the near-term evaluations, on future developments in the component technology areas, and the narrowing selection of plausible candidate solutions.

- Integration of HPSS with various shared file systems. This ambitious activity is focused on incorporating hierarchical storage management (HSM) functionality into the most promising candidate shared file systems through the XDSM/DMAPI interface and capability of HPSS. This will require extensions not only to the candidate file systems, but also to HPSS. The integration may also involve changes to all the direct HPSS movers to natively mount the shared file systems and to transfer and/or migrate data to and from the file system by directly reading and writing files on it. As of the end of FY 2003, only IBM's GPFS provided any usable DMAPI support. ADIC has decided not to expose its DMAPI interface and wants to eliminate it. Consequently, we probably will focus on GPFS for these integration activities. Even more extensive integration of HPSS with the shared file systems is also being explored, although it is probably beyond the direct scope of the GUPFS project.

Some of the second-phase planned activities are:

- Finishing up the GUPFS Requirements Document.
- Planning and preparing for the procurement of the GUPFS solution, including preparations for RFI and RFP activities.
- Narrowing the selection of candidate technology components in preparation for the final selection of the technologies to be deployed.
- Planning for the consolidation of disk storage for the production computational and support systems, including organizational arrangements and operational methodologies for supporting centralized storage and a center-wide shared file system.
- Initiating planning of the phased rollout and deployment of the center-wide shared file system and consolidated, centrally managed storage.

7.4 Conclusion

The GUPFS project made a substantial amount of progress during FY 2003. We continued to developed and refine our evaluation methodology and benchmark suites and updated the GUPFS testbed system. The project did a substantial number of investigations and evaluations involving many vendors and products. Bridging between multiple fabric technologies is not only doable but is performing reasonably well, InfiniBand and iSCSI are good options for storage traffic, sustained file system performance in excess of 1 GB/s was achieved with several shared file systems (sustained read performance of 1140 MB/s with the ADIC StorNext file system; sustained read rate of 1484 MB/s with IBM's GPFS 1.3 file system).

Our evaluations also confirmed that most vendors and many of the products are more focused on the commercial market. Vendors lack the understanding of, or do not have the resources to pay enough attention to, the needs of high-performance computing environments such as NERSC.

Much remains to be done in both the near term and longer term to ensure a successful conclusion to the GUPFS project. With careful planning, diligent effort, and continued technological advances, particularly in the file system arena, this should be achievable. Our findings in all three critical component technology areas (shared file systems, storage, and SAN fabrics) give us confidence that we will attain our goal to provide a scalable, high-performance, high-bandwidth, shared file system for all of the NERSC production computing and support systems.

Appendix A: GUPFS Testbed Configuration

The GUPFS project uses a testbed system to conduct investigations and evaluations of the component technologies needed for a center-wide shared file system, and to explore these components interactions. In addition to these uses, we have employed the testbed to develop the GUPFS benchmark methodology and the actual benchmark codes used to conduct the technology evaluations. The testbed continues to be a useful resource in attracting the attention of component technology vendors and developing relationships with a number of these vendors.

The testbed we used during FY 2003 was the expanded testbed upgraded at the end of FY 2002. This upgrade is detailed in the GUPFS Project FY 2002 report [1]. This testbed was designed and built to provide sufficient hardware and computational resources to support the evaluation of multiple new component technologies, and to provide the underlying SAN fabric and storage resources with sufficient aggregate performance to stress-test existing and emerging shared file system technologies. This design emphasized the extensibility of the testbed system in order to accommodate future technology developments. In this regard, the testbed proved to be very effective. A variety of shared-file systems were successfully tested, a number of fabric components were integrated and tested throughout the year, and additional storage solutions were bought in and evaluated.

The base configuration of the GUPFS testbed during FY 2003, and the changes in that configuration throughout the year are presented in the following sections.

A.1 FY 2003 Initial Testbed Configuration

The GUPFS FY 2003 testbed system presented a microcosm of a parallel scientific cluster — dedicated computational nodes, special-function service nodes, and a high-speed interconnect for message passing. It used an internal jumbo frame Gigabit Ethernet as the primary high-speed message passing interconnect. An internal 10/100 Mb/s Fast Ethernet LAN was employed for system management and NFS distribution of the user home file systems. The testbed supplied Fibre Channel as the base SAN fabric, as well as Fibre Channel storage, and a variety of alternative fabrics and bridges between these fabrics.

During FY 2003, the testbed was configured as a Linux parallel scientific cluster, with a management node, a core set of 16 dedicated dual Pentium-4 compute nodes, a set of six special-purpose dual Pentium-4 nodes, and a reserve of five auxiliary dual Pentium-3 compute nodes from the original FY 2002 testbed. The Fibre Channel SAN fabric was expanded extensively with the addition of two 16-port 2 Gb/s FC switches. The Gigabit Ethernet used as the message passing interconnect for parallel jobs was also expanded to support the increased number of nodes and iSCSI testing. A picture of the FY 2003 testbed appears on the following page as Figure A-1. The FY 2003 testbed configuration is shown in Figure A-2 (page 53).

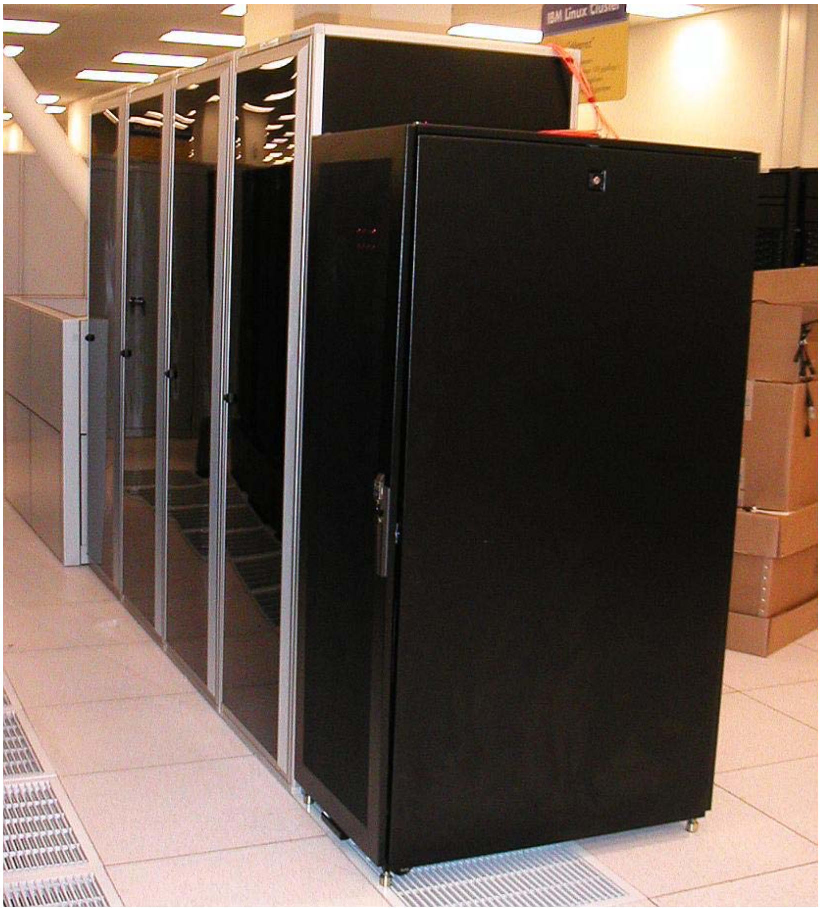


Figure A-1. The FY 2003 testbed, with the NetStorager shown in front.

The following major components were included in the FY 2003 testbed:

System nodes

- Twenty-two dual Pentium-4 nodes: sixteen in 2U cases and six in 4U cases (these are described in greater detail later in this section)
- Six dual Pentium-3 nodes in 4U cases

Fabric

- Ethernet
  - One 32-port Extreme 7i Gigabit Ethernet switch
  - One 16-port Extreme 5i Gigabit Ethernet switch
  - Two 10/100 Ethernet switches for system management
- Fibre Channel
  - Two 16 port 2 Gb/s Fibre Channel Switches (Brocade SilkWorm 3800 and Qlogic SANbox2-16)
  - One 16 port 1 Gb/s Fibre Channel Switch (Brocade SilkWorm 2800)
  - One Cisco SN5428 iSCSI Router fabric bridge to Ethernet
- Myrinet
  - One Myrinet 2000 8-port switch with eight host interface cards
- InfiniBand

- One InfiniCon ISIS InfinIO 7000 1x InfiniBand switch, with eight 1x HCA host adapters, and fabric bridge modules for Fibre Channel and Gigabit Ethernet

Storage

- A EMC CLARiiON CX600 disk subsystem
- A Dot Hill 7124 RAID disk subsystem
- A Silicon Gear Mercury II RAID subsystem
- A Chaparral A8526 RAID subsystem with attached storage

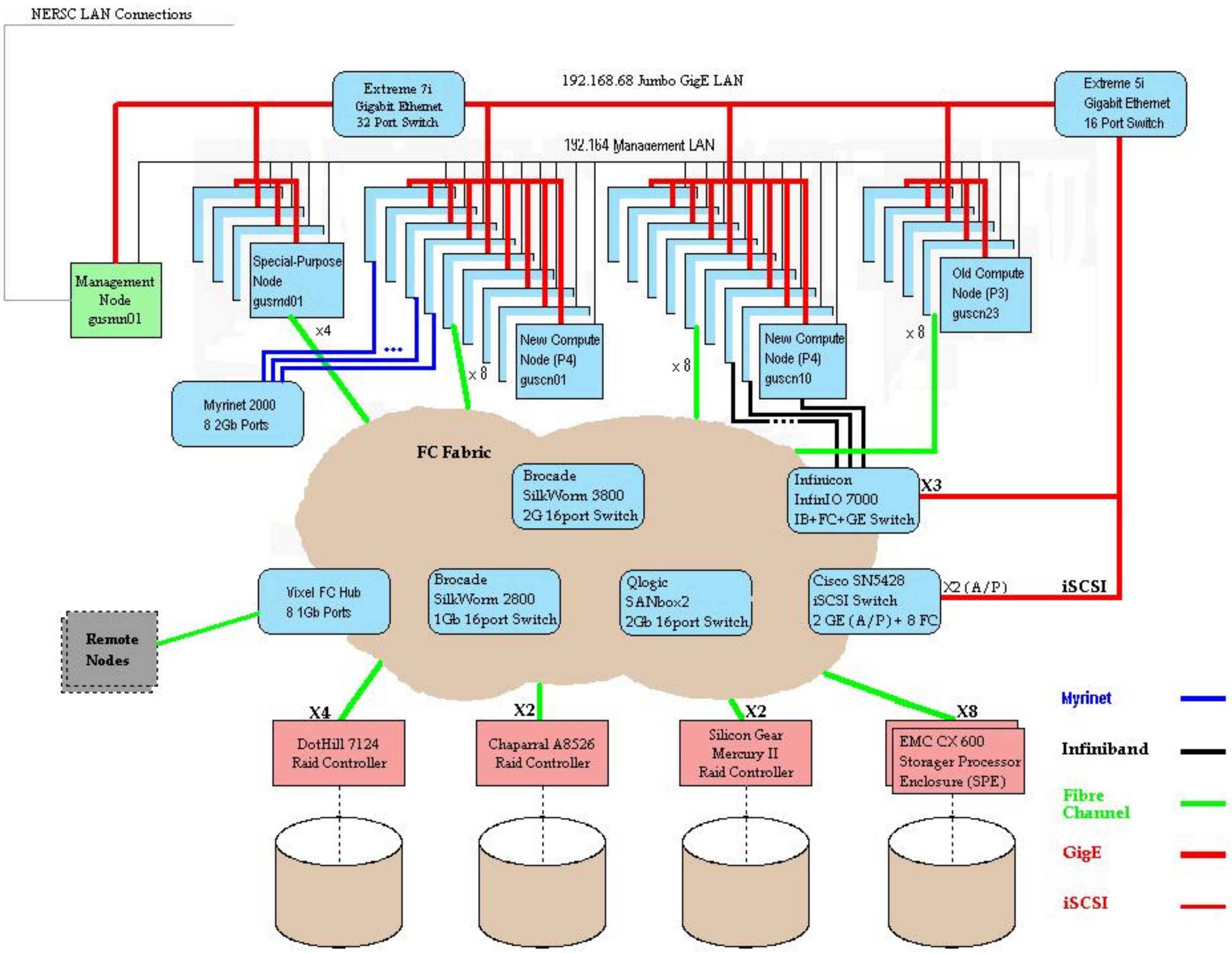

NERSC LAN Connections
Extreme 7i Gigabit Ethernet 32 Port Switch
192.168.68 Jumbo GigE LAN
Extreme 5i Gigabit Ethernet 16 Port Switch
192.164 Management LAN
Management Node gusmn01
Special-Purpose Node gusmd01
x4
Myrinet 2000 8 2Gb Ports
x 8
New Compute Node (P4) guscn01
x 8
New Compute Node (P4) guscn10
x 8
Old Compute Node (P3) guscn23
FC Fabric
Brocade SilkWorm 3800 2G 16port Switch
Infinicon InfinIO 7000 IB+FC+GE Switch
X3
Vixel FC Hub 8 1Gb Ports
Brocade SilkWorm 2800 1Gb 16port Switch
Qlogic SANbox2 2Gb 16port Switch
Cisco SN5428 iSCSI Switch 2 GE (A/P) + 8 FC
X2 (A/P)
iSCSI
Remote Nodes
X4
DotHill 7124 Raid Controller
X2
Chaparral A8526 Raid Controller
X2
Silicon Gear Mercury II Raid Controller
X8
EMC CX 600 Storager Processor Enclosure (SPE)
Myrinet
Infiniband
Fibre Channel
GigE
iSCSI

Figure A-2. FY 2003 base GUPFS testbed configuration.

The new Pentium-4 nodes all utilize the same motherboard and are configured similarly. The only differences among them are the sizes of the cases in which they are installed. Sixteen of the new technology nodes were put in 2U cases in order to save space, eliminating the need to buy more than one additional cabinet. Six of the new technology nodes were put in 4U cases to allow standard-height-profile peripheral component interconnect (PCI) cards to be installed for the Myrinet 2000 Host interfaces, Intel PRO/1000 T IP iSCSI cards, and early 1x InfiniBand HCAs for the InfiniCon InfinIO 7000 fabric bridge. The need to have PCI-X slots on the motherboard to support the high-performance FC, InfiniBand, and Gigabit Ethernet cards dictated the class of motherboards and processors that were acquired, as the only motherboards available with PCI-X buses were relatively high-end server motherboards.

- All Pentium-4 nodes, regardless of the size of their case had the same base configuration. This configuration consisted of the same motherboards, dual 2.2 GHz Pentium IV Prestonia Xeon CPUs, 2 GB of DDR memory, 10/100 and Gigabit Ethernet interfaces, 36 GB SCSI disks, and Qlogic 2340 2 Gb/s Fibre Channel HBAs.
- All Pentium-3 nodes had the same base configuration. This consisted of identical motherboards, dual Pentium III 1 GHz CPUs, 1 GB of memory, 10/100 and Gigabit Ethernet interfaces, and 18 GB SCSI disks. One of the Pentium-3 nodes was configured as a management node and had additional 10/100 and Gigabit Ethernet interfaces. The remaining five Pentium-3 nodes were configured as auxiliary compute nodes with Qlogic 2310 2 Gb/s Fibre Channel HBAs

All Pentium-4 nodes, regardless of the size of their case, are configured with:

- Supermicro P4DP6 motherboards with six PCI-X slots, two of which are 133 MHz capable
- Dual 2.2 GHz Pentium IV Prestonia Xeon CPUs
- 2 GB of DDR PC2100 ECC memory
- Dual onboard Intel PRO/100 Ethernet interfaces
- Dual onboard U160 Adaptec SCSI controllers
- Onboard VGA graphics
- One 36 GB Ultra 160 LVD 10K RPM SCSI disk drive
- One Qlogic qla2340 133 MHz PCI-X Fibre Channel HBA (low or standard profile)
- One Intel PRO/1000 XT 133 MHz PCI-X Gigabit Ethernet NIC (low or standard profile)

All Pentium-3 nodes shared a common base configuration. The management/interactive node and the computational nodes differed only in that the computational nodes contained a 2-Gb/s Fibre Channel interface card, while the management node contained additional Fast Ethernet cards and an additional Gigabit Ethernet card.

All of the Pentium-3 nodes were installed in 4U rack mount cases and had the following base configuration:

- Intel Server Board STL2 motherboards, with two 64-bit 66 MHz PCI slots with additional 32-bit PCI slots
- Dual Pentium III 1 GHz CPUs
- 1 GB of PC133 ECC memory

- Onboard VGA (video graphics array) graphics
- Onboard U160 Adaptec SCSI controllers
- One onboard Intel PRO/100 Ethernet interface
- One 18 GB Ultra 160 LVD 10 k RPM SCSI disk drive
- One Qlogic qla2200 64-bit Fibre Channel Optical HBA (compute nodes only)
- One Intel PRO/1000 T 64-bit PCI Gigabit Ethernet NIC (the management node had two)

The identical base configuration of all the Pentium-4 nodes, including those intended as special-purpose nodes, allowed them to be used at times as compute nodes for the purpose of scalability testing. Four of the new nodes were configured to be special-purpose nodes. These special-purpose nodes had the same basic hardware and software configuration as the dedicated compute nodes. The four special-purpose nodes are configured to perform the following functions:

- Code development and benchmark debugging
- Metadata and lock manager services
- A dedicated installation target for developing and testing new kickstart configurations
- A storage server for testing distribution of shared file system with NFS gateways

Two additional Pentium-4 nodes were initially reserved for transient special-purpose usage, such as running the InfiniCon InfiniBand subnet manager, which initially ran under Windows 2000. When InfiniCon’s subnet manager became able to run under Linux, both of these nodes were reconfigured as general purpose compute nodes and were usable in evaluations.

The increased scale, many advanced technology components, and flexible and expandable design of the updated GUPFS testbed will enable many interesting and important evaluations to be conducted over the next several years. These evaluations should lead to the selection of the best and most appropriate component technologies for the rollout of a high-performance shared file system during the second phase (FY 2005–2006) of the GUPFS project.

All testbed nodes ran Linux, based on the RedHat 7.1, 7.2, 7.3, 8.0, or 9.0 distribution, depending on the requirements of the file system tested. The testbed supported parallel job submission and execution using Open PBS and utilized MPICH as the MPI implementation for parallel jobs. Portland Group C, C++, and various flavors of FORTRAN compilers provided the compilation and execution environment for the parallel jobs.

Independent Linux systems on individual nodes are automatically installed through PXEboot kickstart mechanisms. This allowed for multiple, completely different system images to be present on each of the nodes, enabling rapid reconfiguration of the testbed so that it could quickly switch among different software environments, each of which was needed to conduct a different evaluation.

All nodes except the management node were connected to a switched 2 Gb/s Fibre Channel SAN fabric by 2 Gb/s Qlogic 2300 family FC Host Bus Adapters (HBAs). The 2 Gb/s FC fabric was entirely optical. The 1 Gb/s FC copper fabric was retired and replaced with an optical fabric, except as necessary to attach the original 1 Gb/s FC storage to the 1 Gb/s Brocade SilkWorm 2800 switch.

In addition to the testbed nodes being attached to the Fibre Channel SAN fabric, the various storage devices, both permanent and under evaluation, were attached to the same switched FC fabric, as were a number of fabric bridges. These fabric bridges included the Cisco SN5428 Storage Router bridging between Gigabit Ethernet iSCSI storage traffic from hosts and FC fabric attached storage devices, and the InfiniCon InfinIO fabric bridge between InfiniBand attached hosts and FC fabric attached storage.

A.1.1 Storage Configuration

The disk storage devices connected to the Fibre Channel SAN fabric were:

- A dual-controller DotHill 7124 RAID subsystem, with an expansion cabinet
- A dual-controller Silicon Gear Mercury II RAID subsystem
- A single-controller Chaparral A8526 RAID subsystem with attached storage
- A dual-controller EMC CLARiiON CX 600 RAID subsystem with storage

Each of the RAID controllers had two or more Fibre Channel ports for connecting to the switch, and these FC ports could be used simultaneously. All four storage devices supported various RAID configurations and utilized similar 10,000 RPM 73 GB disk drives. The DotHill contained 20 drives, the Silicon Gear 12 drives, the Chaparral 10 drives, and the EMC 30 drives. Total unformatted SAN attached storage capacity was approximately 5.3 terabytes (TB), with a nominal maximum of 4.2 TB of formatted RAID 5 storage. The DotHill and Silicon Gear were both limited to 1 Gb/s FC interfaces, while the Chaparral and EMC supported 2 Gb/s FC interfaces

The storage configuration was chosen to enable the exploration of the relative performance, reliability, and interoperability of multiple storage vendors’ products. The quantity and character of the storage was dictated by:

- The technology available at the time of acquisition
- The desire to be able to achieve maximum performance from each storage controller
- The desire to be able to explore the Linux support for file systems greater than 2 TB on 32-bit architectures when such support became available
- Price

A.1.2 Testbed Configuration Changes during FY 2003

A number of changes in the testbed configuration occurred during FY 2003. These included upgrading the InfiniCon InfinIO switch and HCAs from 1x (2.5 Gb/s) to 4x (10 Gb/s) InfiniBand, exchanging the Myrinet 2000 PCI based Rev C host interface adapters for higher performance PCI-X Rev D cards, and connecting the Alvarez cluster 10/100 management network with the GUPFS testbed Gigabit Ethernet fabric. Another modification was the exchange of three of the five Intel iSCSI HBAs for a newer version that could run with more up to date Linux kernels, allowing the iSCSI HBAs to be tested in conjunction with the newer file systems.

The Myrinet 2000 host adapter exchange allowed all eight Myrinet 2000 host adapters to be installed and the full capabilities of the testbed Myrinet fabric to be investigated. The new Rev D adapters were available in low profile form factor (2U) which allowed all of them to be installed in nodes. The previous adapters were full height (4U). Since the testbed only had six 4U Pentium-4 nodes, only six of the eight original adapters could be installed. Once all eight Rev D adapters were installed, the GUPFS project proceeded with plans to install the GUPFS 8 port Myrinet switch blade with the Alvarez Myrinet switch, making the eight connected GUPFS nodes part of the Alvarez system. This allowed testing of GPFS 1.3 for Linux using Alvarez compute nodes with GUPFS nodes as high-performance Network Storage Devices (NSDs) in lieu of Alvarez's low performance storage. This also allowed GPFS testing at a large scale (64 or more nodes) in conjunction with 1600 MB/s storage bandwidth, a combination unachievable by either system alone. In addition, it provided us an opportunity to begin investigating shared file systems running on multiple systems.

During FY 2003, the GUPFS testbed InfiniBand configuration received several updates. InfiniCon upgraded the HCAs and switch modules to 4X (10 Gb/s), enabling early investigation of storage transfers over 4X InfiniBand. As in the case of the 1x IB HCAs, the initial 4x HCAs were full height (4U). Because the 4x HCAs required PCI-X slots, and because the testbed only had six 4U Pentium-4 nodes with PCI-X slots, the InfiniBand fabric deployment was limited to six systems, although components were available for eight systems. In one of a number of software upgrades, it became possible to run the InfiniCon IB subnet manager under Linux. This allowed the system running the subnet manager under Windows 2000 to be converted back to Linux and used as a compute node in evaluations.

A second InfiniCon hardware upgrade brought in a second generation of 4x HCAs. In addition to providing higher performance, these HCAs were available in low profile (2U) form factor. This made it possible for us to install all eight of the new HCAs in 2U Pentium-4 nodes, to fully populate the configuration for the first time. This allowed more meaningful scalability numbers to be obtained, enabling more direct and clear interconnect/fabric performance comparisons.

Appendix B: Updating the Testbed Configuration for FY 2004

The testbed system has provided us with a useful facility for developing the benchmark methodology and special benchmark codes for the GUPFS project. It has also been useful in helping to establish the credibility of GUPFS with technology vendors, and in building relationships with various technology vendors. However, with the inexorable progression of technical advancements, it became apparent that the testbed was inadequate in size for conducting the types and levels of technology evaluations needed for the GUPFS project in FY 2004

Technological advancements over the last year have outstripped the ability of the existing testbed to incorporate them. Experience with the testbed and attempts to integrate new storage and fabric technology into it demonstrated that more nodes were needed in the testbed to allow emerging advanced technologies to be integrated into it for evaluation, even in the near term. Given the technologies that the GUPFS project plans to begin evaluating in FY 2004, it was clear that the existing testbed could not accommodate their inclusion.

In addition to the testbed being limited in accommodating new technologies, practical experience during FY 2003 indicated to us that the testbed could be improved in a number of ways that would increase the speed at which file system evaluations could be conducted, in conjunction with specific combinations of fabrics and high-performance storage. Foremost of these was that the testbed should be able to conduct multiple, simultaneous, independent evaluations with node sets of various sizes. Next was that the testbed be able to reconfigure the fabric and storage connections much more easily. Other important areas to improve were to greatly increase the Gigabit Ethernet connectivity for iSCSI and cross-fabric testing, and to only have a single node type used in testing. Based on these needs, we designed a testbed upgrade to facilitate more simultaneous evaluations and much more rapid and easy reconfiguration. The design considerations for the update to the testbed and discussed in the following section.

We completed the improved testbed system design at the end of the third quarter of FY 2003. We then identified, tested, and procured the components during the fourth quarter. These were assembled and integrated them into the existing testbed system at the end of the fourth quarter of FY 2003 in preparation for the planned FY 2004 activities. The configuration of the expanded testbed system is discussed in detail later in this section.

B.1 Updated Testbed Design Considerations

The design of the upgrade to the GUPFS project testbed system was predicated on a number of factors. These included (1) the lessons learned and (2) limitations encountered from using the FY 2003 testbed to integrate and test new technologies, and (3) the new and emerging technologies expected to be investigated in the next year; these are discussed below, along with (4) a brief review of other factors affecting the upgrade.

Lessons Learned

The FY 2003 testbed included many design features that improved its usability and utility over the FY 2002 testbed. However, a number of lessons were learned through the use of the updated FY 2003 testbed and the evaluations conducted on it. These directly impacted the design of the testbed upgrade for FY 2004. The major lessons learned are:

- **Multiple evaluations need to be done simultaneously.** An evaluation of a technology component is complex, and time consuming to both set up and conduct. The reality of the situation is that a technology component cannot be evaluated in isolation. In general, one of the GUPFS component technology types can only be evaluate in conjunction with one or more of the other component types. For example, to evaluate a file system, it is necessary to test it in conjunction with storage and some fabric or interconnect connecting the clients to the storage, In addition, each component usually has required versions of software, necessitating client systems being configured with specific OS and driver versions and specific fabric connectivity. Setting up the required configurations to conduct specific evaluations is time consuming, often in the order of several weeks or a month. Often, a single evaluation takes several months and requires dedicated resources. Because of the number of important component technology issues that need to be investigated, particularly in the file system arena and multi-cluster configurations, and because of the rapidly changing component technologies, it is vital to have enough resources to be able to have extended unrelated evaluations in progress simultaneously.
- **Only a single type of compute node should be used.** The FY 2003 testbed contained two types of compute nodes — the 18 dual Pentium-4 compute nodes and the 5 legacy dual Pentium-3 compute nodes. Because of the limited number of available compute nodes and the high volume of items to evaluate or investigate, it was necessary to use both types of nodes for evaluations. This predictably caused problems in a number of areas. First, the different node architectures required distinct software configurations be built, installed, and tested for each type, thereby increasing the administrative burden for the testbed. Secondly, the hardware components and performance differences between the types of nodes made it very difficult to compare results obtained from evaluations, and made evaluations using both types of nodes at the same time too difficult to understand. The solution to this difficulty is to standardize on a single node type for use as compute nodes in evaluations, and to obtain adequate numbers of them to conduct the necessary number of evaluations simultaneously at adequate scale.
- **A separate interactive node is needed.** The FY 2003 testbed used a single dual Pentium-3 node as both a management node for running and maintaining the testbed environment and as an interactive node for the project members to log into and from which to conduct tests and evaluations. With the increased size and complexity of the testbed and the increased number of evaluations in progress at one time, it became apparent to us that a single node of this type could not perform both functions without impacting one or the other. In addition to periodic failures and substantial delays caused by overloading the management node, the launching of long-running benchmarks frequently prevented timely maintenance activities, such as rebooting the management node to clear software problems or activate patches. Another reason to separate the management and interactive functionality was to reduce the possibility of the management node accidentally being destroyed by the setting up evaluations and running of benchmarks by the project members, both activities that frequently required running with elevated privileges, leading to a number of close calls. Management node functionality is much more difficult and time consuming to configure and install than interactive node functionality. The necessity of separating the management and interactive node functionality contributed to the decision to stop using any of the Pentium-3 nodes as compute nodes, and to assign them to testbed support roles.

Limitations

In addition to lessons learned, the design of the FY 2004 testbed was also influenced by certain *limitations* encountered while using the FY 2003 testbed. These limitations include:

- **The EMC CX600 did not meet expectations.** The EMC CX 600 storage device did not live up to expectations regarding performance scalability. This was partly the result of unexpected architectural limitations and partly the result of configuration constraints. As a consequence, neither the expected aggregate bandwidth nor the desired scalability was achieved. This made the GUPFS project dependent on other scalable storage that was being evaluated, such as the Yotta Yotta NetStorager and the 3PARdata Inserv. With the completion of the 3PARdata evaluation and the Yotta Yotta extended beta test at the beginning of the 4th quarter of FY 2003, the GUPFS project faced the prospect of not having storage that was of high enough performance or sufficiently scalable to use for the file system evaluation planned for the end of FY 2003 and for all of FY 2004.
- **The number of Gigabit Ethernet switch ports was inadequate.** The GUPFS testbed's Gigabit Ethernet fabric quickly became limited by the 48 available switch ports. The available ports were quickly consumed by the testbed systems, inter-switch links between the two Gigabit switches, original iSCSI router and Intel iSCSI HBA connections, and the InfiniCon InfiniBand to Gigabit Ethernet bridges. With the introduction of additional equipment requiring Gigabit Ethernet connections, the number of available ports was oversubscribed by at least a factor of two. This resulted in the serialization of evaluations and made it necessary to disconnect various equipment in order to connect and test other equipment. The additional equipment included Topspin InfiniBand to Gigabit Ethernet fabric bridges, Panasas storage devices, Adaptec iSCSI HBAs and TOE cards, and the inter-switch links to the Alvarez management Ethernet switch. Based on the need to maintain adequate inter-switch bandwidth, a 3x expansion of the Gigabit switch ports was needed.
- **Reconfiguring the physical connections between storage, fabric elements, and client systems became extremely difficult.** In order conduct evaluations of various file systems, fabric components and bridges, and storage combinations, and to conduct evaluations of various loaner equipment, it was necessary to change the physical fiber optical connections of the equipment connected to the three 16 port Fibre Channel switches in order to connect the correct set of components. This was made extremely difficult by the rigidity of the bundles of fiber cables and the fragility of the connectors. Moving a fiber between one switch and another often required major efforts to obtain adequate slack to permit the connector to plug into another switch. This was particularly a problem when we were evaluating new equipment such as Fibre Channel switches, which might be physically mounted in a different cabinet. Making such connections required that substantial time be devoted to rebundling fibers or stringing new ones. In addition, replugging the connectors exposed them to mechanical failure (the 2 Gb/s SFPs are especially fragile), and to contamination of the optics with dust. Another problem with making changes to the fiber configuration was that it soon be came very difficult to determine what was connected to what and which fibers were active. A fiber patch panel is needed to resolve these problems.
- **A single dedicated metadata server node is not enough.** It became apparent that a single special-purpose node acting as a dedicated metadata server was inadequate. Nearly all of the shared file systems being tested required either a metadata server or a centralized lock manager. As a consequence, testing of these file systems became

serialized because of the single metadata/lock server. Frequently, more than one of these file systems was in some stage of the installation and evaluation cycle; sometimes a file system would be undergoing several tests at once, each instance having different hardware and/or software configurations. This required either very careful alternating of test segments, or the suborning other nodes to become additional metadata servers. Using other testbed nodes as secondary metadata servers impacted other activities by limiting the number of nodes available to them. Similar constraints applied to testing configurations supporting metadata/lock server redundant operation and failover. At least one more special purpose Pentium-4 node with an identical configuration needs to be dedicated to the metadata/lock server role.

- **A full complement of eight 4U Pentium-4 nodes is needed.** The GUPFS testbed only had six 4U Pentium-4 nodes. Most new technologies are initially implemented on standard height (4U) PCI/PCI-X cards, and only in the second or third generation of the technology do low-profile (2U) cards become available. The initial Gigabit Ethernet, 1x InfiniBand, 4x InfiniBand, Myrinet 2000, iSCSI HBA, Gigabit Ethernet TOE cards, and 1 Gb/s Fibre Channel cards were all standard height cards, requiring 4U cases. Most fabrics provide switching capabilities as powers of 2, and frequently with 8 ports as a minimum, leading to a standard purchase of an 8-port switch and 8-host interface cards, as 4-host cards provide little in the way of insights about scalability. Earlier constraints limited the testbed to six 4U Pentium-4 nodes, preventing eight-way evaluations when standard height interface cards were required. Adding two more 4U Pentium-4 nodes would enable eight-way evaluations requiring standard height interface cards to be conducted, permitting more direct comparisons with other evaluations.

New and Emerging Technologies

The design changes for the FY 2004 testbed were influenced by the *new and emerging technologies* impacting the GUPFS solution, which are expected to be available for evaluation during the coming year. In this regard, several important issues need to be investigated in the near term, including:

- **Conducting cross-platform file system tests.** The GUPFS project plans to conduct cross-platform file system tests to explore functionality and deployment issues in a heterogeneous environment that involves multiple hardware and different OS architectures, which is designed to mimic the NERSC environment in which GUPFS will be deployed. These tests will require either incorporation of additional systems into the testbed, or opening up the testbed to other NERSC systems, both of which require additional fabric-switching capabilities.
- **Conducting multiple cluster file system tests.** The GUPFS project plans to conduct file system tests involving multiple clusters accessing the same file system and storage simultaneously, as is expected in the NERSC environment at deployment. PDSF, Alvarez, and Dev2 are likely candidate peer systems. This will require opening up and connecting the testbed to these other NERSC systems.
- **Evaluating 4 Gb/s and 10 Gb/s Fibre Channel.** Both 4 Gb/s and 10 Gb/s production quality Fibre Channel equipment will be becoming available in the time frame of the initial phase of GUPFS deployment. Because of the anticipated aggregate performance needs for production use — and as it is likely that the backend storage controllers, if not the storage itself, for most shared file system solutions will be Fibre Channel connected

— these technologies are likely to be important to a successful GUPFS deployment. As such, they need to be evaluated and understood.

- **Evaluating 10 Gb/s Ethernet.** The 10 Gb/s Ethernet technology is expected to be deployed during FY 2004. Because it is most likely that PDSF will be accessing the GUPFS file system over the Ethernet, 10 Gb/s Ethernet is a likely component of the deployed GUPFS solution and needs to be understood in a storage fabric context.
- **Evaluating Panasas file system and storage.** The Panasas ActiveScale File System is a very interesting object-based file system implemented over the Ethernet. Architecturally, it is quite similar to Lustre, but is more standards based, being implemented with a variant of iSCSI. The Panasas file system offering is integrated with Ethernet-attached storage devices specific to the file system, and can be accessed either through integrated NFS and CIFS gateways, or as part of a shared file system through the DirectFlow client software. The Panasas file system should be accessible over any IP-based fabric that can bridge to the Ethernet. This is a promising candidate file system and needs to be evaluated for GUPFS.
- **Evaluating the IBRIX file system.** The IBRIX file system is an interesting potential GUPFS file system solution that is based on federating the individual file systems of storage engines (SEs). The IBRIX file system is distributed over IP networks. It utilizes back-end SAN based storage. The IBRIX file system was originally scheduled to be available in preproduction versions for evaluation in FY 2003. This schedule has slipped into FY 2004.
- **Evaluating the IBM TotalStorage SANFS (StorageTank) file system.** The IBM StorageTank file system was renamed the TotalStorage SANFS file system at the end of FY 2003. SANFS is expected to become available as a product in the first half of FY 2004. It targets very large numbers of client systems, and supports multiple hardware architectures and operating systems. It uses metadata servers in conjunction with block storage accessed via iSCSI over IP networks (making it largely fabric agnostic), or accessed directly by Fibre Channel. The ability to access remote tanks over the WAN is being developed. SANFS is an extremely promising GUPFS candidate, although quite young, and needs to be investigated thoroughly.
- **Further iSCSI investigations.** The iSCSI protocol is making an appearance in several promising shared file systems. It also provides a cheap mechanism for accessing block storage over inexpensive fabrics, although at the expense of the higher processor overhead. With the ability of most fabrics and interconnects to perform IP transfers, and with the ability of most fabrics to bridge to the Ethernet, iSCSI may facilitate the implementation of heterogeneous fabrics directly tied into cluster interconnects. However, it needs much more investigation as its availability and use expand.

Additional Considerations

Other considerations that affected the design of the FY 2004 testbed include:

- **InfiniBand technology refresh needed.** The testbed InfiniBand technology needs to be refreshed. Current second-generation 4x InfiniBand equipment, particularly HCAs and Fibre Channel and Gigabit gateways need to be acquired. The original 1x IB equipment and the loaner first generation 4x IB equipment are no longer supported.
- **Multiple management nodes needed.** The testbed needs at least two management nodes. The management node is currently a single central point of failure in the testbed,

and is extremely difficult and complex to configure. A second management node is needed to ensure the testbed and GUPFS project evaluations and investigations can continue if the existing management node fails. In addition, the availability of a second management node would allow the management node software versions and configurations to be upgraded one at a time without disruption.

- **Gigabit Ethernet emerging as the standard fabric-to-others bridge.** Gigabit Ethernet is emerging as the common fabric to which all other fabrics and interconnects bridge. Because of this, a large number of Gigabit Ethernet switch ports are needed in the testbed, particularly in conjunction with the iSCSI, cross-platform, and multi-cluster file system testing planned for FY 2004.
- **The Myrinet to Gigabit Ethernet bridge would expand the file systems that can be tested with Alvarez.** Myricom announced a Gigabit Ethernet bridge blade for their Myrinet switches, with 8 Gigabit Ethernet ports. With such a fabric bridge, the number of shared file systems that could be tested on and in conjunction with the LBNL Alvarez Linux cluster increases substantially. IP-based file systems such as Panasas and IBRIX could be evaluated for scalability. Block-based file systems, such as StorNext, could be tested for scalability using iSCSI bridged to storage in the GUPFS testbed. In addition, a Myrinet upgrade to the Rev D card in late FY 2003 allowed low-profile PCI-X Myrinet 2000 cards to be installed, enabling the Myrinet network to be installed in 2U nodes, thus freeing up the 4U nodes for other uses and enabling the full use of the Myrinet switch with 8 hosts.

B.2 Updated Testbed Configuration for FY 2004

Design of the updated testbed configuration was completed at the end of the third quarter of FY 2003. This design was based on all of the considerations presented in the previous section. The central tenet of the updated configuration was to increase the number of simultaneous evaluations that could be conducted, increase the maximum scale of these evaluations, and simplify the process of physically reconfiguring the connectivity from testbed nodes to storage devices through various fabric components.

The updated configuration expanded the total number of Pentium-4 nodes from 22 to a total of 36. As in FY 2003, four of these Pentium-4 nodes were retained as dedicated special purpose nodes, although there were some changes in assigned functions. The remaining 32 Pentium-4 nodes were assigned as compute nodes dedicated to running benchmarks and conducting other investigations. To facilitate conducting multiple simultaneous independent evaluations, the 32 compute nodes were logically partitioned into four sets of eight. This logical partitioning allows up to four independent 8-way investigations to be conducted simultaneously, or in various size combinations, such as one 32-way, two 16-way, or one 16-way and two 8-way tests.

While the partitioning of the compute nodes into groups was at a logical level, there were some physical characteristics related to their partitioning. Each group of eight compute nodes was connected to a separate Dell Power Connect Gigabit Ethernet switch, which allowed the nodes in the group maximum communication performance among themselves. The Dell switches for each of the groups were then connected by four-way trunks to a central Extreme 7i switch. This allowed nodes in any of the groups to communicate with each other, but reduced aggregate bandwidth and increased latency. Another physical characteristic related to the partitioning of the nodes was the additional PCI-X fabric interface cards each node had. For a variety of reasons, the GUPFS project conducts evaluations of fabric interfaces/interconnects with a minimum of eight hosts for each fabric. In addition to Fibre Channel interfaces, present on all nodes except the management nodes, the GUPFS testbed contains three other sets of high-performance fabrics, each of which is connected to eight compute nodes. These additional fabrics are a Myrinet 2000 interconnect, and after the testbed upgrade, two 4x InfiniBand fabrics from different vendors. The three sets of nodes with extra fabric connections are each put into logically separate groups to facilitate the independent testing of these extra fabrics.

An additional element of the updated configuration was the dispensation of the original six Pentium-3 nodes. One of these has always been used as the testbed management node. The remaining five were used as compute nodes in FY 2002 and 2003, although in an auxiliary role during FY 2003. With the addition of more Pentium-4 nodes in the updated configuration, it was possible to stop using the Pentium-3 nodes as compute nodes and assign them to other supporting duties. One was to become a second testbed management node for redundancy and simplifying upgrades. Another was to become a dedicated interactive node, offloading this function from the management nodes for the reasons discussed earlier. The remaining three Pentium-3 nodes became dedicated development nodes for benchmark and analysis code development, and possible auxiliary HPSS integration investigation roles.

Another part of the upgraded configuration included the installation of a fiber-optic patch panel, allowing all fiber-optic ports to be centrally connected in a static configuration and then cross connected as necessary using easily movable fiber-optic patch cords. All fiber-optical Gigabit Ethernet, Myrinet, and Fibre Channel host adapters, switch ports, and device connections were hardwired into the central patch panel to simplify physical reconfiguration of the fabrics and connections.

The other major element of the updated configuration included purchasing the Yotta Yotta NetStorager as the standard high-performance storage to be used in evaluations in lieu of the disappointing CX 600, the previously mentioned Gigabit Ethernet switching capacity, additional Fibre Channel switching capacity, and 4x InfiniBand technology refresh from two vendors. A front view of the updated testbed for FY 2004 appears as Figure B-1. A rear view of the testbed, showing the nodes and cable connections, appears in Figure B-2. The updated FY 2004 testbed configuration is shown in Figure B-3. The Port Fibre optical patch panel is shown in Figure B-4.

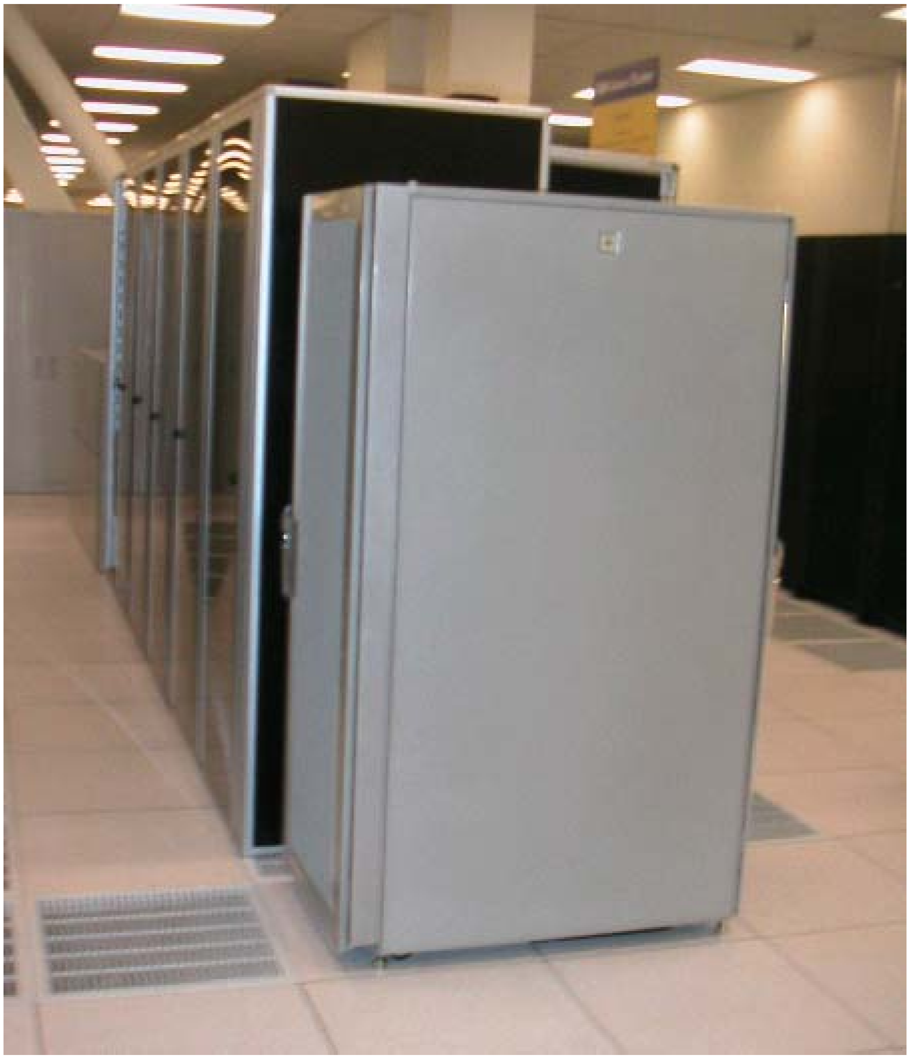


Figure B-1. The FY 2004 testbed, with the Refurbished NetStorager in front.

The following major components were added to the testbed as part of its technology upgrade for FY 2004:

- Fourteen additional dual Pentium-4 nodes: 12 in 2U cases and 2 in 4U cases (these nodes were identical to those already in the testbed.
- One 64-port 2 Gb/s Fibre Channel Qlogic SANbox2-64 switch with 48 ports
- Five 24-port Dell Power Connect 5224 Gigabit Ethernet switches
- One Yotta Yotta NetStorager GSX 2400 Disk Storage Subsystem
- InfiniCon ISIS InfinIO 7000 switch and fabric bridge 4x InfiniBand upgrade
- One Topspin TS90 4x InfiniBand switch and Fibre Channel gateway
- One Myrinet 2000 MS-SW16-8E switch line card with 8 Gigabit Ethernet ports
- A Fiber Optic patch panel and cables for Gigabit Ethernet, Myrinet, and Fibre Channel

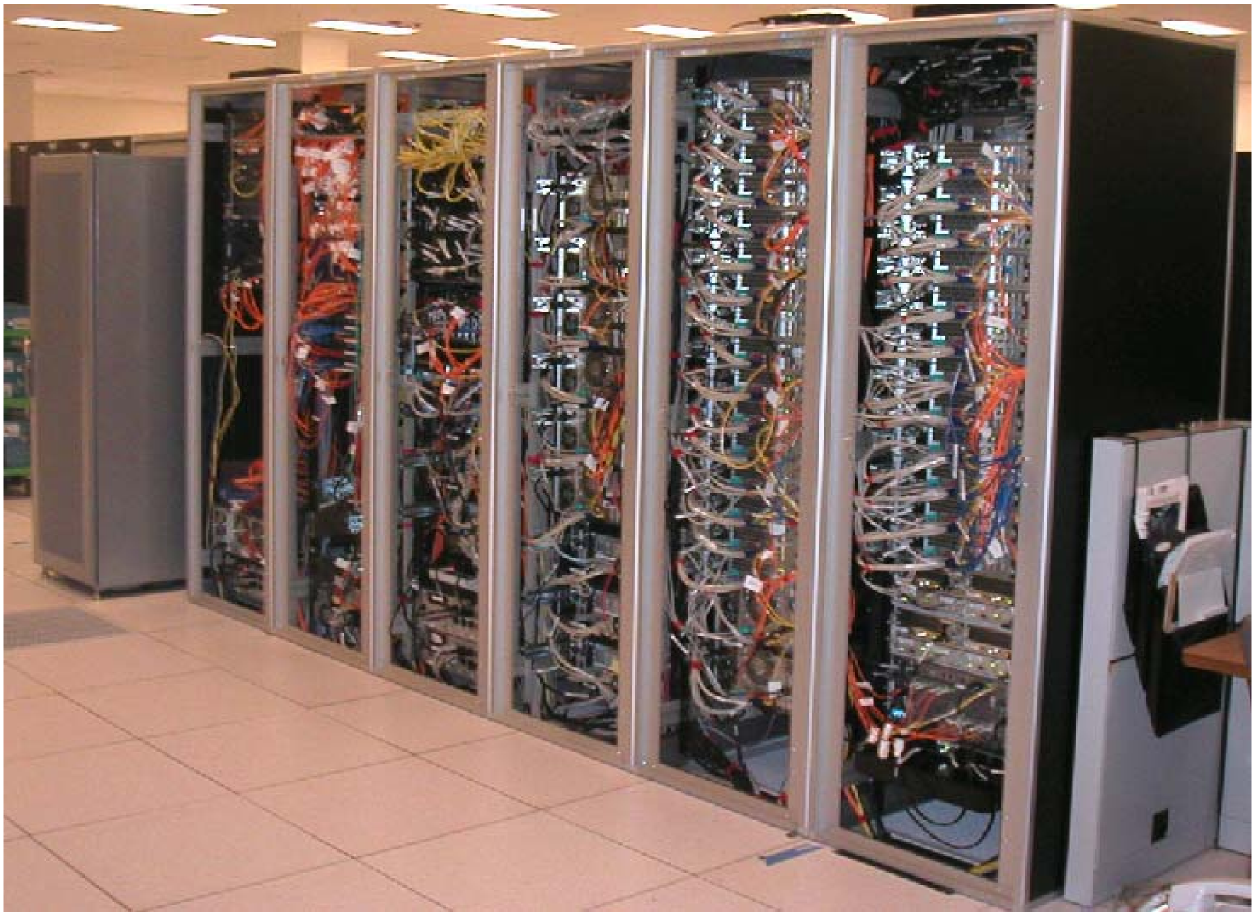


Figure B-2. Rear view of the FY2004 testbed.

The new Pentium-4 nodes obtained as part of the upgrade were as identical as possible to those obtained the previous year. They were configured with the same motherboard, the same quantity and performance grade of memory, the same 2.2 GHz Xeon CPU, and the same Intel Gigabit Ethernet and Qlogic Fibre Channel PCI-X cards. All of these components differed from the originals only in revision numbers. The nodes were all equipped with the same speed (10,000 RPM) and capacity (36 GB), and U160 SCSI disks, but from a different manufacturer.

Special efforts were made to configure the new Pentium-4 nodes to be identical to those obtained earlier. This was done to ensure uniformity of performance so that results obtained from both sets would be comparable and they could be intermixed without affecting evaluation results. The motherboard proved to be the most difficult to obtain as it was being phased out. However, a newer revision of the motherboard was acquired that showed nearly identical performance, which allowed the new and existing Pentium-4 nodes to be intermixed with negligible impact on the benchmark results.

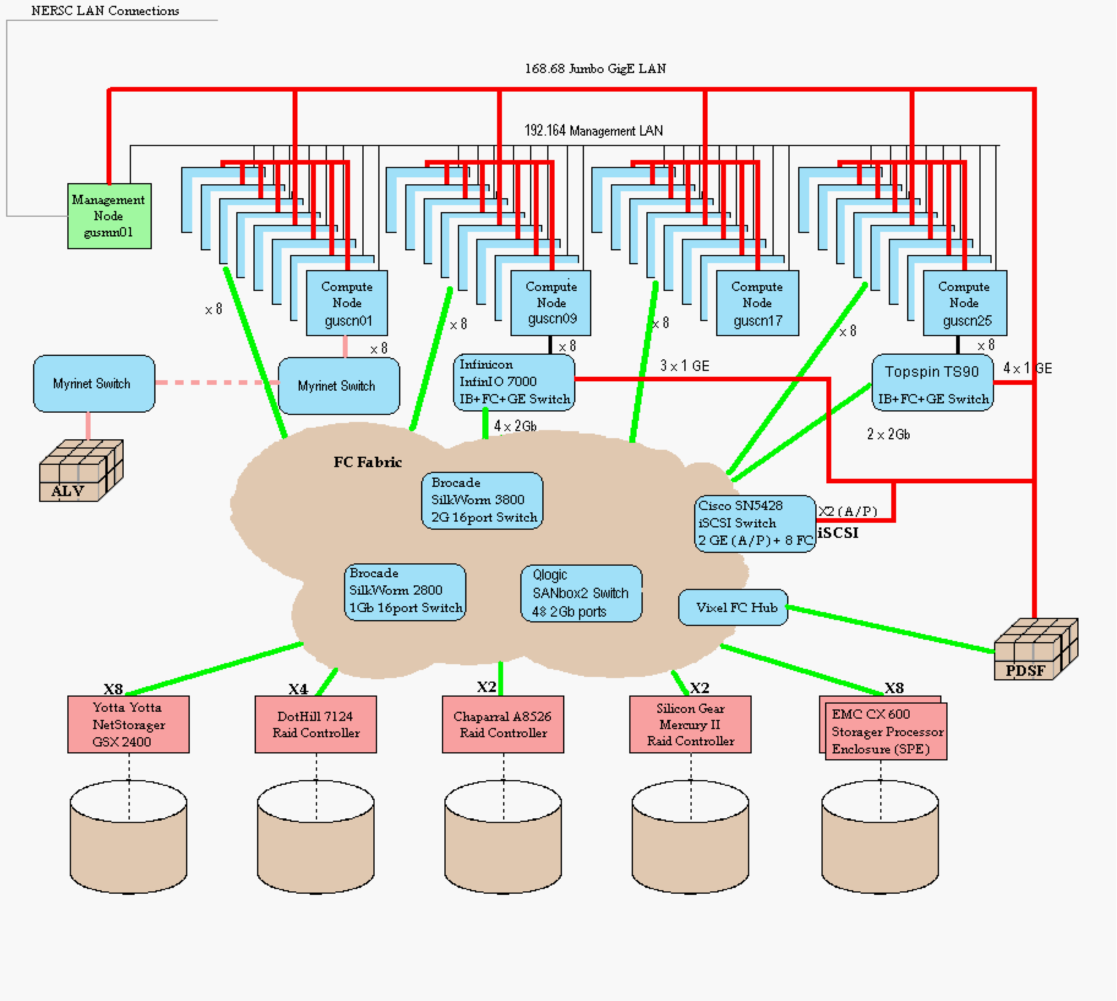

NERSC LAN Connections
168.68 Jumbo GigE LAN
192.164 Management LAN
Management Node gusmn01
x 8
Compute Node guscn01
x 8
x 8
Compute Node guscn09
x 8
x 8
Compute Node guscn17
x 8
Compute Node guscn25
x 8
Myrinet Switch
Myrinet Switch
ALV
Infinicon InfiniIO 7000 IB+FC+GE Switch
4 x 2Gb
3 x 1 GE
Topspin TS90 IB+FC+GE Switch
4 x 1 GE
2 x 2Gb
FC Fabric
Brocade SilkWorm 3800 2G 16port Switch
Cisco SN5428 iSCSI Switch 2 GE (A/P) + 8 FC
X2 (A/P)
iSCSI
Brocade SilkWorm 2800 1Gb 16port Switch
Qlogic SANbox2 Switch 48 2Gb ports
Vixel FC Hub
PDSF
X8
Yotta Yotta NetStorager GSX 2400
X4
DotHill 7124 Raid Controller
X2
Chaparral A8526 Raid Controller
X2
Silicon Gear Mercury II Raid Controller
X8
EMC CX 600 Storager Processor Enclosure (SPE)

Figure B-3. Updated GUPFS testbed configuration for FY 2004.

The major components of the updated GUPFS testbed for FY 2004 included:

System nodes

- 36 dual Pentium-4 nodes: 28 in 2U cases and six in 4U cases; 32 for use as compute nodes and 4 for use as special-purpose nodes
- 6 dual Pentium-3 nodes in 4U cases, used as management, interactive, and auxiliary testbed support nodes

Fabric

- One 240 connector Fiber Optical patch panel with SC connectors and patch cables (see Figure B-4)
- Ethernet
  - One 32-port Extreme 7i Gigabit Ethernet switch
  - One 16-port Extreme 5i Gigabit Ethernet switch
  - Five 24-port Dell Power Connect 5224 Gigabit Ethernet switches
  - Two 10/100 Ethernet switches for system management

- Fibre Channel
  - One 48-port 2 Gb/s Fibre Channel Switch (SanBox2-64)
  - One 16-port 2 Gb/s Fibre Channel Switch (Brocade 3800)
  - One 16-port 1 Gb/s Fibre Channel Switch (Brocade 2800)
  - One Cisco SN5428 iSCSI Router fabric bridge to Ethernet
- Myrinet
  - One Myrinet 2000 8-port switch with 8 Revision D host interface cards
  - One Myrinet 2000 MS-SW16-8E switch card with 8 Gigabit Ethernet ports for bridging between Myrinet and Gigabit Ethernet
- InfiniBand
  - One InfiniCon ISIS InfinIO 7000 4x InfiniBand switch, 8 4x HCA host adapters, and single fabric bridge modules for Fibre Channel and Gigabit Ethernet
  - One Topspin TS90 4x InfiniBand switch, 8 4x HCA host adapters, and single fabric bridge modules for Fibre Channel and Gigabit Ethernet

Storage

- Yotta Yotta NetStorager GSX 2400
- EMC CLARiiON CX600 disk subsystem
- Dot Hill 7124 RAID disk subsystem
- Silicon Gear Mercury II RAID subsystem
- Chaparral A8526 RAID subsystem with attached storage

The expanded testbed, with its increased scale, updated and new technologies, and features to support easier reconfiguration, will facilitate evaluations to be conducted during FY 2004. The increased scale of the testbed will enable both more simultaneous small-scale initial evaluations and single larger scale evaluations. The testbed's multiple fabrics and the bridges between them will allow the issues involving heterogeneous fabric environments, such as those expected at NERSC, to be investigated and understood.

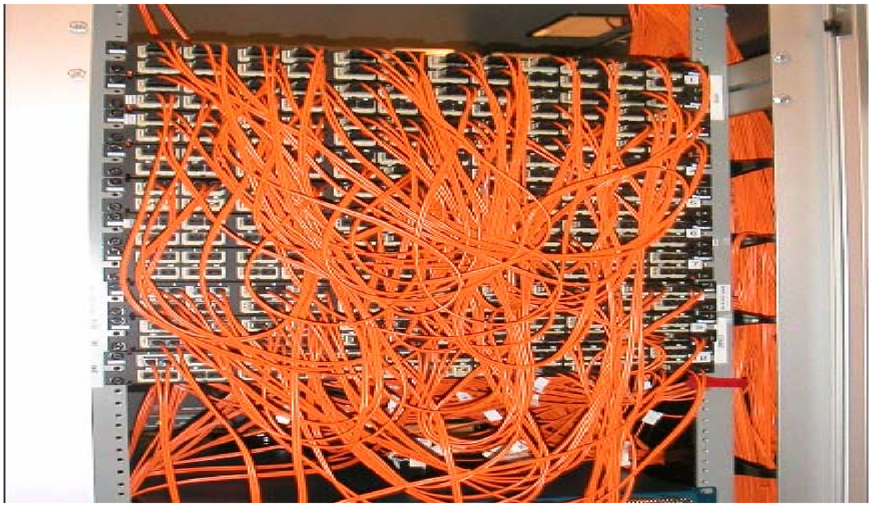


Figure B-4. 240-Port Fibre optical patch panel.

The increased number of testbed nodes will also facilitate the conducting of cross-platform and multiple OS tests of promising file systems supporting such capabilities. A great deal of important information and experience stands to be gained through these tests, which will explore the issues associated with deployment in a heterogeneous environment, as is expected at NERSC. The increased Gigabit Ethernet fabric switching capacity and additional improved fabric bridging capabilities will further facilitate this testing and will enable multiple-cluster testing to be conducted with both the Alvarez cluster and PDSF systems. This will allow phased deployment to be simulated, and we can then begin addressing the networking issues associated with deployment.

Appendix C: Yotta Yotta GSX 2400 Performance

C.1 Overview

Performance benchmarks were run in three configurations to measure Yotta Yotta GSX 2400 performance, using NERSC *MPTIO* benchmark in raw mode:

- A point-to-point (single channel) test
- A single blade (two hosts connecting to the same blade) test
- A single system performance test

C.2 Test Results

C.2.1 Point-to-Point Performance

The *point-to-point* (single channel) test was run between a Linux host and a NetStorager port via the Qlogic SANbox2-64 Fibre Channel switch.

For in-cache tests, the Linux host read/wrote a 512 MB file, and for out-of-cache tests, the host read/wrote an 8 GB file. The I/O size used in the benchmark was 1 MB. The test configuration details are provided in Appendix C.3.

Point-to-point test	Single Shared LUN
In-cache reads	190 MB/sec
In-cache writes	188 MB/sec
Out-of-cache reads	176 MB/sec
Out-of-cache writes	188 MB/sec

C.2.2 Single-Blade Performance

The single-blade test was run between two Linux hosts and two NetStorager ports on the same blade via the Qlogic SANbox2-64 Fibre Channel switch.

For in-cache tests, each host read/wrote a 512 MB file, and for out-of-cache tests, each host read/wrote a 4 GB file. The I/O size used in the benchmark was 1 MB. The test configuration details are provided in Section A.3.

Single blade test	Singe Shared LUN	Two LUNs, Two RAIDs
In-cache reads	372 MB/s	376 MB/s
In-cache writes	375 MB/s	376 MB/s
Out-of-cache reads	300 MB/s	241 MB/s
Out-of-cache writes	277 MB/s	254 MB/s

C.2.3 System Performance

The system performance test was run between eight Linux hosts and eight NetStorager ports over four blades (two of the four blades were on loan from Yotta Yotta via the Qlogic SANbox2-64 Fibre Channel switch.

For in-cache tests, each host read/wrote a 512 MB file while for out-of-cache tests, each host read/wrote a 4 GB file. The I/O size used in the benchmark was 1 MB. The test configuration details are provided in Section A.3.

Many factors influence system performance. The expected performance numbers stated above are relative to the corresponding baseline test configuration specified in the Appendix. Should NERSC experience significant deviation from the stated numbers for similar or alternative configurations, Yotta Yotta will assist with performance tuning and results interpretation.

System performance	Single Shared LUN	8 LUNs, 8 RAIDs
In-cache reads	1513 MB/sec	1514 MB/s
In-cache writes	1425 MB/sec	1497 MB/s
Out-of-cache reads	711 MB/sec	840 MB/s
Out-of-cache writes	714 MB/sec	927 MB/s

C.2.4 16-Node Scalability Test

The 16-node scalability test was run with sixteen Linux hosts connected to eight NetStorager ports, over four blades the Qlogic SANbox2-64 Fibre Channel switch. The NERSC *pioraw* benchmark was used for the test, with different I/O packet sizes (1 KB, 4 KB, 16 KB, 64 KB, 256 KB, 1MB, 4 MB, and 16 MB). The test spawned four I/O processes on each host and ran for 5 minutes for each I/O size. NetStorager was able to handle 64 simultaneous connections without dropping any I/O request. A possible symptom of dropping I/O requests is when any I/O process on any of the host has an outstanding I/O request that cannot complete within 1 minute.

	I/O Size	PIORAW Results (64 PEs)	
		Read	Write
In-cache	16 MB	1579 MB/s	1626 MB/s
	4 MB	1567 MB/s	1605 MB/s
	1 MB	1581 MB/s	1642 MB/s
	256 KB	1566 MB/s	1603 MB/s
	64 KB	1542 MB/s	1636 MB/s
	16 KB	936 MB/s	855 MB/s
	4 KB	270 MB/s	244 MB/s
	1 KB	69 MB/s	62 MB/s
Out-of-cache	16 MB	604 MB/s	288 MB/s
	4 MB	593 MB/s	427 MB/s
	1 MB	575 MB/s	409 MB/s
	256 KB	587 MB/s	393 MB/s
	64 KB	591 MB/s	726 MB/s
	16 KB	406 MB/s	113 MB/s
	4 KB	219 MB/s	94 MB/s
	1 KB	66 MB/s	56 MB/s

C.3 Test Configurations

C.3.1 Test Configuration for Single Shared LUN Tests

Disk details:

- Two enclosures (each enclosure is split into two sets of eight disks)
- Even number of disk from each loop in the RAID

RAID details:

- One single 16-disk wide RAID 0
- 32 K block size
- 8 K stripe depth
- 1 900 GB partition

Access Group details:

- Eight access groups
- One port per access group (AG all have different ports)
- One WWN per AG
- One LUN per WWN
- Partition is exported to all LUNs/hosts

Host details:

- Intel® Xeon Pentium® 4 CPU 2.2 GHz with 2 GB of RAM
- Linux 2.4.18 SMP kernel
- QLogic 2300 series HBA
- QLogic 5.38 driver with Bounce Buffer patch and Transfer Size patch

Front-end connectivity:

- Direct attach, or
- Fibre Channel switch — SANbox2-16 or SANbox2-64

Examples of MPTIO parameters:

- Point-to-point test:
  - o
    ```
    mpirun -np 1 -machinefile MPTIO_yy/bin/hosts-1th
    MPTIO_yy/bin/MPTIO -F 8192M --recsz 1024K --num-thread 8
    --raw /dev/raw/raw1 --test=0 --test=3
    ```
- Single blade test:
  - o
    ```
    mpirun -np 2 -machinefile MPTIO_yy/bin/hosts-1th
    MPTIO_yy/bin/MPTIO -F 16384M --recsz 1024K --num-thread
    8  --raw /dev/raw/raw1 --test=0 --test=3
    ```

- System performance test:

```
o mpirun -np 8 -machinefile MPTIO_yy/bin/hosts-1th
  MPTIO_yy/bin/MPTIO -F 65536M --recsz 1024K --num-thread
  8  --raw /dev/raw/raw1 --test=0 --test=3
```

C.3.2 Test Configuration for Multiple LUN, Multiple RAID Tests

Disk details:

- Two enclosures (each enclosure is split into two sets of eight disks)
- Even number of disks from each loop in the RAID

RAID details:

- Multiple single disk RAID 0
- 16 K block size
- One partition (entire disk)

Access Group details:

- Eight access groups
- One port per access group (AG all have different ports)
- One WWN per AG
- One LUN per WWN
- Each RAID is exported to one Host/LUN

Host details:

- Intel® Xeon Pentium® 4 CPU 2.2 GHz with 2 GB of RAM
- Linux 2.4.18 SMP kernel
- QLogic 2300 series HBA
- QLogic 5.38 driver with Bounce Buffer patch and Transfer Size patch

Front-end connectivity:

- Direct attach, or
- Fibre Channel switch — SANbox2-16 or SANbox2-64

Examples of MPTIO parameters:

- Point-to-point test:

```
o mpirun -np 1 -machinefile MPTIO_yy/bin/hosts-1th
  MPTIO_yy/bin/MPTIO -F 8192M --recsz 512K --num-thread 4
  --raw /dev/raw/raw1 --raw /dev/raw/raw2 --raw
```

```
/dev/raw/raw3 --raw /dev/raw/raw4 --test=0 --test=3  --
overlap t
```

- Single blade test:

    -
      ```
      mpirun -np 2 -machinefile MPTIO_yy/bin/hosts-1th
      MPTIO_yy/bin/MPTIO -F 16384M --recsz 512K --num-thread 4
      --raw /dev/raw/raw1 --raw /dev/raw/raw2 --raw
      /dev/raw/raw3 --raw /dev/raw/raw4 --test=0 --test=3  --
      overlap t
      ```
- System performance test:

    -
      ```
      mpirun -np 8 -machinefile MPTIO_yy/bin/hosts-1th
      MPTIO_yy/bin/MPTIO -F 65536M --recsz 512K --num-thread 4
      --raw /dev/raw/raw1 --raw /dev/raw/raw2 --raw
      /dev/raw/raw3 --raw /dev/raw/raw4 --test=0 --test=3  --
      overlap t
      ```

Appendix D: 3PARdata S400 Storage I/O Performance Results

D.1 Overview

Appendix D documents the timing results of the performance tests done on the 3PAR InServ S400 Storage Server. The primary focus of the performance test was to measure the maximal single-port, single-node, and single-system performance (bandwidth in MB/second). The tests included single-stream and multi-stream tests with different file sizes and I/O sizes. The evaluation period was from August 2003 to October 2003.

D.2 Test Configuration

We used NERSC MPTIO and PIORAW benchmarks to test read and write small (in-cache) and large (out-of-cache) files. Except for the file system tests, all the tests were performed using the raw device interface (e.g., /dev/raw/raw1) to avoid the use of host system buffer cache.

The block sizes used in the tests include: 1 KB, 4 KB, 16 KB, 64 KB, 256 KB, 1 MB, 4 MB, and 16 MB.

For multi-stream parallel I/O tests, the aggregate bandwidth was calculated as the total number of bytes read or written divided by the longest elapsed time.

All tests were run on a quiet system. There were no other activities, neither on the clients nor on the 3PARdata storage when the tests were running.

D.2.1 Linux Host Configuration

The test host had the following configuration:

- Dual 2.2 GHz Xeon P4 processors, SuperMicro motherboard
- 2 GB 133 MHz ECC memory
- RedHat 7.3 with 2.4.18-10smp kernel
- Qlogic QLA2340 2 Gb HBA with qlogic driver v6.1b2 (with a patch that fixed the bounce buffer problem)

D.2.2 3PAR Storage Configuration

The 3PARdata InServ S400 Storage Server consisted of:

- Four controller nodes, each with 4 GB memory
- Each controller node had six PCI slots for host connectivity and back drive chassis
- Only four PCI slots were populated, two for front-end host connections and two for backend drive chassis connections
- Each slot sits dual 2 Gb FC ports

D.3 Performance Results

D.3.1 Single FC Port Performance

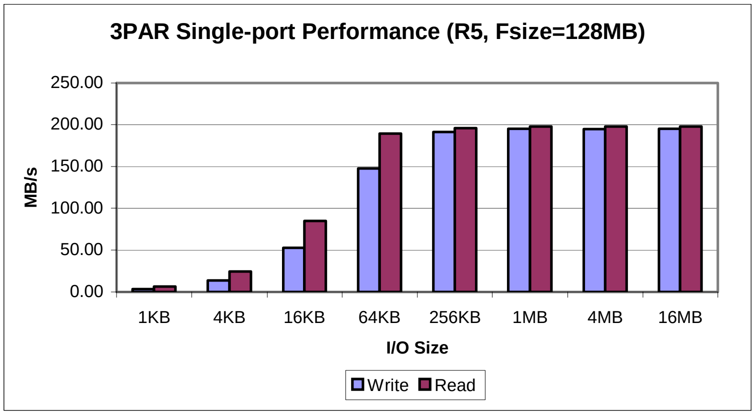
Objective: To measure S400 single FC port performance.

Test Setup:

- Client: Linux 2.4.18-10smp, Dual 2.2 GHz Xeon P4 with QLA2340 HBA
- Storage: 3PARdata
- # of FC Ports and Speed: one 2 Gb FC port
- LUN Configuration: One RAID-5 or RAID-1 LUN
- I/O Size: 1 KB to 16 MB
- # of Client: 1
- # of I/O Processes per Client: 4
- Benchmark: PIORAW

RAID-5 Test Result:

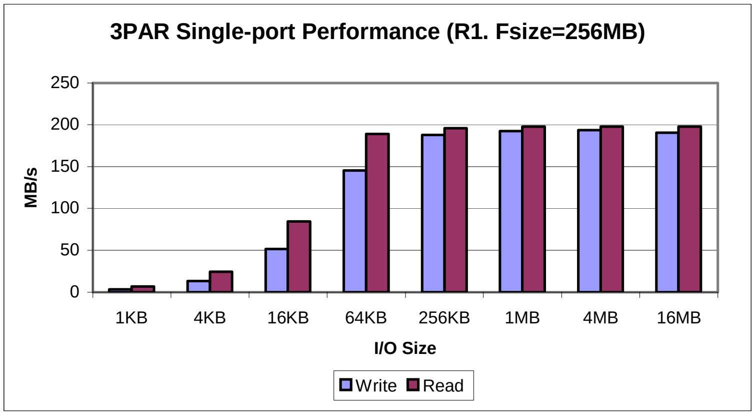
- LUN Configuration: RAID-5 LUN
- File Size: 128 MB



Detail Results: 1-port performance (R5)		
I/O Size	Write	Read
1 KB	3.28 MB/s	6.65 MB/s
4 KB	13.82 MB/s	24.49 MB/s
16 KB	52.86 MB/s	84.94 MB/s
64 KB	147.47 MB/s	189.14 MB/s
256 KB	191.00 MB/s	195.88 MB/s
1 MB	194.87 MB/s	197.67 MB/s
4 MB	194.66 MB/s	197.61 MB/s
16 MB	195.01 MB/s	197.61 MB/s

RAID-1 Test Result:

- LUN Configuration: RAID-1 LUN
- File Size: 256 MB



Detail Results: 1-port performance (R1)		
I/O Size	Write	Read
1 KB	3.31 MB/s	6.77 MB/s
4 KB	13.50 MB/s	24.43 MB/s
16 KB	51.73 MB/s	84.44 MB/s
64 KB	145.25 MB/s	188.96 MB/s
256 KB	187.57 MB/s	195.88 MB/s
1 MB	192.18 MB/s	197.63 MB/s
4 MB	193.25 MB/s	197.62 MB/s
16 MB	190.51 MB/s	197.64 MB/s

Result Summary:

The above results show 3PARdata’s S400 single FC port performance for cached I/O’s on both a RAID-1 LUN and a RAID-5 LUN. The results were the sequential I/O performance, for reading

and writing of a 128 MB file, using different I/O sizes. Since the memory cache on each of the InServ nodes is 8 GB, the content of the 128 MB region can be entirely cached in the InServ controller node. With the file content cached, the performance would show the best possible transfer rate between a host and a FC port on the storage controller.
The best single-port performance was about 195 MB/s for writes and 197 MB/s for reads. Since the entire file could be cached in the controller, we expected to see similar performance results on RAID-1 and RAID-5.

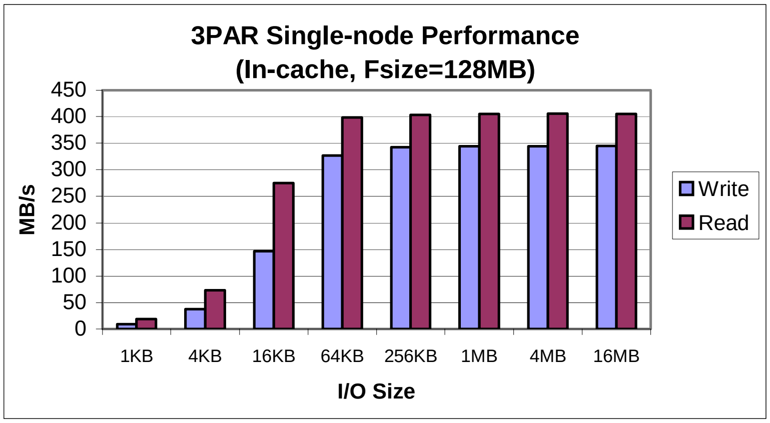
D.3.2 Single-Controller Performance

Objective: To measure S400 single-controller performance.

Test Setup:

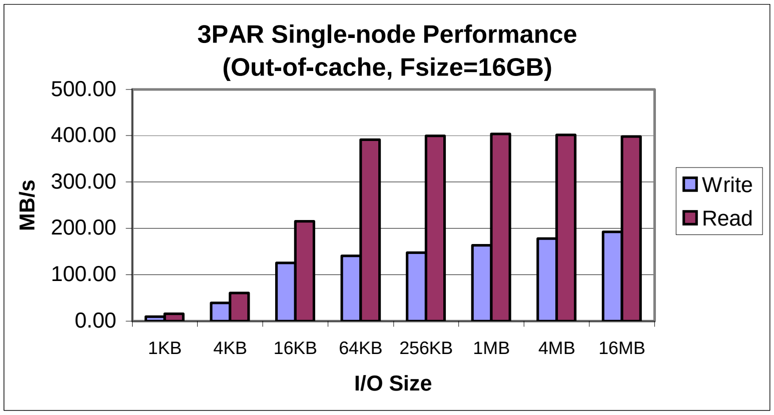
- Clients: Linux 2.4.18-10smp, Dual 2.2 GHz Xeon P4 with QLA2340 HBA
- Storage: 3PARdata
- # of FC Ports and Speed: 4 2 Gb FC ports
- LUN Configuration: 1 RAID-5 LUN
- File Size: 128 MB and 16 GB
- I/O Size: 1 KB to 16 MB
- # of Clients: 4
- # of I/O processes per host: 4
- Benchmark: PIORAW

In-Cache I/O Test Result:



Detail Results: 1-node performance (In-cache, Fsize = 128 MB)		
I/O Size	Write	Read
1KB	9.69 MB/s	19.17 MB/s
4KB	38.11 MB/s	73.52 MB/s
16KB	146.87 MB/s	274.92 MB/s
64KB	326.41 MB/s	398.33 MB/s
256KB	342.08 MB/s	403.07 MB/s
1MB	344.12 MB/s	404.92 MB/s
4MB	344.09 MB/s	405.47 MB/s
16MB	344.94 MB/s	404.95 MB/s

Out-of-Cache I/O Test Result:



Detail Results: 1-node performance (Out-of-cache — 16 GB)		
I/O Size	Write	Read
1KB	9.93 MB/s	15.85 MB/s
4KB	39.50 MB/s	60.81 MB/s
16KB	125.19 MB/s	215.46 MB/s
64KB	140.53 MB/s	390.81 MB/s
256KB	147.55 MB/s	399.43 MB/s
1MB	163.55 MB/s	403.53 MB/s
4MB	177.59 MB/s	401.06 MB/s
16MB	192.73 MB/s	398.12 MB/s

Result Summary:

The above results show 3PAR's single-node performance for reading and writing of 128 MB and 16 GB files using four clients. Each client was connected to one of the four front-end 2 Gb ports on the same 3PAR node (so theoretically a single 3PAR node could sustain a data transfer rate as high as 8 Gb/s).

The in-cache I/O test measures the best possible performance a single controller can achieve. The out-of-cache I/O test measures the sustained I/O performance when the data need to be flushed to or read from the backend disks.

The best single controller performance for in-cache I/O's was about 345 MB/s for writes and 405 MB/s for reads, and for out-of-cache I/O's, it was about 193 MB/s for writes and 404 MB/s for reads. Note that the out-of-cache read performance was very similar to the in-cache read performance. These results seem to indicate that the controller may have a very intelligent read-ahead implementation such that the out-of-cache reads become very similar to in-cache reads.

D.3.3 Performance Scalability of a Single Controller

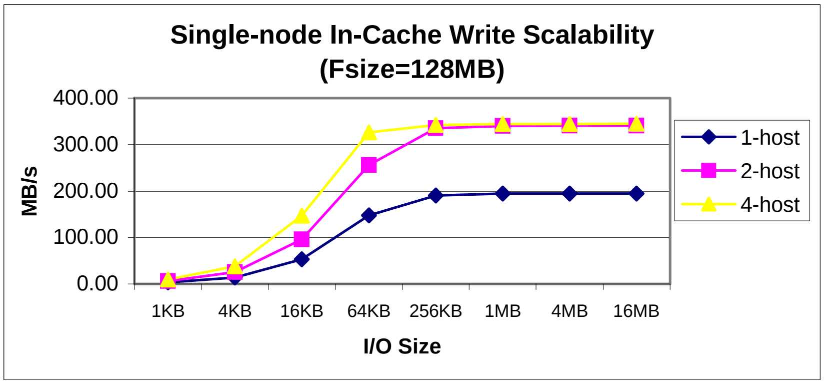
Objective: To test the I/O performance scalability of a single controller.

Test Setup:

- Clients: Linux 2.4.18-10smp, Dual 2.2 GHz Xeon P4 with QLA2340 HBA
- Storage: 3PARdata
- # of FC Ports and Speed: 4 2 Gb FC ports
- LUN Configuration: 1 RAID-5 LUN
- File Size: 128 MB, 16 GB
- I/O Size: 1 KB to 16 MB
- # of Clients: 1, 2, and 4
- # of I/O processes per host: 4
- Benchmark: PIORAW
- Host Connectivity Assignments (based on the Node_Slot_Port position):
  Hosts 1 – 4: 0_1_0, 1_1_0, 2_1_0, 3_1_0

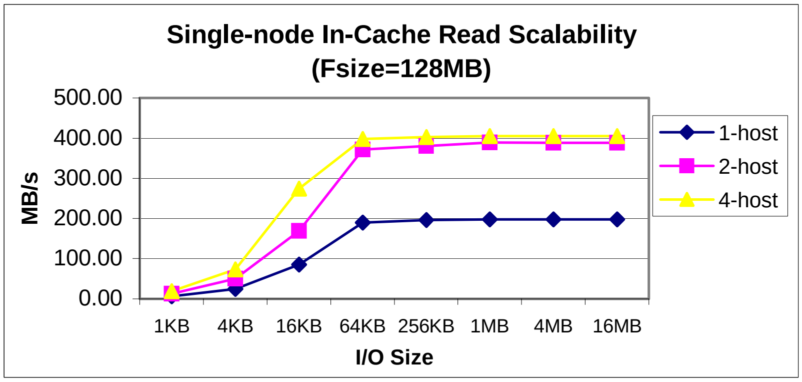
In-Cache I/O Performance Scalability (1-Controller)

In-Cache Write Result:



In-Cache Write Scalability (1-node)			
I/O Size	1-host	2-host	4-host
1KB	3.28 MB/s	6.41 MB/s	9.69 MB/s
4KB	13.82 MB/s	25.70 MB/s	38.11 MB/s
16KB	52.86 MB/s	95.70 MB/s	146.87 MB/s
64KB	147.47 MB/s	256.32 MB/s	326.41 MB/s
256KB	191.00 MB/s	335.16 MB/s	342.08 MB/s
1MB	194.87 MB/s	340.34 MB/s	344.12 MB/s
4MB	194.66 MB/s	340.84 MB/s	344.09 MB/s
16MB	195.01 MB/s	340.95 MB/s	344.94 MB/s

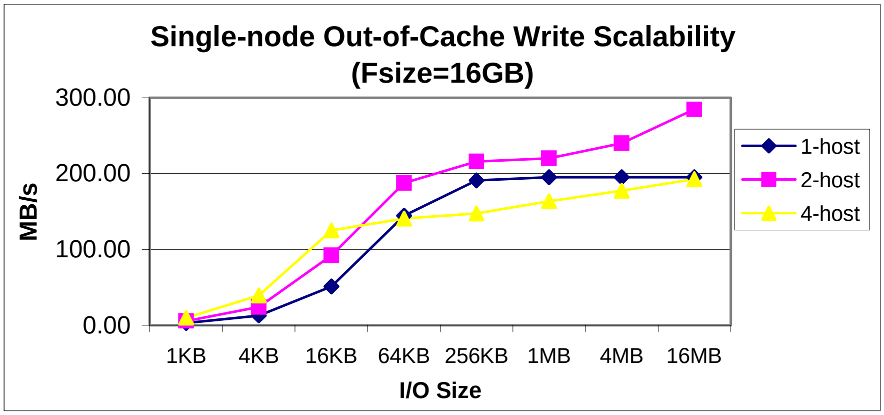
In-Cache Read Result:



In-Cache read scalability (1-node)			
I/O Size	1-host	2-host	4-host
1 KB	6.65 MB/s	12.93 MB/s	19.17 MB/s
4 KB	24.49 MB/s	49.88 MB/s	73.52 MB/s
16 KB	84.94 MB/s	169.03 MB/s	274.92 MB/s
64 KB	189.14 MB/s	372.01 MB/s	398.33 MB/s
256 KB	195.88 MB/s	380.19 MB/s	403.07 MB/s
1 MB	197.67 MB/s	388.95 MB/s	404.92 MB/s
4 MB	197.61 MB/s	388.74 MB/s	405.47 MB/s
16 MB	197.61 MB/s	388.60 MB/s	404.95 MB/s

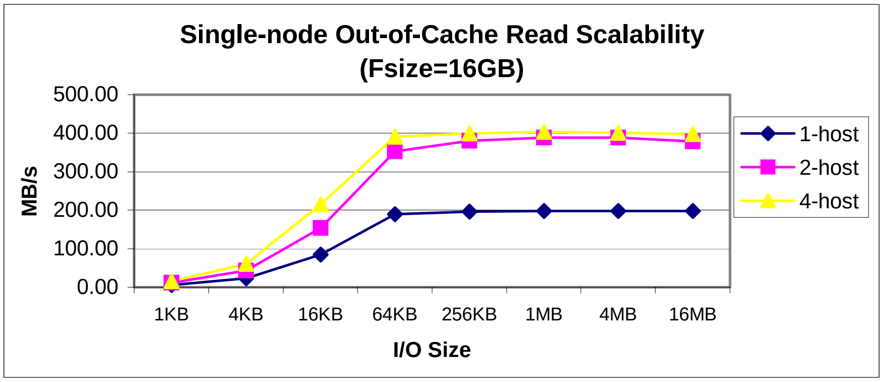
Out-of-Cache Performance Scalability (1-Controller)

Out-of-Cache Write Result:



Out-of-Cache write scalability (1-node)			
I/O Size	1-host	2-host	4-host
1 KB	3.11 MB/s	5.95 MB/s	9.93 MB/s
4 KB	13.07 MB/s	24.39 MB/s	39.50 MB/s
16 KB	51.35 MB/s	92.27 MB/s	125.19 MB/s
64 KB	144.65 MB/s	187.37 MB/s	140.53 MB/s
256 KB	190.95 MB/s	215.62 MB/s	147.55 MB/s
1 MB	195.12 MB/s	219.83 MB/s	163.55 MB/s
4 MB	195.12 MB/s	239.70 MB/s	177.59 MB/s
16 MB	195.20 MB/s	284.38 MB/s	192.73 MB/s

Out-of-Cache Read Result:



Out-of-Cache read scalability (1-node)			
I/O Size	1-host	2-host	4-host
1KB	6.21 MB/s	11.38 MB/s	15.85 MB/s
4KB	23.64 MB/s	42.86 MB/s	60.81 MB/s
16KB	84.58 MB/s	153.94 MB/s	215.46 MB/s
64KB	189.10 MB/s	353.01 MB/s	390.81 MB/s
256KB	195.64 MB/s	380.08 MB/s	399.43 MB/s
1MB	197.64 MB/s	388.82 MB/s	403.53 MB/s
4MB	197.63 MB/s	388.89 MB/s	401.06 MB/s
16MB	197.50 MB/s	378.81 MB/s	398.12 MB/s

Result Summary:

The above results show the I/O scalability of a single node for reading and writing of a 128 MB (in-cache) region and 16 GB (out-of-cache) region of a raw device. Each host was connected to one of the four front-end 2 Gb ports on a single 3PAR controller node.

For in-cache tests with I/O sizes larger than 256 KB, the results indicate that two clients would be able to generate loads to saturate a single controller node. The best I/O performance a single controller node could do was about 345 MB/s for writes and 405 MB/s for reads.

For out-of-cache tests with larger I/O sizes, the results show very different behavior for reads and writes. For reads, the out-of-cache results were very similar to the in-cache results. The read results seem to indicate that the controller may have a very intelligent read-ahead implementation such that the speed of reading data from the backend disks was able to match the read speed from the hosts. As a result, out-of-cache reads became in-cache reads and out-of-cache read performance became very similar to in-cache read performance. For out-of-cache writes, the results were quite different. The best aggregate performance for writes was achieved with two hosts, and the write performance dropped when there were four hosts writing for tests with I/O sizes larger than 256 KB. For tests with I/O sizes that were larger than 256 KB, the aggregate performance with four hosts was actually lower than what a single host was able to achieve. The out-of-cache write forces data to be flushed from the storage controller cache to the backend disks. The performance drop for writes with four hosts might have been caused by a combination of contentions on the controller, parity calculation, and/or the head movement on the backend disks. However, the head movement did not seem to affect the out-of-cache, concurrent-read performance that much.

D.3.4 S400 I/O Scalability (with a Shared LUN)

Objective: To test the I/O performance scalability of a single S400 system.

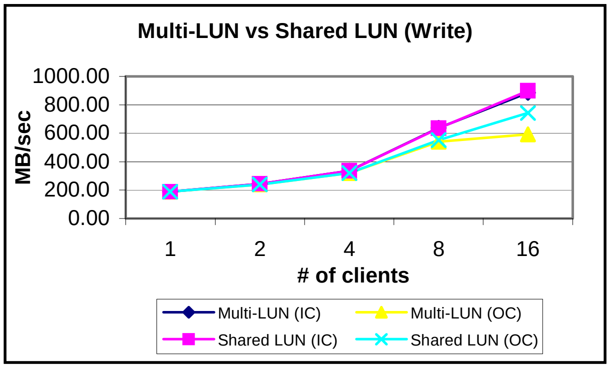
Test Setup:

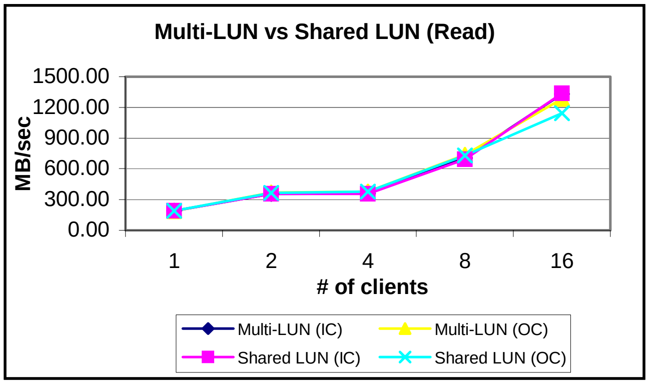
- Clients: Linux 2.4.18-10smp, Dual 2.2 GHz Xeon P4 with QLA2340 HBA
- # of FC Ports and Speed: sixteen 2 Gb FC ports
- LUN Configuration: Shared-LUN vs. Multi-LUN (16 LUNs)
- File Size: 2 GB and 50 GB

- I/O Size: 1 MB
- # of Clients: 1, 2, 4, 8, 16
- # of I/O processes per host: 8
- Benchmark: MPTIO
- Host Connectivity Assignments (based on the Node_Slot_Port position):

  Hosts 1 – 4: 0_1_0, 1_1_0, 2_1_0, 3_1_0
  Hosts 5 – 8: 0_5_0, 1_5_0, 2_5_0, 3_5_0
  Hosts 9 – 12: 0_1_1, 1_1_1, 2_1_1, 3_1_1
  Hosts 14 -16: 0_5_1, 1_5_1, 2_5_1, 3_5_1

Test Results:





Multi-LUN Results:

- LUN Configuration: sixteen 100 GB RAID-5 LUNs, one LUN per FC port

3PARdata S400 I/O Scalability (Multiple LUNs)				
# of Clients	IC Write	OC Write	IC Read	OC Read
1	188.34 MB/s	188.44 MB/s	189.72 MB/s	189.09 MB/s
2	243.48 MB/s	240.01 MB/s	357.85 MB/s	369.33 MB/s
4	336.29 MB/s	319.80 MB/s	361.27 MB/s	378.32 MB/s
8	637.75 MB/s	540.34 MB/s	703.65 MB/s	736.90 MB/s
16	885.60 MB/s	593.00 MB/s	1329.37 MB/s	1284.20 MB/s

Shared-LUN Results:

- LUN Configuration: one 1 TB RAID-5 LUN, exported on all sixteen FC ports

3PARdata S400 I/O scalability (shared LUN)				
# of Clients	IC Write	OC Write	IC Read	OC Read
1	188.36 MB/s	189.75 MB/s	189.75 MB/s	189.66 MB/s
2	243.44 MB/s	240.06 MB/s	357.11 MB/s	367.80 MB/s
4	336.84 MB/s	321.44 MB/s	355.44 MB/s	378.16 MB/s
8	633.48 MB/s	552.38 MB/s	692.19 MB/s	729.68 MB/s
16	900.03 MB/s	742.03 MB/s	1335.77 MB/s	1144.42 MB/s

Result Summary:

The GUPFS project has been evaluating several shared file system technologies. The two most promising shared file system architectures are SAN-based (ADIC StorNext File System and IBM StorageTank) and network-based with directly attached storage (Lustre and NSD-based IBM GPFS). Most of the new storage systems that offer storage consolidation can support both file system implementations. For a SAN-based file system implementation, the underlying storage needs to support shared access to the same LUN from many hosts. For a network-based file system implementation, to deliver the aggregate performance requirement, the storage needs to be able to create many LUNs. The purpose of this test is to compare the performance difference between S400 when running with a Shared-LUN configuration and when running with a Multi-LUN configuration.

The above results show the raw I/O scalability and aggregate performance of a 3PARdata S400 system, with two LUN configurations: multi-LUN and shared-LUN. The performance results were measured for parallel reading and writing of 2 GB and 50 GB regions using 1, 2, 4, 8, and 16 clients. For tests with more than one client, the file region was divided equally among the clients. For example, in the case of 4 clients reading or writing a 2 GB file region, each client was reading or writing a separate 512 MB region.

Each client was connected to one of the sixteen front-end 2 Gb ports on S400, based on a port assignment to minimize slot or node contention to achieve the best possible performance out of a single S400 system. For example, for four-client tests, no two clients were connected to the storage ports on the same controller node, and for eight-client tests, no two clients were connected to the storage ports on the same slot.

The results show that the performance between shared-LUN configuration and multi-LUN configuration was almost identical, except for the sixteen-client tests. On the 3PAR S400 storage, a LUN is created in terms of 256 MB chunks, and the chunks that makes up a LUN will spread across as many spindles as the LUN will take. For example, to create a 10 GB LUN, 40 chunks will be used, and these 40 chunks will spread across all the available spindles to get the best performance.

Note again that the out-of-cache read performance was very similar to the in-cache read performance, probably because of the intelligent read-ahead implementation on the controllers. For writes, the out-of-cache performance was very close to the in-cache performance until more than four clients were involved. For eight-client and sixteen-client tests, the out-of-cache write performance was lower than the in-cache write performance.

For shared-LUN configuration, the S400 performance was about 886 MB/s for in-cache writes, 593 MB/s for out-of-cache writes, and about 1280+ MB/s for in-cache and out-of-cache reads. For multi-LUN configuration, the performance was about 900 MB/s for in-cache writes, 742 MB/s for out-of-cache writes, and about 1144+ MB/s for in-cache and out-of-cache reads.

D.3.5 GPFS on Linux 1.3 Performance Results

Objective: To measure GPFS on Linux performance using a S400 system.

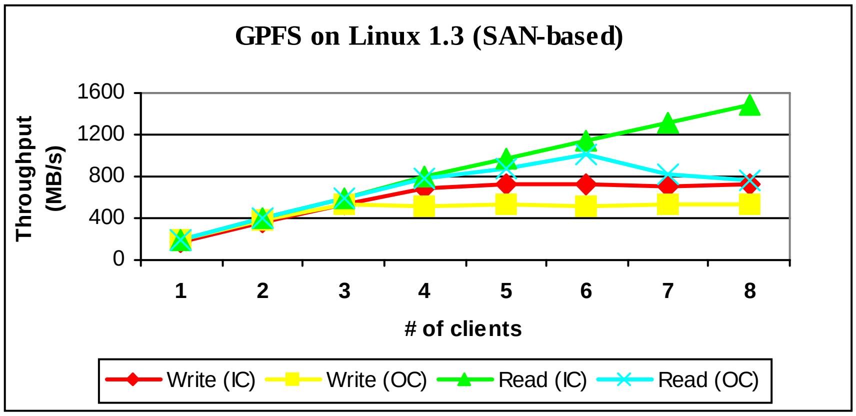
Test Setup:

- GPFS Clients: Linux 2.4.18-10smp, Dual 2.2 GHz Xeon P4, 2 GB Memory
- Interconnect: Myrinet (Rev D) with LAPI
- Client FC HBA: QLA2340 2 Gb/s HBA
- Storage: 3PARdata
- LUN Configuration: 1 1-TB RAID-5 LUN
- File Size: In-cache (IC): 128 MB, out-of-cache (OC): 1073 MB
- I/O Size: 1 MB
- # of Clients: 1 to 8
- # of I/O processes per client: 4
- Benchmark: PIORAW
- Host Connectivity Assignments (based on the Node_Slot_Port position):
  Hosts 1 – 4: 0_1_0, 1_1_0, 2_1_0, 3_1_0
  Hosts 5 – 8: 0_5_0, 1_5_0, 2_5_0, 3_5_0

Test Results:

SAN-based Configuration (with a Single Shared LUN):

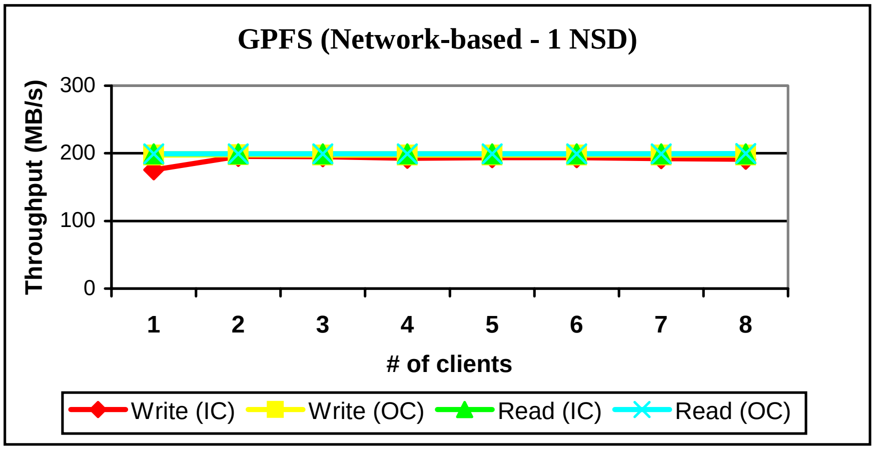
- SAN Configuration: eight 2 Gb FC ports (over a shared LUN)



GPFS on Linux Scalability (SAN-based)				
# of Clients	IC Write	OC Write	IC Read	OC Read
1	179.51 MB/s	196.80 MB/s	198.97 MB/s	199.07 MB/s
2	369.69 MB/s	381.90 MB/s	397.91 MB/s	398.09 MB/s
3	541.96 MB/s	534.86 MB/s	596.67 MB/s	596.15 MB/s
4	679.77 MB/s	509.67 MB/s	794.88 MB/s	781.15 MB/s
5	726.49 MB/s	532.32 MB/s	971.14 MB/s	879.01 MB/s
6	731.05 MB/s	513.67 MB/s	1146.58 MB/s	1005.01 MB/s
7	700.06 MB/s	531.46 MB/s	1318.86 MB/s	819.08 MB/s
8	714.67 MB/s	524.58 MB/s	1484.45 MB/s	757.33 MB/s

NSD-based Configuration (with a single NSD):

- NSD Configuration: one 2 Gb FC port with a direct-attach storage



GPFS on Linux scalability (NSD-based)				
# of Clients	IC Write	OC Write	IC Read	OC Read
1	175.56 MB/s	197.04 MB/s	199.03 MB/s	199.11 MB/s
2	195.14 MB/s	197.42 MB/s	198.94 MB/s	198.98 MB/s
3	194.70 MB/s	197.28 MB/s	199.20 MB/s	199.00 MB/s
4	192.27 MB/s	197.31 MB/s	199.17 MB/s	199.04 MB/s
5	193.44 MB/s	197.17 MB/s	199.24 MB/s	199.18 MB/s
6	193.38 MB/s	197.24 MB/s	199.26 MB/s	199.22 MB/s
7	192.08 MB/s	197.19 MB/s	199.12 MB/s	199.27 MB/s
8	190.90 MB/s	197.17 MB/s	199.27 MB/s	199.32 MB/s

Result Summary:

The above results show the I/O scalability of GPFS on Linux Release 1.3 using a single LUN on the 3PARdata S400 system. Two GPFS file system architectures were tested: a SAN-based implementation and an NSD-based implementation. In the SAN-based implementation, the same LUN was exported on eight FC ports on the four controller nodes of S400. Each client was connected to one of the eight FC ports so that the same LUN appeared on all eight hosts (as /dev/sdb) for shared access. A GPFS file system instance was created on the shared partition, and the I/O tests were run on the eight GPFS clients. In the NSD-based implementation, a LUN should only be exported to a single host (a single LUN can be exported to more than one host to support high availability, but only one host can be active and others are stand-by only). Since we only had a single LUN, a single NSD node with eight GPFS clients was created for the I/O tests.

The performance was measured for parallel reading and writing of 128 MB and 1087 MB files, scaling from one to eight clients. The results show that, in the NSD-based implementation, the single NSD was the bottleneck and the I/O performance was limited to only about 200 MB/s while the SAN-based implementation was able to achieve better performance when the number of clients increased.

The results seem to indicate that the SAN-based implementation scales better. However, since the number of clients used was only eight, which was not a very large cluster, we may not have experienced all the scalability issues of a SAN-based implementation. If the number of clients were increased to hundreds or thousands, we might have encountered different scalability issues. More scalability tests are needed on larger clusters (for example, on NERSC’s Alvarez cluster or PDSF cluster) to really compare between a SAN-based implementation and an NSD-based implementation.

D.3.6 Single-LUN Stress Test

Objective: To stress test a single LUN with up to 128 connections.

Test Setup:

- Clients: Linux 2.4.18-10smp, Dual 2.2 GHz Xeon P4, 2 GB Memory
- Client FC HBA: QLA2340 2 Gb/s HBA
- LUN Configuration: 1 RAID-5 LUN
- File Size: In-cache (IC): 2100 MB
- I/O Size: 1 MB
- # of Clients: 1 to 8
- # of I/O processes per client: 1, 2, 4, and 8
- Benchmark: MPTIO

Test Results:

# Procs per Client	# of Clients	Total # of Procs	In-Cache		Out-of-Cache	
			Write	*Read*	*Write*	*Read*
1	1	1	93.23	110.19	91.76	103.46
	2	2	160.01	186.66	155.34	181.66
	4	4	195.35	189.50	193.98	189.60
	8	8	196.33	189.77	196.13	189.39
	16	16	191.57	189.75	190.57	189.54
2	1	2	180.67	184.96	170.94	181.09
	2	4	195.05	188.90	193.12	189.59
	4	8	195.99	189.77	196.05	189.58
	8	16	195.31	189.75	195.71	189.50
	16	32	190.47	189.71	189.30	189.82
4	1	4	187.50	189.61	186.43	189.41
	2	8	196.31	189.77	196.09	189.61
	4	16	196.32	189.75	196.14	189.56
	8	32	195.76	189.68	194.13	189.81
	16	64	190.41	186.63	189.37	186.73
8	1	8	188.35	189.78	188.37	189.74
	2	16	196.39	189.77	196.16	184.98
	4	32	196.32	189.79	196.15	189.82
	8	64	196.34	189.75	196.10	189.86
	16	128	190.43	186.60	189.38	186.69

Result Summary:

Some storage systems cannot support too many simultaneous connections with active I/O requests on each connection. The test shows that S400 was able to handle 128 connections doing active I/O's at about 190 MB/s.

D.3.7 Performance Comparison between RAID-1 and RAID-5 on S400

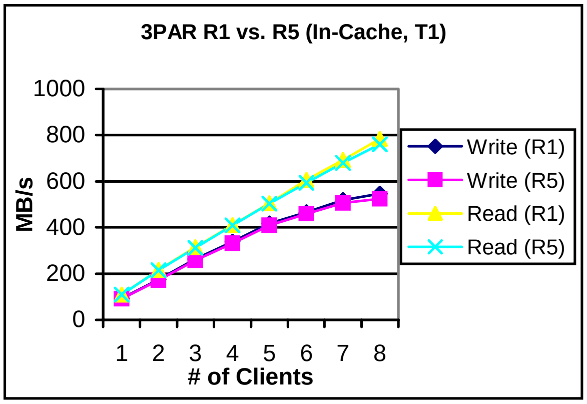
Objective: To compare performance differences between RAID-1 and RAID-5 on S400.

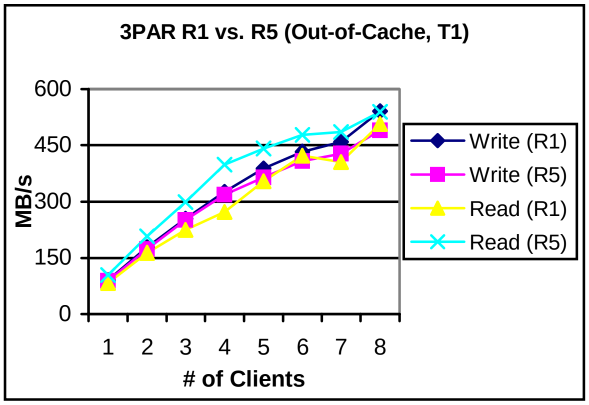
Test Setup:

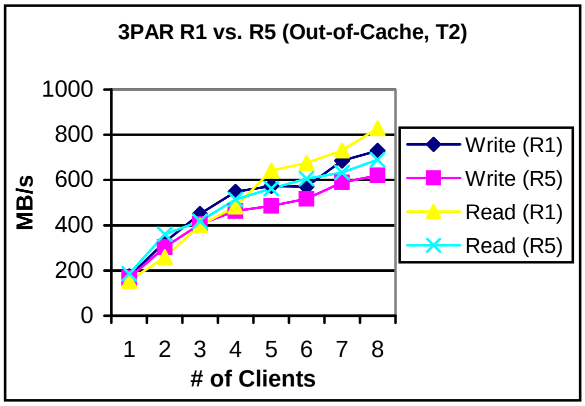
- Clients: Linux 2.4.18-10smp, Dual 2.2 GHz Xeon P4, 2 GB Memory
- Client FC HBA: QLA2340 2 Gb/s HBA
- Storage: 3PARdata S400
- LUN Configuration: 1 RAID-1 LUN and 1 RAID-5 LUN using the same spindles
- File Size: In-cache (IC): 2100 MB, Out-of-cache (OC): 50400 MB
- I/O Size: 1 MB
- # of Clients: 1 to 8
- # of I/O processes per client: 1, 2, and 4
- Benchmark: MPTIO

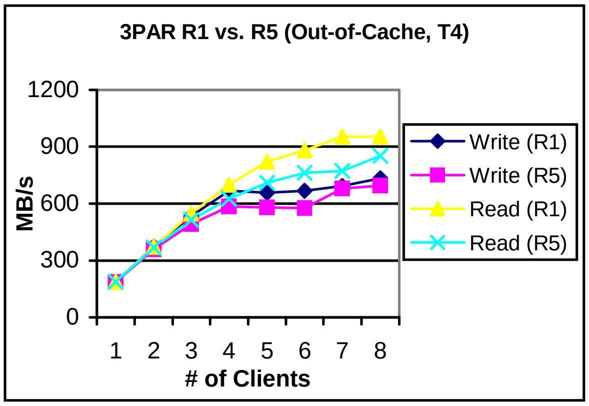
- Host Connectivity Assignments (based on the Node_Slot_Port position):
  Hosts 1 – 4: 0_1_0, 1_1_0, 2_1_0, 3_1_0
  Hosts 5 – 8: 0_5_0, 1_5_0, 2_5_0, 3_5_0

Test Results:









Result Summary:

The above results show the in-cache and out-of-cache I/O performance differences between a RAID-1 LUN and a RAID-5 LUN, using the MPTIO benchmark.

For in-cache tests with 1-thread (In-Cache, T1), the RAID-1 performance was almost identical to the RAID-5 performance, for both reads and writes. This is expected since with in-cache tests the entire file can fit in the controller cache. For out-of-cache (Out-of-Cache, T1) tests, RAID-5 was faster than RAID-1 for reads but was slightly slower for writes. However, for multiple-thread, out-of-cache tests, RAID-1 was generally faster than RAID-5 for both reads and writes on S400. For sequential writes, RAID-5 performance generally was as good as RAID-1 or even RAID-0, when the underlying storage controller could do *full-stripe writes* without the overhead of parity calculation. On S400, we observed a performance difference as high as 18% between RAID-1 writes and RAID-5 writes. The result that RAID-1 performance was faster than RAID-5 for reads was a surprise In general, when a RAID-1 LUN or a RAID-5 LUN is created with the same amount of spindles, the RAID-5 LUN generally is faster on reads and slower on writes than RAID-1. It is not clear why RAID-1 was faster than RAID-5 on reads (with multiple clients and multiple threads per client) on the 3PAR S400 storage. It may be because S400 has a better disk access scheduling algorithm so that reading from a RAID-1 volume can spread across more spindles than RAID-5 when there are enough number of threads doing I/O's at the same time.

Detailed Test Results:

In the following tables, the title on each table describes the test configuration used for the tests. The interpretation for a title R_CC_TT_X is:

- R — Raw device I/O test
- CC — IC (in-cache) or OC (out-of-cache)
- TT — # of I/O thread per client
- X — R for rotate read and Z for sleeping before reads start after writes complete; when X is missing, reads will follow writes immediately

For example, R_IC_T1 means the test was an in-cache (IC) I/O test with one test thread per client. Please see the GUPFS FY 2002 Technical Report for a more detailed description of the MPTIO benchmark.

R_IC_T1					
Fsize (MB/proc)	**# of Clients**	**Seq Write**		**Seq Read**	
		R1	**R5**	**R1**	**R5**
2100	1	92.11	90.91	109.66	109.49
1050	2	174.53	171.64	215.71	214.40
700	3	263.30	257.86	313.76	311.76
525	4	338.03	331.55	410.22	407.93
420	5	416.28	408.51	504.87	503.64
350	6	466.76	459.50	606.75	594.33
300	7	519.15	507.12	691.92	679.85
262	8	547.00	524.46	784.11	759.79

R_IC_T1_R					
Fsize (MB/proc)	**# of Clients**	**Seq Write**		**Seq Read**	
		R1	**R5**	**R1**	**R5**
2100	1	92.27	90.46	109.61	106.83
1050	2	173.43	171.11	212.39	209.77
700	3	263.00	257.63	317.45	314.06
525	4	336.83	327.70	413.72	416.26
420	5	411.85	411.48	498.05	509.70
350	6	462.28	448.58	597.81	599.88
300	7	517.81	505.76	693.59	687.45
262	8	543.20	524.04	761.90	731.81

R_IC_T1_Z

Fsize (MB/proc)	# of Clients	Seq Write		Seq Read	
		R1	R5	R1	R5
2100	1	92.06	90.97	109.85	107.95
1050	2	174.56	171.64	218.93	214.28
700	3	262.00	257.23	318.62	319.84
525	4	338.14	329.00	423.12	423.23
420	5	412.14	407.44	509.58	521.16
350	6	468.83	454.97	622.79	611.53
300	7	515.02	504.98	726.08	710.07
262	8	549.29	521.45	826.48	813.29

R_IC_T2

Fsize (MB/proc)	# of Clients	Seq Write		Seq Read	
		R1	R5	R1	R5
2100	1	177.69	175.30	188.71	186.11
1050	2	338.65	327.72	365.88	366.97
700	3	471.33	449.32	549.11	548.04
524	4	594.95	567.80	714.05	704.35
420	5	673.69	651.66	828.61	825.44
350	6	694.26	673.82	956.18	655.84
300	7	731.73	706.86	1061.80	1042.01
262	8	780.74	759.01	1233.64	1204.70

R_IC_T2_R

Fsize (MB/proc)	# of Clients	Seq Write		Seq Read	
		R1	R5	R1	R5
2100	1	177.29	166.77	188.55	186.65
1050	2	339.23	327.34	368.15	361.67
700	3	471.06	451.88	538.41	538.81
524	4	597.21	571.24	702.15	720.57
420	5	669.14	636.41	853.65	827.79
350	6	700.72	657.61	972.75	962.81
300	7	717.84	695.83	1091.50	1054.91
262	8	774.08	762.06	1096.53	1041.27

R_IC_T2_Z					
Fsize (MB/proc)	# of Clients	Seq Write		Seq Read	
		R1	R5	R1	R5
2100	1	177.14	175.27	189.01	189.65
1050	2	338.82	331.28	375.93	378.73
700	3	466.48	452.41	564.79	568.27
524	4	600.64	567.60	753.18	756.19
420	5	667.31	629.93	916.77	916.65
350	6	695.54	662.99	1063.79	1067.04
300	7	726.30	705.61	1183.38	1188.20
262	8	774.13	755.74	1435.51	1442.70
R_OC_T1					
Fsize (MB/proc)	# of Clients	Seq Write		Seq Read	
		R1	R5	R1	R5
50400	1	90.39	89.06	82.70	104.07
25200	2	178.16	172.68	163.12	207.02
16800	3	253.89	250.52	224.30	299.11
12600	4	326.06	318.68	272.01	398.93
10080	5	388.59	364.63	354.17	441.50
8400	6	433.49	408.34	422.83	477.25
7200	7	458.73	427.60	405.79	485.68
6300	8	541.18	489.95	506.94	539.15

R_OC_T1_R					
Fsize (MB/proc)	# of Clients	Seq Write		Seq Read	
		R1	R5	R1	R5
50400	1	89.96	88.67	82.61	104.07
25200	2	175.34	171.53	165.62	205.99
16800	3	255.57	249.24	218.80	295.62
12600	4	328.54	319.39	266.79	396.40
10080	5	383.59	356.73	349.24	445.04
8400	6	426.95	405.28	422.56	449.43
7200	7	454.15	425.45	415.97	475.92
6300	8	536.01	498.28	505.82	547.93
R_OC_T1_Z					
Fsize (MB/proc)	# of Clients	Seq Write		Seq Read	
		R1	R5	R1	R5
50400	1	90.09	89.39	82.65	103.24
25200	2	175.28	171.89	165.40	208.12
16800	3	253.01	247.25	219.20	306.04
12600	4	324.49	318.03	280.27	404.28
10080	5	381.54	366.73	380.84	445.95
8400	6	425.69	405.09	415.53	453.33
7200	7	455.88	426.14	400.10	485.64
6300	8	512.62	493.68	513.14	573.93

R_OC_T2					
Fsize (MB/proc)	**# of Clients**	**Seq Write**		**Seq Read**	
		R1	**R5**	**R1**	**R5**
50400	1	175.45	168.62	155.89	186.34
25200	2	325.70	303.43	259.33	360.07
16800	3	450.97	404.86	400.39	417.59
12600	4	548.72	462.83	483.36	513.70
10080	5	574.08	485.58	641.50	562.43
8400	6	570.22	517.53	675.84	605.41
7200	7	685.45	588.87	730.11	633.41
6300	8	729.46	620.51	827.18	689.04
R_OC_T2_R					
Fsize (MB/proc)	**# of Clients**	**Seq Write**		**Seq Read**	
		R1	**R5**	**R1**	**R5**
50400	1	175.29	168.13	156.00	186.14
25200	2	321.83	303.98	259.42	360.06
16800	3	434.70	397.78	402.30	436.31
12600	4	549.65	467.44	491.65	514.68
10080	5	562.79	485.32	648.15	594.89
8400	6	566.91	522.89	622.02	620.35
7200	7	677.90	588.83	714.29	640.74
6300	8	719.13	620.48	835.64	688.08

R_OC_T2_Z					
Fsize (MB/proc)	**# of Clients**	**Seq Write**		**Seq Read**	
		R1	**R5**	**R1**	**R5**
50400	1	175.25	168.75	156.90	185.36
25200	2	325.55	305.03	264.57	363.09
16800	3	436.70	401.06	399.10	443.48
12600	4	503.13	461.30	516.99	530.35
10080	5	562.36	483.99	659.31	594.83
8400	6	573.29	521.59	706.11	631.34
7200	7	684.36	588.74	756.49	675.42
6300	8	727.82	621.33	885.95	705.64
R_OC_T4					
Fsize (MB/proc)	# of Clients	Seq Write		Seq Read	
		R1	R5	R1	R5
50400	1	188.25	186.60	188.25	188.04
25200	2	373.39	358.92	372.11	368.08
16800	3	534.69	491.27	547.70	512.76
12600	4	667.41	585.03	699.36	625.62
10080	5	657.57	580.01	821.62	710.03
8400	6	668.07	576.60	881.16	761.92
7200	7	693.88	679.24	953.18	773.27
6300	8	733.30	694.98	952.63	851.83

R_OC_T4_R					
Fsize (MB/proc)	# of Clients	Seq Write		Seq Read	
		R1	R5	R1	R5
50400	1	188.17	186.65	187.97	188.24
25200	2	372.43	358.11	367.74	365.54
16800	3	535.43	483.11	547.40	512.57
12600	4	662.41	586.44	706.29	620.49
10080	5	654.45	581.38	820.81	708.42
8400	6	657.26	575.15	896.00	756.30
7200	7	701.51	680.70	952.21	814.56
6300	8	730.07	691.48	954.14	818.15

R_OC_T4_Z					
Fsize (MB/proc)	# of Clients	Seq Write		Seq Read	
		R1	R5	R1	R5
50400	1	188.20	186.53	188.25	188.26
25200	2	373.25	357.63	371.85	368.99
16800	3	537.14	494.13	560.79	522.59
12600	4	669.41	582.44	726.90	631.78
10080	5	657.21	580.58	842.54	718.65
8400	6	639.19	571.42	904.25	750.71
7200	7	695.27	680.54	952.31	850.24
6300	8	727.32	693.78	1004.07	881.98

Appendix E: Topspin 90 InfiniBand Switch Performance

E.1 Overview

The Global Unified Parallel File System (GUPFS) project is a multiple-phase, five-year NERSC Center project whose focus is to provide a scalable, high-performance, high-bandwidth, shared file system for all the NERSC production systems. The GUPFS project is intended to evaluate emerging storage, fabric, and file system technologies to determine the best solutions for a center-wide shared file system. Appendix E reports the performance evaluation results of the Topspin TS90 InfiniBand switch and the 4x HCA (host channel adapter) as a storage fabric technology candidate for GUPFS.

The emerging InfiniBand (IB) interconnect shows promise as a transport for storage traffic in a storage area network (SAN). IB offers performance (bandwidth and latency) beyond that of either Ethernet or Fibre Channel (FC). NERSC started the evaluation of 1x IB technology in 2002. We have seen significant progress made in performance improvement. Below are our findings and evaluations of Topspin TS90 IB switch and 4x HCA technologies and interoperability between different fabric technologies (namely, IB, FC, and iSCSI).

E.2 Topspin InfiniBand Switch

The Topspin 90, with 12 InfiniBand ports, creates a single 10 Gb/s fabric for inter-process communications, storage, and networking. The Topspin 90 switch can be expanded from the 12-port base configuration by inserting an optional Ethernet Gateway-Router Module and Fibre Channel Gateway Module into the expansion slot. With the 4xHCA adapter, virtual network interface cards (NICs) and host bus adapters (HBAs) can be created in every server, with a full 10 Gb/s of peak bandwidth.

- The Fibre Channel Gateway Module supports up to two 2 Gb/s Fibre Channel (FC) ports.
- The Ethernet Gateway-Router Module supports up to four 1 Gb/s GigE ports.

E.3 Test Configuration

Figure E-1 shows the overall test configuration. The detailed configuration of each component is described below.

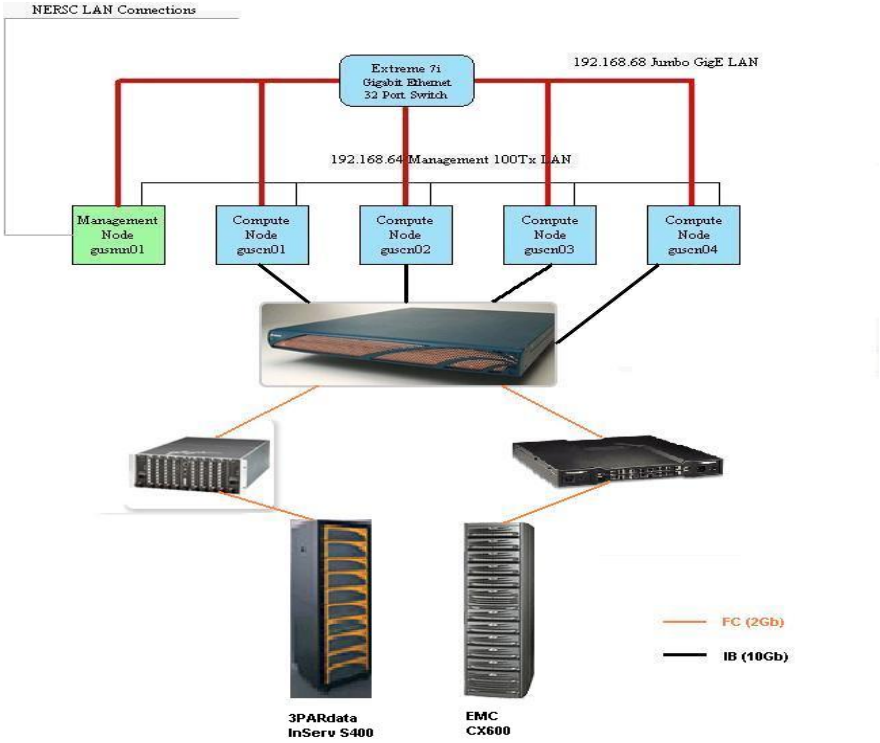


Figure E-1. Diagram showing overall test configuration.

E.3.1 Test Equipment

1. Eight Linux hosts (2.2 GHz, Dual P4 Xeon, 2 GB memory) with the following cards installed:
    - Topspin 4xHCA (driver version 1.1.2)
    - 2 GB Qlogic QLA2300 HBA (driver version 6.05)
    - Intel 1000 GigE
2. Topppin TS 90 Switch
    - IB Module (12 4x ports, firmware version 1.1.2)
    - Ethernet Gateway-Router Module (four GigE ports)
    - Fibre Channel (FC) Gateway Module (two 2 Gb/s FC ports)
3. 2 Gb/s FC Switches (QLogic SANbox2-16 and SANbox2-64)
4. ISCSI Switch (Cisco SN5428)
5. 2 Gb/s FC Storage Devices (EMC CX600 and 3PARdata InServ S400)

6. Other GUPFS equipment: GigE switch, 10/100 Ethernet switch, KVM switch, etc.

E.3.2 Linux Host Configuration

- Supermicro P4DP6 motherboards with six PCI-X slots, two of which were 133 MHz capable
- Dual 2.2 GHz Pentium IV Prestonia Xeon CPUs
- 2 GB of DDR PC2100 ECC memory (512 MB was used to avoid the bounce buffer problem)
- Dual onboard Intel PRO/100 Ethernet interfaces
- Dual onboard U160 Adaptec SCSI controllers
- Onboard VGA graphics
- One 36 GB Ultra 160 LVD 10K RPM SCSI disk drive
- One Qlogic qla2340 133 MHz PCI-X Fibre Channel HBA (low or standard profile)
- One Intel PRO/1000 XT 133 MHz PCI-X Gigabit Ethernet NIC (low or standard profile)
- RedHat 7.3 Linux kernel 2.4.20-13.7smp

E.3.3 Test Software

- NERSC PIORAW (MPI-based Parallel I/O Benchmark), to measure port/fabric performance and scalability
- NERSC's MPTIO (MPI-based Parallel I/O Benchmark), to measure file I/O scalability
- NERSC's METABENCH (MPI-based Parallel File System Meta-Data Benchmark), to measure metadata performance scalability
- lmdd from Lmbench, to measure I/O throughput and scalability
- NetPerf from HP, to measure single-port network throughput
- NERSC's mpiperf, to measure latency

E.4 InfiniBand SRP Performance

E.4.1 Single-Port SRP Performance

This section shows the performance results of a single-port InfiniBand SRP performance, comparing it against the performance of a single 2 Gb/s Qlogic QLA2340 Fibre Channel (FC) HBA.

The benchmark was run with the following two measurements collected:

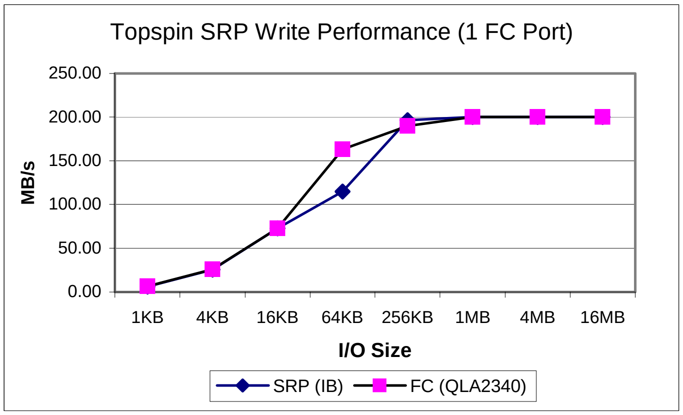
- Raw I/O throughput (MB/sec), and
- CPU utilization (%sy value in the vmstat output)

The single-port SRP performance was obtained from four PIORAW processes running on a single Linux host, accessing a single 2 Gb/s FC storage device, via one of the two FC ports on the FC Gateway Module.

The FC results were obtained from four PIORAW processes, running on the same host, accessing the same 2 Gb/s FC storage device, via a 2 Gb/s Qlogic QLA2340 HBA.

E.4.1.1 Single-port SRP Write Performance

Topspin SRP write performance (MB/s)		
I/O Size	*SRP (IB)*	*FC (QLA2340)*
1 KB	6.35	6.52
4 KB	25.57	25.95
16 KB	72.71	72.81
64 KB	114.68	162.91
256 KB	196.37	189.80
1 MB	200.08	199.80
4 MB	200.09	199.85
16 MB	200.06	199.82

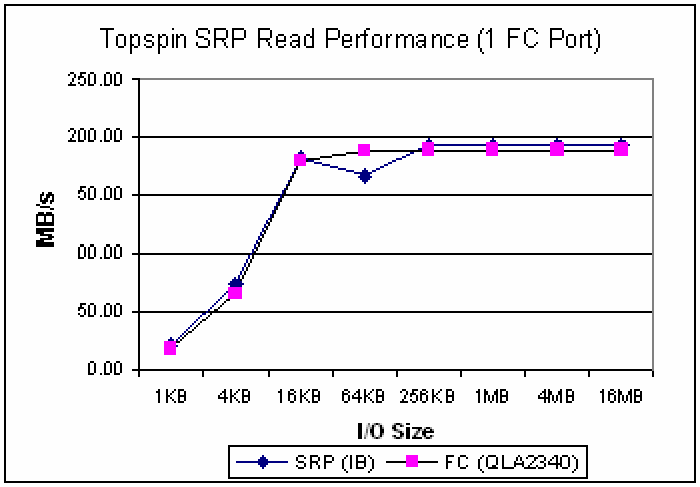


The SRP write performance matches very closely the native FC performance for the different I/O sizes that were used in the tests, except for the 64-KB I/O size. In fact, the SRP performance was slightly better than the native FC performance for the I/O sizes that were larger than 256 KB. This performance difference may have been a manifestation of the implementation or SRP driver and the Qlogic driver. The underlying storage device was capable of sustaining a 200 MB/s write throughput via the 2 Gb/s FC port.

It is not clear why there is a drop in the SRP performance for the 64-KB I/O size. As seen in later test results, there was always a drop in the SRP performance for the 64-KB I/O size.

E.4.1.2 Single-port SRP Read Performance

Topspin SRP read performance (MB/s)		
I/O Size	*SRP (IB)*	*FC (QLA2340)*
1 KB	20.95	17.81
4 KB	73.94	65.14
16 KB	182.50	179.52
64 KB	166.48	188.30
256 KB	193.74	189.06
1 MB	194.06	188.74
4 MB	194.07	188.62
16 MB	194.10	188.70

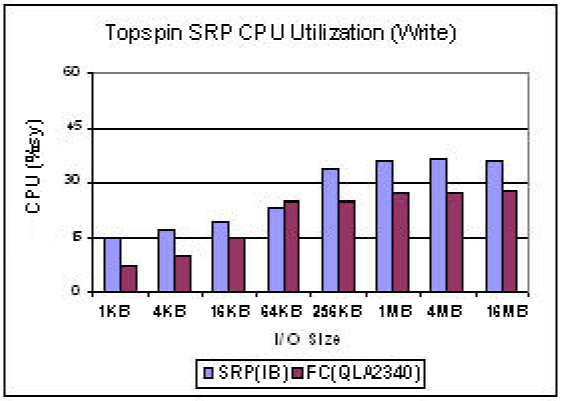


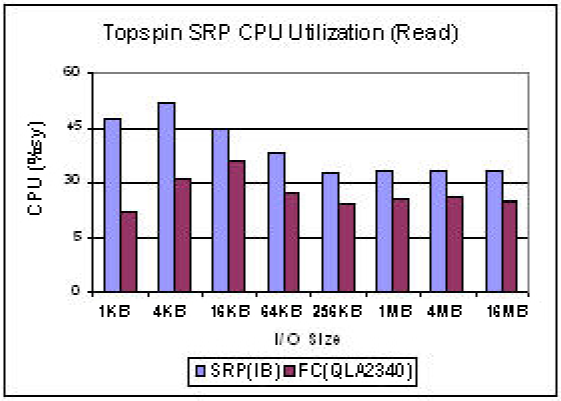
The SRP read performance was slightly better than the native FC performance for the different I/O sizes that were used in the tests, except for the 64-KB I/O size. Again, this performance difference may have been a manifestation of the implementation or SRP driver and the Qlogic driver. The underlying storage device was capable of sustaining a 200 MB/s read throughput via the 2 Gb/s FC port.

Again, we saw the same drop in the SRP read performance for the 64KB I/O size.

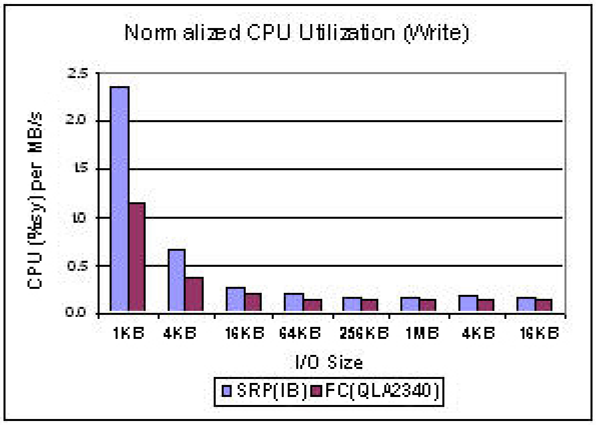
E.4.3 SRP CPU Utilization (single-port)

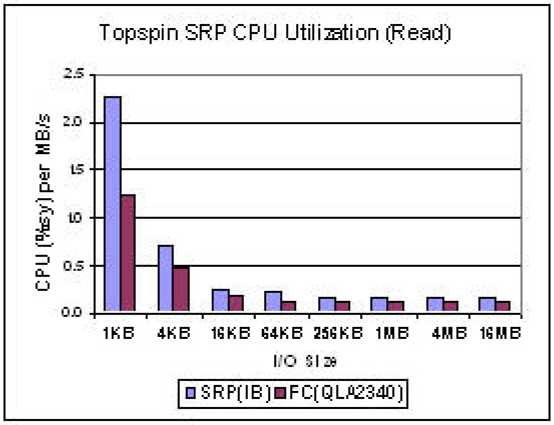
The following two charts show the CPU utilization reported by the vmstat command between the SRP run and the run with the native Qlogic QLA2340 driver.





To do a fair comparison of the CPU overhead between the two test configurations, we normalized the CPU utilization by calculating the CPU utilization per 1 MB/s of data transfer. That is, for each I/O block size, we divided the CPU utilization (%sy) by the throughput number (MB/s). The following two charts show the normalized CPU overhead for both reads and writes.





In general, SRP seemed to consume more CPU cycles than the native Fibre Channel driver. The CPU utilization of the SRP driver was more than 30% higher than that for the QLA2340 driver. The SRP CPU overhead was much higher for reads and writes with I/O sizes smaller than 4 KB.

E.4.4 The Driver Bounce Buffer Problem

During our early tests, the best single-port write performance we were able to achieve was about 21 MB/s. The FC port of the storage device we were using was a 2 Gb/s FC port. This poor performance could have been caused by the known *bounce buffer problem* on a host that has larger than 1 GB system memory.

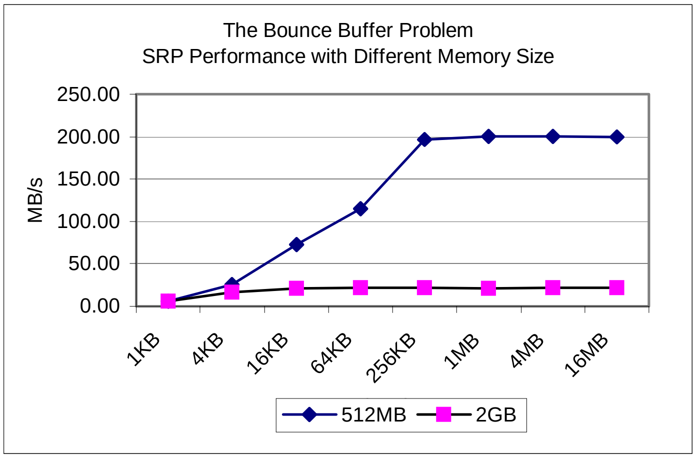
Memory that is above 1 GB is known as high memory. When DMA I/O is performed to or from this high memory, an area known as a bounce buffer is allocated in low memory. During data transfer between a device and high memory, data were first copied through the bounce buffer. It has been observed that systems with intense I/O activity and a large amount of memory can have a performance issue. There are two possible ways to avoid the bounce buffer problem: one is to reduce the size of system memory; the other is to use the latest kernel and the driver with the bounce buffer bypass patch.

The 2-4-20-13.7smp kernel that we were using during the I/O tests had already included a patch for the bounce buffer problem. However, it seemed that the Topspin drivers may not have had the bypass patch. To verify this, we reduced the memory size on the test host from 2 GB to 512 MB.

The following results show the SRP performance running with different memory sizes on the host. It does seem to indicate that we were running into the bounce buffer problem when we were using a 2 GB memory size.

SRP Write performance (MB/s)

I/O Size	Host Memory Size	
	512MB	2 GB
1 KB	6.35	6.11
4 KB	25.57	16.57
16 KB	72.71	20.87
64 KB	114.68	21.68
256 KB	196.37	21.76
1 MB	200.08	21.33
4 MB	200.09	21.59
16 MB	200.06	21.99



E.4.5 Single-HCA SRP Performance

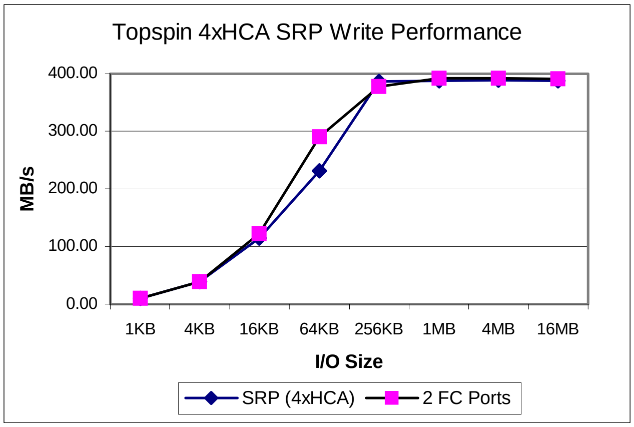
This section shows the I/O benchmark results of the SRP performance from a single 4x HCA (Host Bus Adapter), compared against the aggregate performance obtained from two 2 Gb/s FC ports.

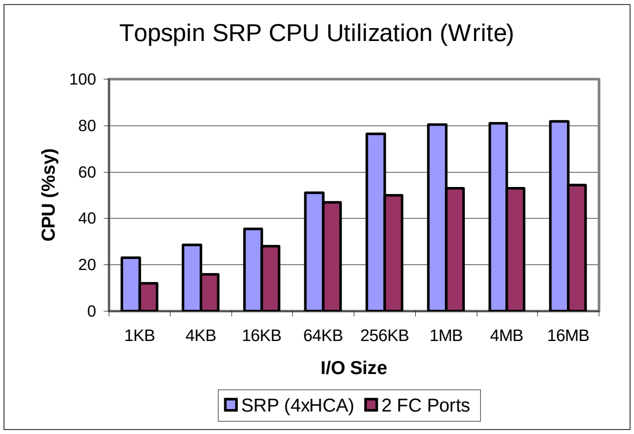
The benchmark was run with the following two measurements collected:

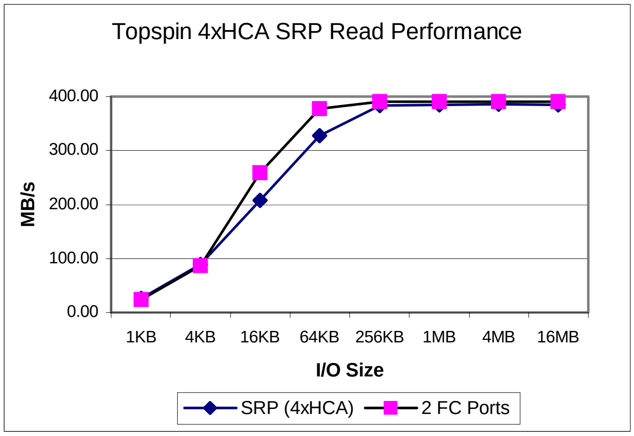
- Raw I/O throughput (MB/sec), and
- CPU utilization (%sy value in the vmstat output)

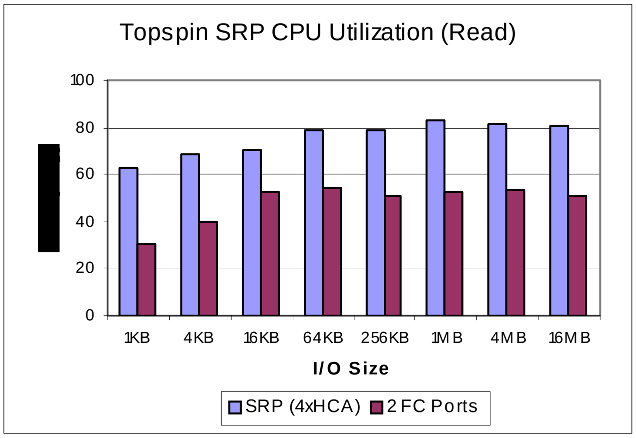
The single-HCA SRP performance was obtained from a single Linux host, using eight PIORAW processes over two FC storage devices, via the two FC ports on the FC Gateway Module.

The FC results were obtained from the tests run individually with each storage device, using a single 2 Gb/s Qlogic QLA2340 HBA.









E.5 Topspin IPoIB (IP over InfiniBand) Performance

Objectives: To measure Topspin IPoIB (IP over InfiniBand) performance in different fabric scenarios.

E.5.1 Test Node Configuration

A. TEST NODES: guscn25, guscn26. Both nodes were configured with:

- Supermicro P4DP6 motherboards with six PCI-X slots, two of which were 133 MHz capable
- Dual 2.2 GHz Pentium IV Prestonia Xeon CPUs
- 2 GB of DDR PC2100 ECC memory
- Dual onboard Intel PRO/100 Ethernet interfaces
- One Intel PRO/1000 XT 133 MHz PCI-X Gigabit Ethernet NIC
- Topspin 4x HCA with IPoIB (IP over InfiniBand) Driver
- Redhat 7.3 Linux 2.4.20-20.7smp kernel

B. SWITCH/PORTS SETUP:

Dell 5224 (gusgs06, Ports 1-4) <=(4)=> Extreme 7i
VLAN 310: Ports 5-12,21-24
VLAN 311: Ports 13-16
VLAN 312: Ports 17-20

Topspin (gusic03, Port 2/1) <-> Dell Port 17
Topspin (gusic03, Port 2/2) <-> Dell Port 21

guscn25 (eth2) <--> Dell (Port 5)
guscn26 (eth2) <--> Dell (Port 6)

C. TEST SOFTWARE: iperf version 1.7.0 (13 Mar 2003) pthreads (default settings, except TCP_WINDOW_SIZE)

D. Topspin TS90 OS Version: 1.1.3-build703

E. Dell Software Version:

```
gusgs06-0#show version
Unit1
 Serial number           :TW03N359704202CS003D
 Service tag             :CMS0221
 Hardware version        :A02
 Number of ports         :24
 Main power status
```

```
      :up
    Redundant power status :not present
   Agent(master)
    Unit id                :1
    Loader version         :1.0.0.0
    Boot rom version       :1.0.0.0
    Operation code version :3.1.0.10
```

E.5.2 Test Scenarios and Test Results

E.5.2.1 10/100 Point-to-Point Performance

Network: 10/100 <-> 10/100 (64 subnet, routed within a 10/100 switch)

```
  TCP_WINDOW_SIZE=85.3 KB (default)

  guscn25 -> guscn26   T1: 94.1 Mb/s
  guscn26 -> guscn25   T1: 94.1 Mb/s
                       T2: 94.1 Mb/s
                       T4: 94.1 Mb/s

  Second Try:

  guscn25 -> guscn26    T1: 94.1 Mb/s, 94.1 Mb/s
                        T2: 94.1 Mb/s, 94.1 Mb/s
                        T4: 94.1 Mb/s, 94.1 Mb/s
  guscn26 -> guscn25   T1: 94.1 Mb/s, 94.1 Mb/s
                        T2: 94.1 Mb/s, 94.1 Mb/s
                        T4: 94.1 Mb/s, 94.1 Mb/s

  TPC_WINDOW_SIZE=256 KB

  guscn25 -> guscn26   T1: 94.1 Mb/s, 94.1 Mb/s
                        T2: 94.1 Mb/s, 94.1 Mb/s
                        T4: 94.1 Mb/s, 94.1 Mb/s
```

Summary: 10/100 performance was 94 Mb/s, independent of how many iperf processes/threads (T1, T2, T4) were used.

E.5.2.2 GigE Point-to-Point Performance

A. Same 68 subnet (routed within a Dell switch):

```
  TCP_WINDOW_SIZE=85.3 KB (default)

  guscn25-ge0 -> guscn26-ge0   T1: 633 Mb/s, 632 Mb/s
                        T2: 989 Mb/s, 989 Mb/s
```

```
             T4: 990 Mb/s, 990 Mb/s
  guscn26-ge0 -> guscn25-ge0   T1: 630 Mb/s, 630 Mb/s
                 T2: 989 Mb/s, 989 Mb/s
                 T4: 990 Mb/s, 990 Mb/s

  TCP_WINDOW_SIZE=256 KB

  guscn25-ge0 -> guscn26-ge0   T1: 990 Mb/s, 990 Mb/s
             T2: 990 Mb/s, 990 Mb/s
             T4: 990 Mb/s, 990 Mb/s
```

B. Between 52 and 68 subnets (Dell <-> E7i <-> Dell):

```
  guscn25-ge0 -> guscn26-ge0   T1: 360 Mb/s, 362 Mb/s
                 T2: 671 Mb/s, 672 Mb/s
                 T4: 989 Mb/s, 989 Mb/s
  guscn26-ge0 -> guscn25-ge0   T1: 361 Mb/s, 360 Mb/s
                 T2: 689 Mb/s, 689 Mb/s
                 T4: 989 Mb/s, 989 Mb/s
```

Summary: iperf was able to achieve ~990 Mb/s with GigE, but it required more than one iperf process/thread. With only one iperf process (T1), the performance was lower (633 Mb/s if the network traffic was within the same Dell switch and 360 Mb/s if the traffic needed to hop through the Extreme 7i switch).

With sufficient numbers of iperf processes, it would saturate the GigE pipe.

By increasing TCP_WINDOW_SIZE to 256 KB, iperf was able to achieve 990 MB/s with one iperf thread.

E.5.2.3 IPoIB Point-to-Point Performance

IPoIB <-> IPoIB (52 subnet, routed within the Topspin switch):

```
  TCP_WINDOW_SIZE=85.3 KB (default)

  guscn25-ib0 -> guscn26-ib0  T1: 967 Mb/s, 989 Mb/s
                 T2: 965 Mb/s, 966 Mb/s
                 T4: 918 Mb/s, 917 Mb/s
  guscn26-ib0 -> guscn25-ib0  T1: 918 MB/s, 888 Mb/s
                 T2: 991 Mb/s, 988 Mb/s
                 T4: 939 Mb/s, 938 Mb/s
                 T8: 923 Mb/s, 920 Mb/s
```

Summary: The IPoIB performance was disappointing. It was even lower than that of the GigE.

E.5.2.4 IPoIB <-> GigE Point-to-Point Performance

A. Between 52 and 68 subnets (Topspin bridging <-> Dell <-> E7i <-> Dell):

```
  TCP_WINDOW_SIZE=85.3 KB (default)

  guscn25-ib0 -> guscn26-ge0  T1: 495 Mb/s, 495 Mb/s
                              T2: 755 Mb/s, 754 Mb/s
                              T4: 758 Mb/s, 759 Mb/s
  guscn26-ge0 -> guscn25-ib0  T1: 797 Mb/s, 796 Mb/s
                              T2: 784 Mb/s, 790 Mb/s
                              T4: 785 Mb/s, 771 Mb/s
```

Summary: This test measured the performance bridging between IB and GigE, with IPoIB and GigE on different subnets. Extreme 7i was used to route the network traffic between the two subnets. The performance was disappointing at around 800 Mb/s. Also, the numbers indicated faster performance from GigE to IB than from IB to GigE. The reason is not clear.

B. Same 68 subnet (Topspin bridging <-> Dell):

```
  TCP_WINDOW_SIZE=85.3 KB (default)

  guscn25-ib0 -> guscn26-ge0  T1: 561 Mb/s, 559 Mb/s
                              T2: 766 Mb/s, 767 Mb/s
                              T4: 767 Mb/s, 766 Mb/s
  guscn26-ge0 -> guscn25-ib0  T1: 794 Mb/s, 794 Mb/s
                              T2: 780 Mb/s, 781 Mb/s
                              T4: 784 Mb/s, 783 Mb/s
```

Summary: This test measured the performance bridging between IB and GigE, with IPoIB and GigE on the same subnet (without hopping through Extreme 7i switch). This performance was also disappointing. Again, it was faster from GigE to IB than from IB to GigE.

Appendix F: Technology Development in FY 2003

The ongoing evaluations conducted by the GUPFS project during FY 2003 included investigations in all three technology areas that are key to the successful deployment of a center-wide shared file system utilizing consolidated storage resources:

- Shared/cluster file systems
- Storage devices
- SAN fabrics

The technologies tested in each of these areas will be discussed in the following sections. In addition to the technologies that were actually tested, alternative and emerging technologies that were investigated and tracked throughout the year will also be discussed.

F.1 File System Technologies

The key technology for the GUPFS project is the shared file system. Without a reliable, scalable, high-performance, shared file system, it will be impossible to deploy a center-wide file system. Historically, there have been two major technologies for sharing storage between systems: client-server based network attached storage (NAS) and storage area network (SAN) [18]. (Another way to view these technologies is as network-based distributed shared file system, and block-storage-based shared file systems, respectively.) FY 2003 saw the emergence of a third major technology for sharing storage in the form of object based shared file systems accessing storage over Internet Protocol (IP) based networks.

Network attached storage is a general term for storage that is accessible to client systems over general-purpose IP networks from the local storage of network-attached servers. Data transfers between storage servers and clients are performed over a network. This results in the file systems being limited by network bandwidth, network protocol overhead, the number of copy operations associated with network transfers, and the scalability of each server. The file system performance is often limited by the bandwidth of the underlying IP network. The most common example of a NAS file system is the Network File System (NFS). NFS exhibits many of the previously mentioned performance limitations.

Storage area networks, by providing a high-performance network fabric oriented toward storage device transfer protocols, allow direct physical data transfers between hosts and storage devices, as though the storage devices were local to each host. Currently, SANs are implemented using Fibre Channel (FC) [19,20] protocol-based high-performance networks employing a switched any-to-any fabric. The emergence of alternative SAN protocols, such as iSCSI [8], FCIP [21], and SRP [22], are enabling the use of alternative fabric technologies, such as Gigabit Ethernet and InfiniBand, as SAN fabrics. Regardless of the specific underlying fabric and protocol, a SAN allows hosts connected to the fabric to directly access and share the same physical storage devices. This permits high-performance, high bandwidth, and low latency access to shared storage. A shared-disk file system will be able to take full advantage of the capabilities provided by the SAN technology.

Sharing file systems between multiple systems is a very difficult problem because file system coherency[‡] must be maintained by synchronizing file system metadata operations and coordinating data access and modification. Maintaining file system coherency becomes very challenging when multiple independent systems are accessing the same file system and physical devices. Shared file systems that directly access data on shared physical devices through a SAN are commonly categorized as being either asymmetric or symmetric. Asymmetric shared file systems allow systems to share and directly access data but not metadata, which is maintained by a centralized server that provides synchronization services for all clients accessing the file system. Symmetric shared file systems share and directly access both data and metadata. Coherency of data and metadata is maintained through global locks, and distributed lock management is performed directly between participating systems.

Shared file systems are not common because implementing them is inherently difficult [23]. Of the two types, asymmetric shared file systems are used more often because centralized metadata servers are easier to implement than distributed metadata management. However, symmetric shared file systems promise better scalability without the bottleneck and single point of failure problems inherent in asymmetric systems.

F.1.1 File System Technology Arena Developments during FY 2003

Over the last year, shared file system technology has continued to undergo turbulent changes. Some once-promising shared file systems have faded; other existing shared file systems continued to make incremental progress; and many new shared file system vendors and alternative technology approached.

The Sistina GFS file system [2] continued to fade, mostly through the failure to add interesting features addressing missing functionality. Some strides were made in scalability, with GFS being supported on systems with between 128 and 256 nodes.

The ADIC StorNext file system [3] received a boost through its selection by Cray as the file system for the X1, improved scalability, and remained the leader in cross-platform and multiple 0S support. Unfortunately, ADIC appears to be planning to drop DMAPI support in StorNext.

IBM’s General Parallel File System (GPFS) [5] made great strides in functionality, multiple platform and OS support, attractive licensing and plans for future functionality and availability including enhanced DMAPI functionality. In addition to being used on NERSC’s SP system, GPFS is installed and being tested on the LBNL Alvarez Linux cluster.

The once promising InfinARRAY file system [24] became static and no longer actively marketed or supported.

IBRIX file system [25] development was very delayed. Its expected beta testing was postponed from early calendar 2003 to late summer 2003, and even then was still not ready for testing. On an encouraging note, IBRIX development has continued.

[‡] File system coherency is the correctness, integrity, and consistency of the data and meta-data structures that comprise the file system.

The Panasas file system (Panasas ActiveScale File System) [26] became a reality, appearing in late summer of 2003. This is a promising object-based file system that uses Ethernet-attached storage, employing a modified version of iSCSI. Panasas currently only offers an integrated software and storage solution, which limits options.

IBM's StorageTank [27] file system began to appear and moved towards becoming a released product rather than just a development project. StorageTank, recently renamed SAN File System, shows promise as a shared file system. Architecturally, it is reminiscent of Panasas' Direct Flow and Lustre, differing mainly in being block based and using iSCSI and direct Fibre Channel.

The Lustre file system [4] was deployed during FY 2003. Pre-releases began as early as November 2002 with Version 0.5. The much-improved Version 0.6 appeared in March 2003, but still was extremely difficult to install and poorly documented. This version was tested and found to perform very poorly and to be unable to complete all of the standard GUPFS benchmarks. Pre-releases of Lustre continued periodically throughout the remainder of the fiscal year, but did not improve markedly. Lustre remains a promising file system technology, but definitely remained a work in progress and in development for the rest of FY 2003.

The GUPFS project was able to conduct evaluations of a number of these important shared file systems during FY 2003. These included tests of Sistina's GFS 5.1 and 5.1.1, a beta test of GFS 5.2, a test of ADIC StorNext 2.0, tests of a variety of Lustre pre-release version with the focus on Lustre 0.6, and tests of IBM's GPFS for Linux 1.3. GPFS was evaluated both on the GUPFS testbed and on a larger scale on the LBNL Alvarez Linux cluster.

F.1.2 Storage Technologies

On the storage technology side, FY 2003 saw substantial movement towards scalable, higher performance storage, with increased connectivity and bandwidth. Despite this, some storage technologies still have problems with scalability in the number of client system host adapters (initiators) that can be simultaneously serviced.

In the arena of scalable, high-performance storage, the year saw continued development of Yotta Yotta [13] and 3PARdata [15] start up companies and products. Yotta Yotta shifted its focus to providing only the storage controllers, without reselling any storage, and emphasized its coherent storage access capabilities for geographically distributed storage.

3PARdata brought its InServ storage offering to market. The InServ product targets large storage capacity and scalability in sustained aggregate bandwidth, with support for a large number of clients.

The year also saw the decline of Maximum Throughput, both as a high-performance storage vendor and as a file system vendor. During the year, Data Direct Networks [28] rose to ascendancy as the preferred provider of scalable, high-performance storage at many U.S. Department of Energy (DOE) national laboratories and National Science Foundation (NSF) centers with its S2A8000 and S2A8500 products.

One of the important trends in storage that developed during FY 2003 was the appearance of Serial ATA (SATA) technology [29]. SATA is an evolutionary replacement for the Parallel ATA storage interface currently in common use for IDE disks. In addition to packaging improvements in connectors and cabling, SATA provides a 1.5 gigabits per second (Gb/s) ((150 MB/s) point-to-point connection to each drive. The ten-year roadmap for SATA calls for 3.0 Gb/s (300 MB/s) and 6.0 Gb/s (600 MB/s) versions to be introduced [30].

The importance of SATA lies in enabling inexpensive, lower performance, lower reliability IDE disk drives to be incorporated in large quantities into storage devices, much as SCSI and Fibre Channel disks have been in the past. This permits large quantities of lower performance bulk disk storage to be made available at much lower cost. However, this comes at the price of less reliability and reduced performance, both in per-spindle transfer rates and I/O operations per second. Storage devices can be designed to overcome the reduced per-spindle transfer performance, but limited per-spindle I/O operation rates are much harder to overcome. This makes such storage far less attractive for small transfer and seek-dominated applications, such as home file systems and database applications.

Overall, the major changes in the storage arena were limited to the introduction of SATA, the general migration to 7200 RPM disks for IDE storage, some modest per-spindle increases in maximum disk capacity, and the migration to complete 2 Gb/s Fibre Channel storage devices (with 2 Gb/s FC links for both hosts and internal disks). Otherwise, the normal annual 50% decrease in price/GB and doubling of per-plater capacity remained in effect.

We conducted a full evaluation of the 3PARdata InServ storage, and began evaluation of Panasas storage during FY 2003. We continued working with Yotta Yotta to steer their product development towards storage controllers with the necessary performance and scalability.

We still harbor concern about the ability of some storage devices to deal with many simultaneous initiators (host interfaces). There is some possibility that this is really a problem with either the FC host bus adapters **(**HBAs) or their drivers. When there are too many initiators issuing I/O operations at the same time, a huge number of I/O requests may fill up the SCSI command queue on the storage system. When the command queue fills up, the storage system returns a “command-full” status in an attempt to slow down the hosts. However, the host driver software frequently does not know how to process the “command-full” status returned by the storage system, and the hosts continue to send I/O requests to the storage system. As a result, some I/O requests are dropped and the hosts hang, waiting for I/O completion. Some storage systems appear to have shorter command queue depths, and this problem shows up more often on these storage systems.

F.2 SAN Fabric Technologies

An important goal of the GUPFS project is to evaluate the performance characteristics of individual components and the interoperability of these components in a heterogeneous multi-vendor and multiple-fabric environment. The ability of the file systems to operate in such a mixed environment is very important to the ultimate success of the GUPFS project as the file system is expected to persist even though new technologies for fabrics and interconnects will be introduced by the new systems installed at NERSC in future years.

FY 2003 saw the completion of the migration to 2 Gb/s Fibre Channel technology in all elements of SAN fabrics: switches, host adapters, storage devices and disk drives. FY 2003 also saw many fabrics evolving rapidly and a trend towards extensive bridging between fabrics, with a convergence towards Ethernet as the common fabric to which all bridge. The year also saw the adoption of the iSCSI standard, which enables block storage transfers to be conducted over any fabric or interconnect capable of supporting IP packets. This, in conjunction with the extensive bridging between fabrics, has greatly increased the options available for SAN fabrics and for the use of the same fabrics in both block storage transfers and message passing.

During FY 2003, there were substantial evolutions in several fabrics. InfinBand evolved from 2.5 Gb/s (1x) to 10 Gb/s (4x), and 1x InfiniBand quickly disappeared. By the end of the fiscal year, 30 Gb/s (12x) InfiniBand products, primarily switches, began to come on the market. During the same time, Gigabit Ethernet became a commodity technology, with great reduction in both switch and Network Interface Card (NIC) costs. 10 Gigabit Ethernet products also began to appear as early beta products and a few became available as full releases by the end of the year.

F.2.1 Fibre Channel

In addition to the migration from 1 to 2 Gb/s for all components from disk drives, switches, and host interfaces, FY 2003 saw the introduction by multiple vendors of much larger, enterprise-class switches with port counts from 64 to in excess of 100. Of particular interest was the entry of Cisco into the FC/SAN market, and the introduction of the MD950x enterprise-class FC switches supporting 128 ports, and 40 Gb/s inter-switch link bandwidth. (An iSCSI blade is also planned.)

During the last half of FY 2003, information about Fibre Channel’s future technology roadmap began to appear. It became apparent that two technology updates were being planned. The first was an increase to 10 Gb/s bandwidth, initially for switches and storage device, but not disk drives. This was a solid element of the future roadmap, and a necessary step for Fibre Channel to survive given that other fabrics such are Ethernet and InfiniBand are moving to 10 Gb/s and beyond. The second was more problematical and without either a clear commitment or consensus as of the end of FY 2003. This was the addition of a 4 Gb/s bandwidth to augment the 10 Gb/s bandwidth. The perceived need for a 4 Gb/s Fibre Channel arises because the 10 Gb/s bandwidth Fibre Channel will not be backward compatible with 1 and 2 Gb/s Fibre Channels due to differences in the signaling hardware needed for 10 Gb/s equipment. The 4 Gb/s Fibre Channel was intended to provide an increase in performance while maintaining backward compatibility with 1 and 2 Gb/s equipment to protect legacy investments in such equipment. With solid plans for, at minimum, a 10 Gb/s bandwidth channel on its roadmap, Fibre Channel seems assured of surviving at least one more generation.

It is important to understand the improved Fibre Channel technology and determine what level of storage performance can be achieved. This is because backend storage, the storage attached to disk controllers and servers, for most plausible GUPFS file system solutions will likely be Fibre Channel–connected in the near term, even if client storage access is not through Fibre Channel and/or the disks are not Fibre Channel disks.

F.2.2 iSCSI

Alternative SAN fabric technologies are beginning to appear. Using Ethernet as a SAN fabric is now becoming possible due to the iSCSI standard, which is a block storage transport protocol. The iSCSI protocol allows the standard SCSI packets to be enveloped in IP packets and transported over standard Ethernet infrastructure, which allows SANs to be deployed on IP networks.
This option is very attractive as it allows lower-cost SAN connectivity than can be achieved with Fibre Channel, although with lower performance. It will allow large numbers of inexpensive systems to be connected to the SAN and to use the cluster file system through commodity-priced

components. While attractive from a hardware cost perspective, this option does incur a performance impact on each host due to increased traffic through the host's IP stack.

In many respects, FY 2003 was the year in which iSCSI became a reality. The iSCSI standard coalesced at the end of the first quarter of FY 2003, and was adopted as a standard in January 2003. Prior to the adoption of the standard only a few speculative iSCSI products existed, and software drivers for Linux and other OS's were changing rapidly as the standard progressed. After iSCSI became an official standard, the software drivers for iSCSI using normal system IP stacks and Ethernet NICs quickly converged on the finalized standard. The movement toward support for the official standard by the speculative iSCSI products was less swift and more uneven.

The GUPFS project obtained a Cisco SN5428 Storage Router at the end of FY 2002 in order to begin evaluating iSCSI and bridging between Ethernet and Fibre Channel fabrics. At the time it was obtained, the SN5428 router supported provisional versions of the iSCSI standard. In conjunction with appropriate compatible Linux drivers, the SN5428 performed a successful initial evaluation of iSCSI and fabric bridging. Upon adoption of the iSCSI standard, the SN5428 and Linux iSCSI drivers quickly supported the finalized standard and demonstrated single node disk data transfer rates approaching the limits of Gigabit Ethernet. However, achieving such I/O rates imposed a sever system CPU load on a node. This was not unexpected, and the GUPFS project was interested in evaluating the iSCSI HBAs that were available or expected to be available during the fiscal year.

The iSCSI HBAs are designed to offload the system CPU overhead associated with processing the iSCSI protocol and traversing the IP software stack. They do this by presenting the system with a device that appears to be a SCSI HBA with connected disks rather than an Ethernet NIC. The CPU overhead associated with SCSI HBAs is significantly less (often totaling only a few percent) than the CPU overhead incurred during processing the same amount of data through an IP software stack and Ethernet stack. An iSCSI HBA moves the iSCSI protocol handling and the entire IP software stack processing to the HBA card, relieving the system CPU from these tasks. In this regard iSCSI HBAs are similar to TCP Offload Engine (TOE) cards which offload the IP stack processing to the TOE cards. Both iSCSI HBAs and TOEs are intended to reduced the CPU overhead of Ethernet traffic.

We evaluated the very early Intel IP Pro 1000 iSCSI HBA, obtained before the iSCSI standard was adopted. The drivers for these iSCSI HBAs were initially proprietary closed source distributed as binary modules. Only antique RedHat Linux distributions (7.2) and kernel versions were supported. Updates were extremely infrequent. Because of the binary modules and antique Linux kernels required, none of the shared file systems being evaluated could be tested with the iSCSI HBA. However, testing was successfully conducted using raw I/O to storage. Details of this testing appear later in this document. The iSCSI HBA performance was lackluster, being appreciably less than that of the pure iSCSI software solution. Although the iSCSI HBA had less CPU overhead than the purely software solution, it was surprisingly high, coming in at 50% of that of the software solution.

In the second half of FY 2003, Intel released a updated iSCSI HBA, and exchanged some of the GUPFS project's existing cards. The new cards used Linux Open Source drivers. This was a great step forward, but unfortunately the drivers were once again for a kernel version too old to be used with any of the kernels needed for file system and fabric testing. To our disappointment, because the kernel version supported by the drivers was so old, it could not be rebuilt for the newer kernels due to major changes in kernel architecture. Attempts to port the driver to the newer kernel architecture were unsatisfactory, and testing was abandoned in the face of schedule pressures.

Although the iSCSI HBA arena has been disappointing so far, the prospects for iSCSI fabric bridges is looking substantially better. The SN5428 has provided a successful demonstration of the use of the iSCSI protocol for storage access over IP networks and of bridging between Gigabit Ethernet and Fibre Channel networks. Fortunately, Cisco has announced plans to support iSCSI on their large enterprise class MD9500 series Fibre Channel switches. This will be done through the use of line cards with Gigabit Ethernet connections incorporating Fibre Channel bridges that can be plugged into the switches in addition to Fibre Channel line cards, to provide an integrated Fibre Channel/Ethernet fabric. Several other vendors are pursuing similar plans.

With the advent of fabric bridges from several interconnects and fabrics to Ethernet (such as InfiniBand, Fibre Channel, and Myrinet), iSCSI shows promise as being a standard mechanism for accessing storage through complex heterogeneous networks.

F.2.3 InfiniBand

The newly emerged InfiniBand (IB) interconnect continued to mature and evolve during FY 2003.Also during FY 2003, however, the InfiniBand vendor landscape continued to shrink, most notably by the departure of IBM from the ranks of InfiniBand manufacturers, leaving Mellanox as the primary silicon provider. In spite of this, InfiniBand continued to progress and make market inroads.

During FY 2003, InfiniBand progressed from 1x (2.5 Gb/s) host adapters and switches to 4x (10 Gb/s) adapters and switches. Many vendors not only moved to 4x InfiniBand, but produced two generations of 4x equipment. By the end of the FY 2003, 1x InfiniBand had not only been completely supplanted by 4x IB, it had been discontinued. Towards the end of FY 2003, vendors were preparing plans for, and in some cases producing early versions of, 12x (30 Gb/s) IB switches. Because the highest performing host bus for which IB adapters were made, PCI-X, is barely able to service 10 Gb/s data rates, no vendors were planning to produce 12x IB host adapters until the more scalable PCI-Express becomes available in the second half of FY 2004. Although AMD’s HyperTransport was capable of servicing 4x and 12x IB adapters, no HyperTransport Host Channel Adapters (HCAs) were produced. This appears to have been the result of pressure on Mellanox by Intel.

In addition to plans for 12x IB switches, InfiniBand vendors were making plans at the end of FY 2003 for larger switches. The next commonly planned switch size was 96 4x IB ports. Configurations with 32 12x ports and various combinations of 4x and 12x ports also were being planned. One switch configuration targeted for building large clusters was designed to have 64 4x ports for connecting hosts, and 10 12x ports for inter-switch links for creating mesh fabrics for large clusters.

As the year progressed, a number of integrators, such as APPRO and Linux Networks, began offering Linux clusters with integrated InfiniBand interconnects, as well as clusters based on InfiniBand blade systems. More vendors offering InfiniBand based clusters appeared throughout the year. This is encouraging as it indicates a growing market and increased acceptance of InfiniBand. The InfiniBand vendors have indicated that their major markets at the end of the Fiscal Year were scientific clusters and the large financial institutions. Overall, InfiniBand seems to be progressing steadily, although its general acceptance, rate of adoption, and movement towards commodity status are less than previously expected.

F.2.4 High-Speed Interconnect

With the increased interest in large-scale shared file systems, several promising shared file system architectures have appeared. Many of these new file system architectures are designed for the high-performance cluster environments, typically with some kind of high-speed interconnect for the messaging traffic. Many of the new architectures perform storage transfers between client nodes and storage nodes over the high-speed interconnect.

During FY 2003, there were incremental improvements in system interconnects. Myricom introduced a PCI-X version of Myrinet, Revision D, with somewhat improved latency and transfer rates. InfiniBand 4x began to appear as a cluster interconnect and as an interconnect for blade server systems. Quadrics improvements were planned with Elan 4, but Quadrics remained very expensive. Gigabit Ethernet became a commodity, making it a usable medium-performance system interconnect.

From the perspective of the GUPFS project, the most important development in high speed interconnects during FY 2003 was the development of fabric bridges between the various interconnects and other fabrics, such as Ethernet and Fibre Channel. All the interconnects supporting fabric bridges have bridges to Gigabit Ethernet at a minimum. This has a number of important ramifications. First, system interconnects can now be directly connected to Ethernet LANs and the WAN at the switch level, without having to be forwarded through systems. Second, this means that Ethernet, at both one and ten gigabits, is likely to be the common bridge segment used to tie all other interconnects and fabrics together. Third, either network-based or block-storage-based shared file systems can be deployed on the same set of interconnects and fabrics. These are likely to be important considerations for deploying a shared file system in a heterogeneous fabric environment with multiple systems having different interconnects.

Appendix G: Acronyms

AG	access group
ASCI	Advanced Simulation and Computing Program (formerly Accelerated Strategic Computing Initiative)
CIFS	Common Internet File System
CLI	command line interface
CPU	central processing unit
CRC	cyclical redundancy check
CVFS	CentraVision File System (ADIC)
DAE	disk array enclosure
DMA	Direct Memory Access Protocol
DMAPI	Data Management Application Program Interface
DMEP	Device Memory Export Protocol
DMF	Data Migration Facility
DOE	U.S. Department of Energy
ECC	error-correcting code
ERCAP	Energy Research Computing Allocation Process
FC	Fibre Channel
FCIP	Fibre Channel over IP
FY	fiscal year
GB	gigabyte (8 gigabits)
Gb	gigabit
GB/s	gigabyte per second (8 Gb/s)
Gb/s	gigabit per second
GFS	Global File System (Sistina)
GigE	Gigabit Ethernet
GNBD	global network block device
GPFS	General Parallel File System (IBM)
GUI	graphical user interface
GULM	Global Universal Lock Manager (Sistina)
GUPFS	Global Unified Parallel File System
HBA	host bus adapter
HCA	host channel adapter
HDF	Hierarchical Data Format

HP	Hewlett-Packard
HPC	high-performance computing
HPSS	High-performance Storage System
HSM	hierarchical storage management
HSSDC	high speed serial data connector
IB	InfiniBand
IBx	InfiniBand Expansion
IC	in-cache
I/O	input/output
IOPS	I/O operations per second
IP	Internet Protocol
iSCSI	Internet small computer system interface
ISL	Inter-Switch Link
KVM	keyboard, video, mouse switch
LAN	local area network
LBNL	Lawrence Berkeley National Laboratory
LUN	logical unit number
LVM	logical volume management
MDS	metadata server
MHz	megahertz
MPI	Message Passing Interface
MPI-I/O	Message Passing Interface-I/O
MPP	massively parallel processing
MPTIO	MPI program that performs I/O operations
NAL	network abstraction layer
NAS	network attached storage
NERSC	National Energy Research Scientific Computing Center
NIC	network interface card
NFS	Network File System
NSD	Name Server Daemon
OC	out-of-cache
OLTP	online transaction processing
OS	operating system
OSF	Oakland Scientific Facility
OST	object storage target

PCI	peripheral component interconnect
PDF	IBM Profile-Directed Feedback
PCI-X	peripheral component interconnect extension
PDSF	Parallel Distributed Systems Facility
PI	principal investigator
PIORAW	MPI benchmark test to measure raw I/O
POSIX	Portable Operating System Interface
PXE	Pre-Execution Environment (Intel)
RAID	redundant arrays of independent disks
RAM	random access memory
RDMA	Remote Direct Memory Access Protocol
RISC	reduced instruction set computing
RFI	Request for Information
RFP	Request for Proposal
SAN	storage area network
SANFS	IBM file system, formerly known as the IBM StorageTank file system
SCSI	small computer system interface
SE	storage engine
SFP	Small Form Pluggable
SGSFS	Scalable Global Secure File System
SP	storage processor (not to be confused with IBM SP supercomputer)
SPE	storage processor enclosure
SRP	SCSI RDMA Protocol
SRU	storage resource unit
TB	terabyte
TCP	Transmission Control Protocol
TLB	translation lookaside buffer
TOE	TCP Offload Engine
UDP	User Datagram Protocol
VEx	Virtual Ethernet Exchange
VFx	Virtual Fibre Channel Exchange
VGA	video graphics array
XDSM	Data Storage Management Application Program Interface
WWN	Word Wide Name
YY	Yotta Yotta

Appendix H: References

1. The Global Unified Parallel File System (GUPFS) Project: FY 2002 Activities and Results, http://www.nersc.gov/aboutnersc/pubs/GUPFS_02.pdf.

2. Global File System (GFS), Sistina Software, Inc., http://www.sistina.com/downloads/datasheets/GFS_datasheet.pdf.

3. StorNext File System, Advanced Digital Information Corporation (ADIC), http://www.adic.com/us/collateral/stornextfs.pdf.

4. "Lustre: A Scalable, High-Performance File System," Cluster File Systems, Inc., November 2002, http://www.lustre.org/docs/whitepaper.pdf.

5. IBM General Parallel File System (GPFS), IBM Corporation, http://www-1.ibm.com/servers/eserver/ pseries/software/sp/gpfs.html and http://www-1.ibm.com/servers/eserver/clusters/software/gpfs.html.

6. Brocade SilkWorm 3800 Enterprise Fabric Switch, Brocade Corporation, http://www.brocade.com/ products/silkworm/silkworm_3800/sw_3800.jsp.

7. Qlogic SANbox2 16-port Fibre Channel Fabric Switch, Qlogic Corporation, http://www.qlogic.com/ products/sanbox/sanbox_2gb.asp.

8. Julian Satran, "iSCSI" (Internet Small Computer System Interface) IETF Standard, January 24, 2003, http://www.ietf.org/internet-drafts/draft-ietf-ips-iscsi-20.pdf.

9. Cisco SN 5428 Storage Router, Cisco Corporation, http://www.cisco.com/warp/public/cc/pd/ rt/5420/ps2162/prodlit/s5428_ds.pdf.

10. Intel PRO/1000 T IP Storage Adapter, http://www.intel.com/network/connectivity /resources/doc_library/data_sheets/pro1000_T_IP_SA.pdf.

11. InfinIO Shared I/O System, InfiniCon Systems, Inc., http://www.infinicon.com/pdf/ InfinIO_7000_Data_Sheet.pdf.

12. Topspin 90 Switched Computing System, Topspin Communications, http://www.topspin.com/ solutions/t90.html.

13. Yotta Yotta NetStorager GSX 2400, Yotta Yotta, Inc., http://www.yottayotta.com/pages/products/overview.htm.

14. EMC CLARiiON CX600, EMC Corporation, http://www.emc.com/products/systems/ clariion_cx246.jsp.

15. 3PAR InServ Storager Server S400, 3PARdata, Inc., http://www.3pardata.com/documents/ 3PAR_InServ_techspecs_01.1.pdf.

16. Brocade SAN Plan Guide, Brocade Communications Systems, Inc., 2004, http://www.brocade.com/products/WYSP/images/WYSPII_guide_0225.pdf.

17. Building Open SANs with Multi-Switch Fabrics, Qlogic Corporation, 2001, http://www.qlogic.com/documents/datasheets/knowledge_data/whitepapers/building_open_sans.pdf

18. “Comparing Storage Area Networks and Network Attached Storage.” White paper, Brocade Communications Systems, Inc., 2001, http://www.brocade.com/san/white_papers/pdf/SANvsNASWPFINAL3_01_01.pdf.

19. A. Benner, Fibre Channel: Gigabit Communications and I/O for Computer Networks. Boston, McGraw Hill, 1996.

20. FC: Fibre Channel Protocol, ftp://ftp.t10.org/t10/drafts/fcp/fcp-r12.pdf.

21. Raj Bhagwat, Murali Rajagopal, and Ralph Weber, “Fibre Channel Over TCP/IP (FCIP),” IETF Draft Standard, August 28, 2002, http://www.ietf.org/internet-drafts/draft-ietf-ips-fcovertcpip-12.pdf.

22. SRP: SCSI RDMA Protocol, ftp://ftp.t10.org/t10/drafts/srp/srp-r16a.pdf.

23. Chandramohan A. Thekkath, Timothy Mann, and Edward K. Lee, “Frangipanni: A Scalable Distributed File System,” Proceedings of the Sixteenth ACM Symposium on Operating System Principles, pp. 224–237, October 1997.

24. InfinARRAY File System, Max-Throughput, Inc., http://www.max-t.com/html/products/infinarray.html.

25. IBRIX File System, IBRIX, Inc., http://www.ibrix.com/products.php.

26. Panasas ActiveScale File System, Panasas Inc., http://www.panasas.com/docs/ Panasas_ActiveScale_DS.pdf.

27. IBM TotalStorage SAN File System (SANFS, also known as Storage Tank), IBM Corporation.

28. DataDirect Networks S2A 8500 Storage Networking Controller, DataDirect Networks, Inc., http://www.datadirectnetworks.com/pdfs/ddn_s2a_8500_datasheet.pdf.

29. Serial ATA, The Serial ATA Working Group, http://www.serialata.org/collateral/index.shtml.

30. About SATA, The Serial ATA Working Group, http://www.serialata.org/about/index.shtml.