Considerations and Performance Evaluations
Of Shared Storage Area Networks
At NASA Goddard Space Flight Center

Hoot Thompson
Patuxent Technology Partners, LLC
11030 Clara Barton Drive
Fairfax Station VA 22039-1410
hoot@ptpnow.com

Curt Tilmes
NASA GSFC
Greenbelt MD 20771
Curt.A.Tilmes@nasa.gov

Robert Cavey
NASA / ICS, Inc.
9111 Edmonston Road
Greenbelt MD 20770-1544
rcavey@ltpmail.gsfc.nasa.gov

Bill Fink
NASA GSFC
Greenbelt MD 20771
William.E.Fink@nasa.gov

Paul Lang
NASA GSFC
Greenbelt MD 20771
lang@eiger.nasa.atd.net

Ben Kobler
NASA GSFC
Greenbelt MD 20771
Benjamin.Kobler-1@nasa.gov

Abstract

The NASA Goddard Space Flight Center (GSFC) in Greenbelt Maryland is exploring advanced storage architectures for retaining and distributing its large data holdings. As a research vehicle, a multi-building Storage Area Network (SAN) was deployed at GSFC in early 2002. The initial objective was to demonstrate the feasibility and advantages of fibre channel-connected, centralized storage as it applies to a campus installation. The secondary objective was to illustrate the advantages of a heterogeneous SAN shared file system that would allow a single instance of data to be globally shared amongst multiple SAN-connected clients on different platforms. The GSFC SAN has since been extended to include off-campus connections for evaluating the Internet Protocol (IP) as an option for connecting a broader, more geographically dispersed user base. The focus of this paper is the series of tests conducted to characterize distance data sharing using both native FC and IP based technologies. These experiments include both standard I/O benchmarks as well as representative GSFC applications.

1.0   Introduction

Storage Area Network (SAN) technology has been maturing and evolving rapidly over the last five years, pushing both the bandwidth and interconnectivity fronts. Individual fibre channel (FC) links, the interconnect technology of choice, are now running at 2 Gigabit/sec with 10 Gigabit/sec links on the horizon. Enterprise class switches are available that support upwards of 256 ports.   Using cascading and various interconnect strategies, thousands of ports can be joined together into a single logical SAN.

Several key factors continue to drive SAN adoption, among them being:

- Recognized cost savings associated with consolidated storage,
- Greater utility of centralized data
- Data protection through replication for disaster recovery.

Most SANs have been relatively 'local' from a topology point of view. However, several emerging technologies are expanding the traditional distance boundaries. In this study, the SAN consisted of logically centralized disk-based storage that is addressable at the block level by multiple heterogeneous computers or clients interconnected through a switch fabric. Figure 1 is a simplified overview of a representative SAN. Connection between clients, switches and storage are predominantly fibre channel (FC) with the SCSI protocol used to manage block-level transfers. Installed in the client computers are host bus adapters (HBA) and their associated drivers. The HBA's drivers log into the switch fabric that is typically comprised of one or more multi-ported FC switches, with the most common products having either 8 or 16 ports. These switches are usually full-crossbar allowing any-to-any connectivity. Storage is generally RAID (redundant array of independent disks) or JBOD (just a bunch of disks) and it is usually partitioned into logical units or LUNs for presentation to the clients through the fabric. Representative file systems can be built on individual LUNs or the LUNs can be striped or concatenated as required for performance or capacity reasons.
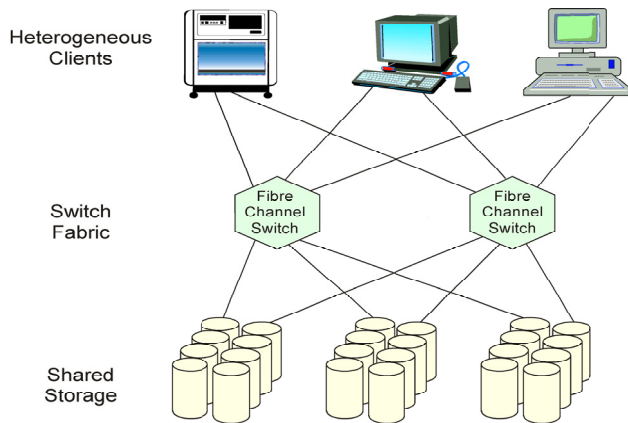
Figure 1 – Typical SAN Architecture

In most SANs, portions of the shared storage are assigned to a given client with file systems and data owned specifically by that client. This is usually accomplish through fabric zoning, LUN masking or a combination of the two. However there are products available that facilitate direct data sharing by multiple clients at the block level. The CentraVision™ File System (CVFS) by ADIC and the Global File System (GFS) by Sistina are two such products. CVFS, or StorNext File System as it is now called, is a heterogeneous SAN file system, meaning clients running different operating systems can simultaneously mount, write and read the same file system and directly manipulate data in that file system. CVFS utilizes a centralized metadata function to permit the sharing. GFS is a Linux-only design and takes a distributed approach to managing metadata.

For completeness, the management and administrative aspects of SANs also require attention. Such concerns are loosely categorized under the term storage resource management (SRM). Security is also of vital concern when deploying a SAN.

2.0 Geographically Distributed SANs: Design Considerations

The decisions associated with deploying a geographically distributed SAN are very similar to those made when installing a SAN in a well-defined, local environment. Numerous design points need to be addressed including:

- Transport technology
- Performance – functional and transfer characteristics
- Client Heterogeneity
- Data sharing requirements

- Security policies
- Reliability, Maintainability and Availability
- Management and administrative policies
- Budget constraints

As distance increases, the following considerations gain importance:

- Increased latency between SAN elements as a function of actual data network/data routing.
- Impacts due to protocol processing and/or conversion.

The challenge of building a geographically distributed SAN equates to the task of extending 'the reach' of SCSI while maintaining architectural integrity. In its native form, FC supports connections of over six miles and even over sixty miles using specialized GigaBit Interface Converters (gbics) in the transmission link [1]. The use of dedicated links is assumed, which in most cases is either impractical or prohibitively expensive or both. An attractive alternative is to leverage the wide availability of in-place IP-based networks for data transmission between computer clients and storage that may not be in the same location. Competing but complementary FC-IP products are coming to market which accomplish this, namely Internet SCSI (iSCSI), FC Over IP (FCIP) and the Internet FC Protocol (iFCP).

iSCSI is a routing technology that not only leverages IP but also client-resident network interface cards (NIC) as well. Hosts participating in an FC SAN run an iSCSI driver on top the normal TCP/IP stack while a targeted iSCSI router located somewhere on the network converts block level commands issued through the NIC back and forth to FC. The hosts see Disks/LUNs as locally attached devices. iSCSI routers typically allow mapping through selected disks and/or LUNs.

FCIP is a bridging technology that connects SAN islands at the FC switch E-port level. The result is a single, logical SAN with clients and storage behaving as if they were local to one another irrespective of distance. Implementation requires equipment at both ends of the wire for conversion between the protocols.

IFCP lies somewhat in between iSCSI and FCIP. It is a gateway-to-gateway protocol that works with and/or eliminates the need for FC fabrics.

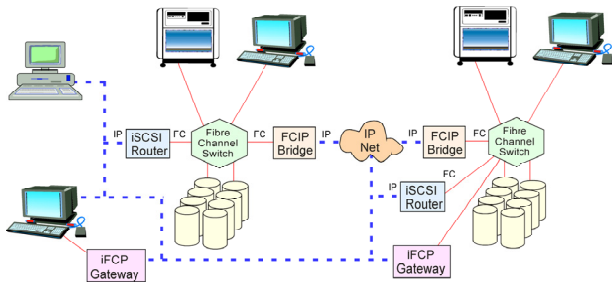Tom Clark [2] provides a more complete discussion.

Figure 2 – Alternative SAN Architectures

3.0 GSFC On-Campus

GSFC's SAN technology evaluation initially focused on standard FC implementations, much like that described in section 1.0. Testing later expanded to include iSCSI and FCIP technologies as well as traditional file-serving alternatives such as NAS. This section describes 'on-campus' pilot SAN activities involving equipment distributed primarily across three buildings at GSFC (figure 3).

The pilot SAN is a loose confederation of hardware and software structured, and restructured, with equipment added and removed as required to accommodate different evaluation objectives. The pilot has remained largely a 1 Gb/sec topology centered on the Brocade 2xxx family of FC switches and the Qlogic 22xx family of HBAs. Recently, some 2Gigabit/sec components have replaced their older counterparts. The Cisco SN5420 is currently deployed in the iSCSI role but a Technomages, Inc. DTP2000 router has also been tested. A DataDirect S2A6000 RAID performs most of the storage duties. However, an ADIC Scalar 100/LTO robotic tape library is also in the mix and an EMC Symmetrix storage unit was once part of it. Linux is the predominant operating system for the host computers, but Solaris, IRIX™ and Windows® equipment have also been available when needed.
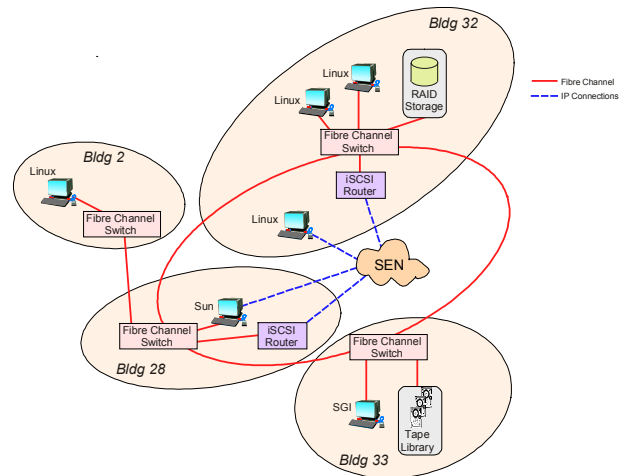

Figure 3 – GSFC Pilot SAN

3.1 GSFC Campus – Generic Benchmarks

Initial benchmarks were designed to exercise the core infrastructure of the GSFC pilot SAN using a representative product suite. *lmdd* [3]*,* a simple benchmarking tool that is supported by multiple operating systems, provided a quick assessment of performance of the various links and SAN components – FC and iSCSI. Test runs stressed large file (multiple GB) transfers and showed transfer rates primarily in the 50 to 80MB/sec range for FC connections and the 20 to 45 MB/sec range for iSCSI connections. The numbers varied as a function of client type, underlying client hardware architecture and storage configuration. iSCSI tests used a nominal MTU size of 1500. Other benchmarks biased more towards small files and metadata operations were run to gain an overall appreciation of the technology. These included *bonnie++* [4] and *postmark* [5]. For the most part, testing was performed using 'out of the box' settings. The following graphs provide comparative *lmdd* data for write and read operations using both an FC connected Linux client and an iSCSI connected Linux client. The results are stated in terms of bandwidth as a function of block size with file size held constant at 3 GB, large enough to avoid caching effects. The numbers include tests using native file system as well as the CVFS and GFS shared file systems.

Note that this data should be viewed as representative rather than definitive, since no major efforts were made at tuning. The numbers, however, are important in that they illustrate that although native FC outperforms iSCSI, the performance of iSCSI is encouraging, especially given the ease of implementation and overall cost.
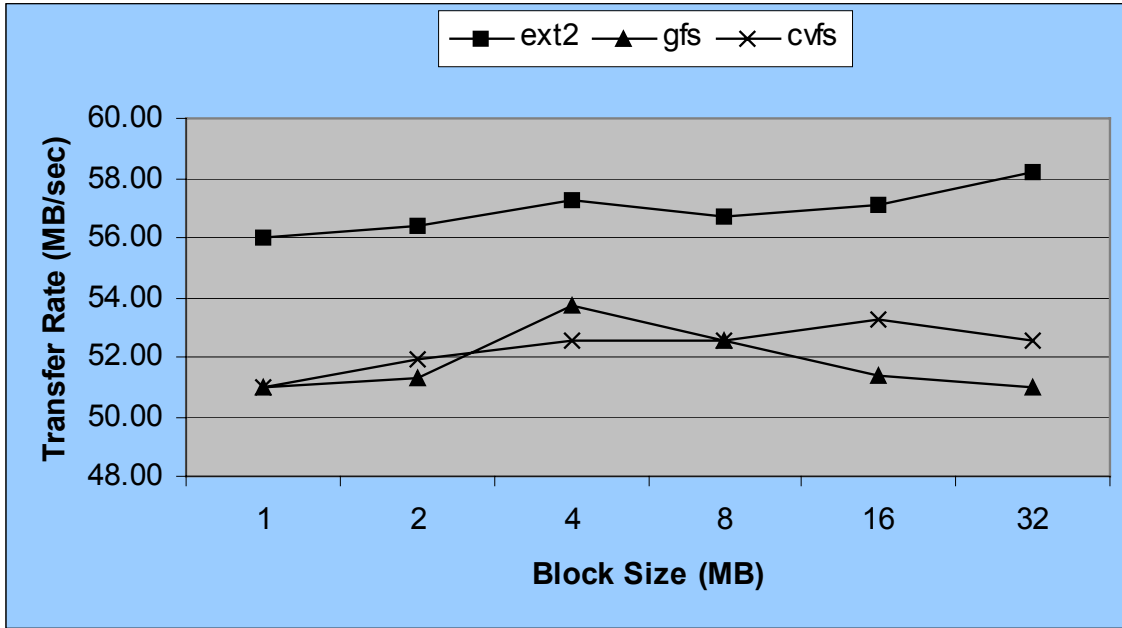
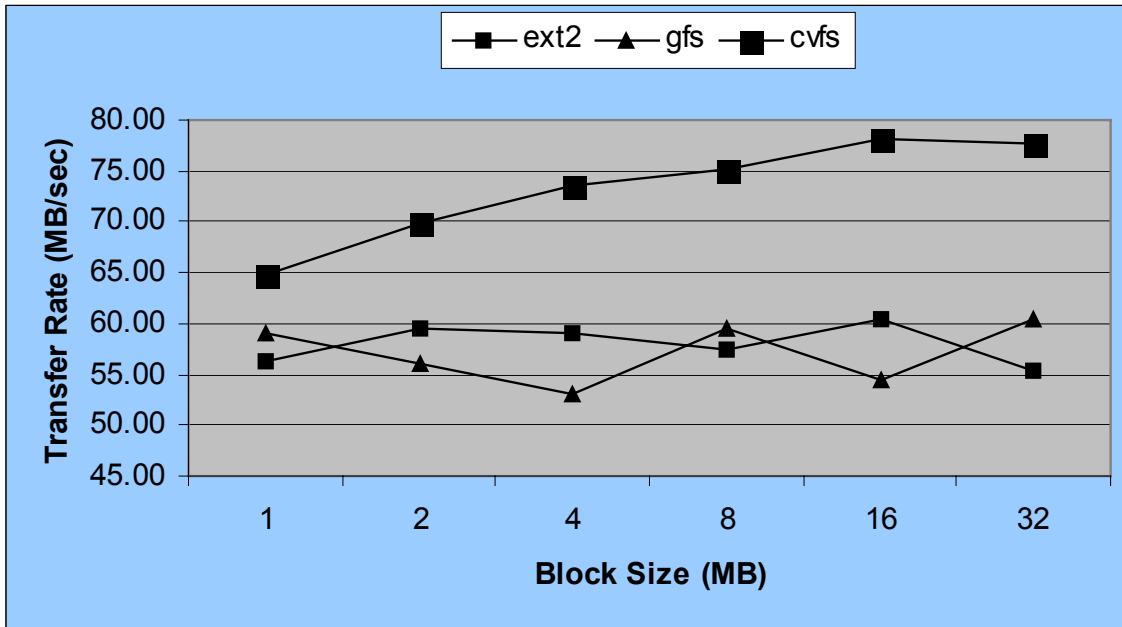**Figure 4 – FC Connected Linux Client: Writes**

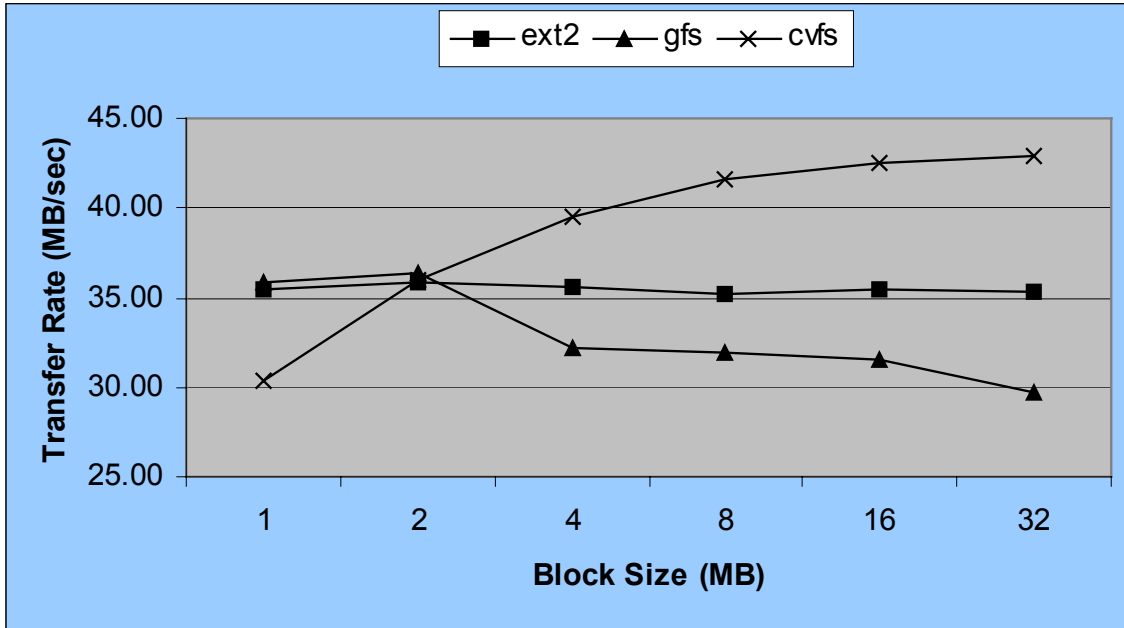**Figure 5 – FC Connected Linux Client: Reads**

Figure 6 – iSCSI Connected Linux Client: Writes
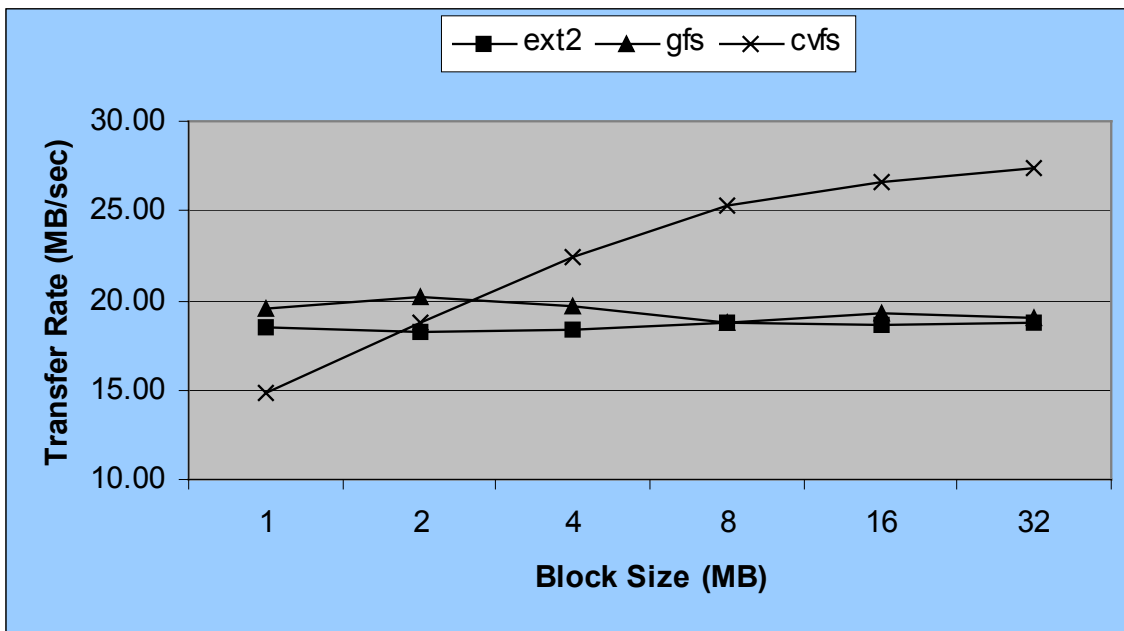
Figure 7 – iSCSI Connected Linux Client: Reads

A potential enhancement to an iSCSI implementation is the use of a TCP offload engine (TOE) card in place of a standard NIC. As the name implies, a TOE shifts some or all of the protocol processing to card-resident silicon thereby removing the burden from the host resident CPU. Several manufactures have such offerings. An early-implementation TOE card was tested in a Windows environment. Results show a definite CPU

load saving, but at the expense of bandwidth. This may be peculiar to the particular implementation. It must be noted that TOE cards are considerably more expensive than NICs, or just software.
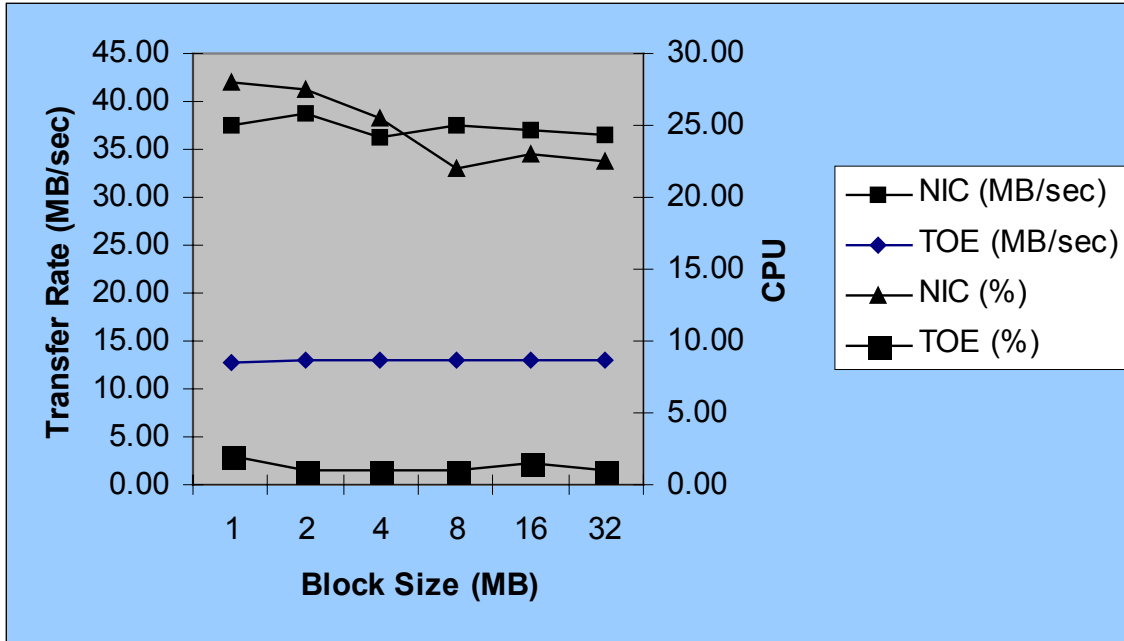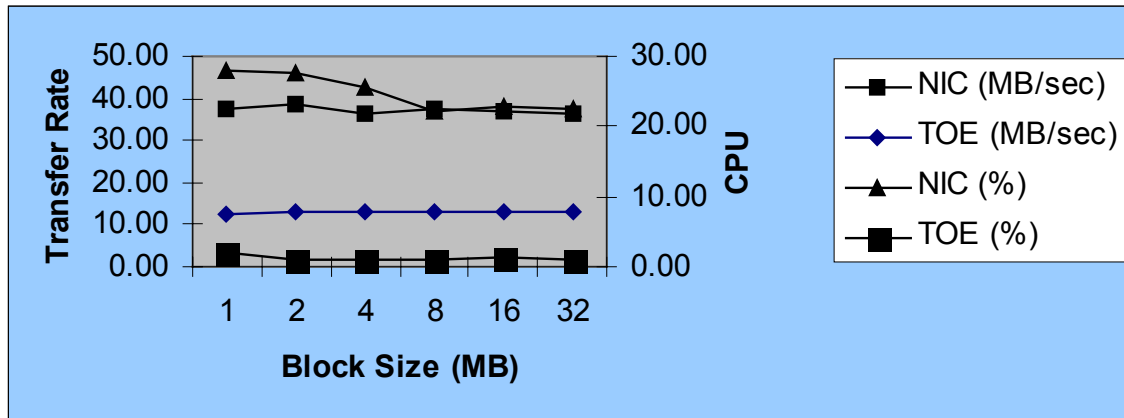
Figure 8– TOE Card: Writes

Figure 9 – TOE Card: Reads

bonnie++ and *postmark* benchmarks characterize differences in the file systems more than the underlying transport technology – FC versus iSCSI. CVFS centralized metadata imposes a performance penalty compared with the more distributed metadata approach of GFS.

Number of Files = 10:120:80/44
Chunk Size = 2GB

		Sequential Output						Sequential Input				Random Seeks		Sequential Create						Random Create					
		Per Char		Block		Rewrite		Per Char		Block				Create		Read		Delete		Create		Read		Delete	
system	file system	KB/sec	%CPU	KB/sec	%CPU	KB/sec	%CPU	KB/sec	%CPU	KB/sec	%CPU	/sec	%CPU	/sec	%CPU	/sec	%CPU	/sec	%CPU	/sec	%CPU	/sec	%CPU	/sec	%CPU
FC Client	ext2fs	9428	99	60735	52	3225	32	9012	98	70047	36	6716.5	45	17420	98	+++++	+++	+++++	+++	17621	99	+++++	+++	+++++	+++
FC Client	gfs	8145	97	40740	65	31937	49	8893	98	64626	39	494.6	4	1330	35	+++++	+++	938	55	1230	71	18745	98	952	53
FC Client	cvfs	7584	86	32691	48	4848	32	7647	87	40208	38	260.5	15	24	1	192	17	25	0	24	2	205	18	48	0
iSCSI Client	ext2fs	7042	99	41860	42	11340	18	6353	97	16164	12	385.2	4	11231	100	+++++	+++	+++++	+++	12182	99	+++++	+++	+++++	+++
iSCSI Client	gfs	5640	97	25467	63	13737	34	6206	95	20683	19	1263.4	19	818	35	15870	100	767	52	437	23	12569	100	731	52
iSCSI Client	cvfs	4289	68	10007	28	2114	23	3752	62	7967	19	237.5	22	23	2	146	16	25	0	23	2	98	12	24	1

Figure 10 – bonnie++ Results

FC Client

			Time			Files																Data			
File Range	Block Size	FS	Total	Trans	Trans /sec	Crtd	Crtd /sec	Crtd Alone /sec	Mixed w/trans /sec	Read	Read /sec	App	App /sec	Del	Del /sec	Del Alone /sec	Mixed w/trans /sec	MB read	MB Read /sec	MB Written	MB Writtten /sec
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---
16K - 1M	512	ext2	6	2	250	730	121	500	166	230	115	252	126	248	124	730	121	460	460	270	135	139.76	23.29	423.93	70.66
16K - 1M	512	gfs	19	8	62	730	38	500	83	230	28	252	31	248	31	730	38	460	92	270	33	139.76	7.36	423.93	22.31
16K - 1M	512	cvfs	90	37	13	730	8	500	13	230	6	252	6	248	6	730	8	460	27	270	7	139.76	1.55	423.93	4.71
1M - 8M	4096	ext2	93	51	9	742	7	500	12	242	4	252	4	248	4	742	7	484	484	258	5	1210.73	13.02	3604.88	38.76
1M - 8M	4096	gfs	147	75	6	742	5	500	7	242	3	252	3	248	3	742	5	484	121	258	3	1210.73	8.24	3604.88	24.52
1M - 8M	4096	cvfs	318	176	2	742	2	500	4	242	1	252	1	248	1	742	2	484	25	258	1	1210.73	3.81	3604.88	11.34

Figure 11 – FC Client postmark Results

3.2 Ozone Monitoring Instrument System

At GSFC, the Atmospheric Chemistry and Dynamics Branch is utilizing the inter-building SAN to share access to high-speed FC disk located in building 32 (figure 12). Scientists in building 33 are developing algorithms and studying datasets from the Total Ozone Mapping Spectrometer (TOMS) instruments and the Solar Backscatter Ultraviolet (SBUV) instruments. As the datasets are reprocessed by a computational system in building 32, the products are pushed onto the shared disk so they will be immediately available to scientists evaluating them in building 33. This shared disk will also support the adaptation of the algorithms for the Ozone Monitoring Instrument (OMI), which will be launched on the EOS Aura spacecraft in 2004. The implementation is FC and includes CVFS to allow multiple hosts simultaneous access to the various scientific datasets. Secondary clients are NFS-mounted off directly mounted CVFS clients.

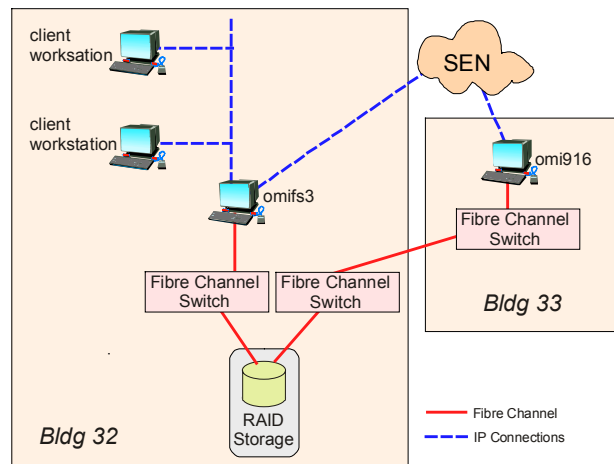The OMI configuration is currently operational. Tests are in progress.

Figure 12 – OMI System Diagram

4.0 GSFC Off-Campus Testing

The pilot SAN, although distributed across a campus, does not present a distance challenge to the SAN technologies described in earlier sections. The longest on-campus run is under a mile. The following sections describe tests conducted with facilities outside the boundaries of GSFC to better characterize the functional and operational issues with geographically distributed portions of a SAN.

4.1 University of Maryland

In cooperation with the University of Maryland Institute for Advanced Computer Studies (UMIACS), GSFC performed distance testing using iSCSI technology.

With a Linux client located on the Maryland campus, some six miles away, the client successfully mounted, wrote and read data using storage located in building 32 on the GSFC campus (figure 13). This includes using not only the native ext2 file system but CVFS and GFS as well. In the case of CVFS the metadata function (Files System Services or FSS) was located in building 32 as well. The same was true for the GFS lock manager. Both CVFS and GFS need a low bandwidth IP channel to communicate with a central function (FSS and Lock Manager, respectively).
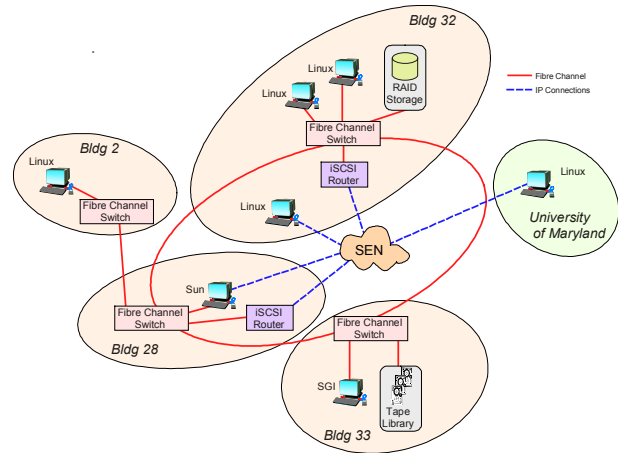

Figure 13 – SAN With UMD Connection

The benchmarks described in section 3 were run on the client at the University of Maryland (UMD). Transfer rates fell, but performance is still acceptable for a remote connection. With tuning, jumbo frames (MTU = 9000), etc. better numbers may be expected.
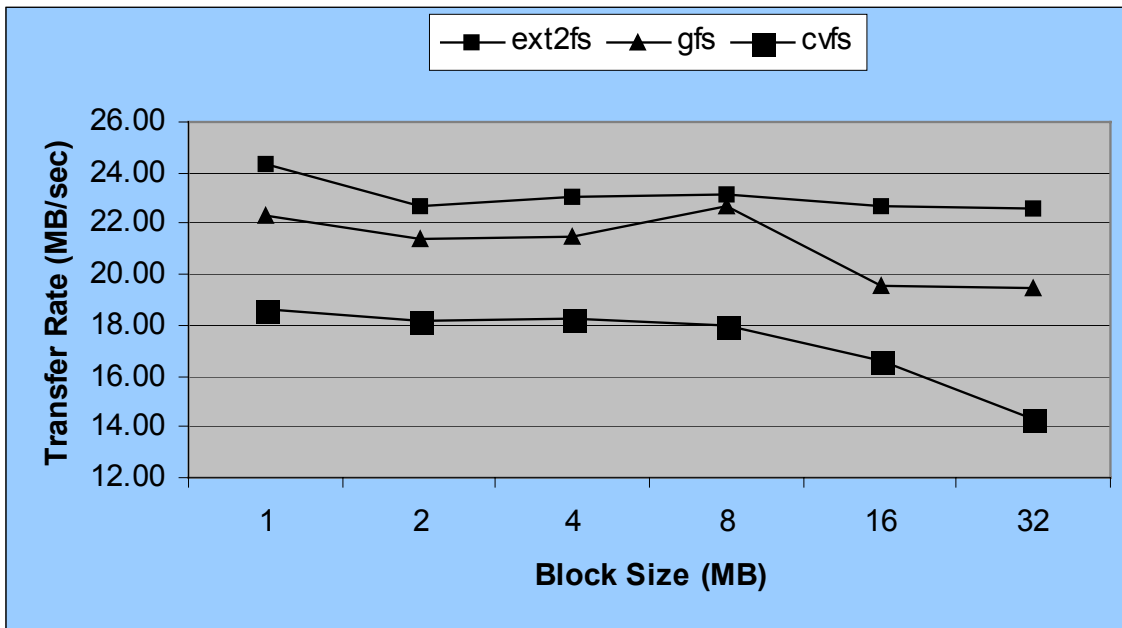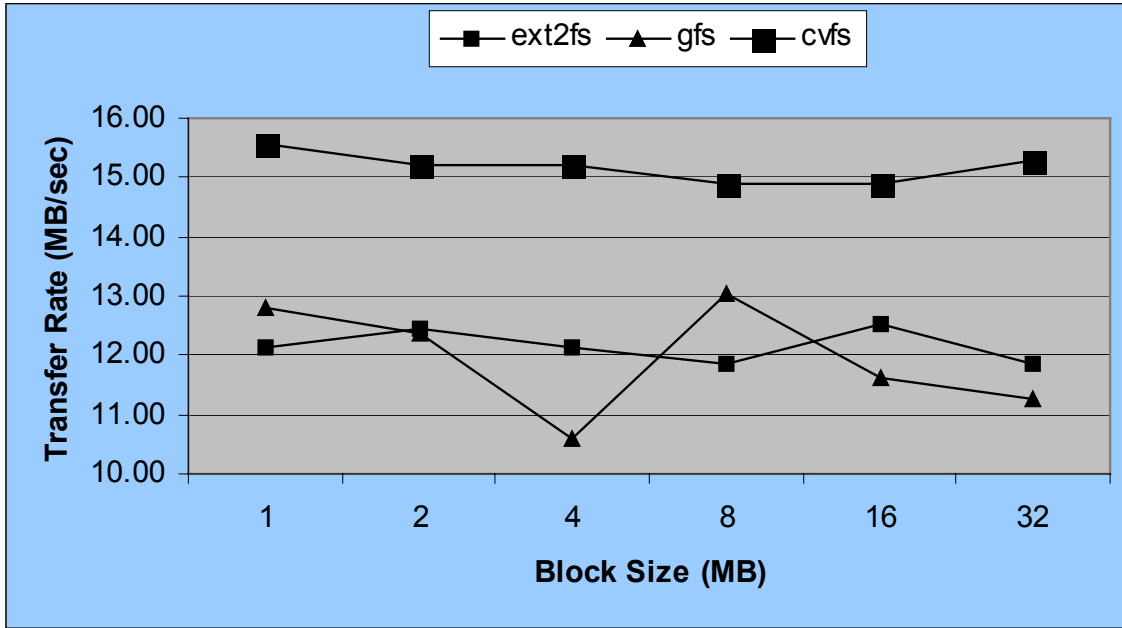

Figure 14 – UMD Client: Writes

Figure 15 – UMD Client: Reads

Number of Files = 10:120:80/44
Chunk Size = 2GB

system	file system	Sequential Output						Sequential Input				Random Seeks		Sequential Create						Random Create					
		Per Char		Block		Rewrite		Per Char		Block				Create		Read		Delete		Create		Read		Delete	
		KB/sec	%CPU	KB/sec	%CPU	KB/sec	%CPU	KB/sec	%CPU	KB/sec	%CPU	/sec	%CPU	/sec	%CPU	/sec	%CPU	/sec	%CPU	/sec	%CPU	/sec	%CPU	/sec	%CPU
UMD Client	ext2fs	11179	58	20761	9	6345	3	10664	65	8678	3	91.1	0	+++++	+++	+++++	+++	15208	13	+++++	+++	+++++	+++	+++++	+++
UMD Client	gfs	10839	64	23343	20	8680	7	11368	70	13551	5	339.5	2	369	6	+++++	+++	283	4	690	11	+++++	+++	281	4
UMD Client	cvfs	3277	19	2762	2	753	3	3226	19	7618	7	188.8	7	18	1	82	3	24	0	18	0	74	3	24	0

Figure 16 – UMD Client: bonnie++

UMD Client

File Range	Block Size	FS	Time			Files																	Data			
			Total	Trans	Trans /sec	Crtd	Crtd /sec	Crtd Alone	Crtd Alone /sec	Mixed w/trans	Mixed w/trans /sec	Read	Read /sec	App	App /sec	Del	Del /sec	Del Alone	Del Alone /sec	Mixed w/trans	Mixed w/trans /sec	MB read	MB Read /sec	MB Written	MB Writtten /sec
16K - 1M	512	ext2	3	1	500	730	243	500	250	230	230	252	252	248	248	730	243	460	460	270	270	139.76	46.59	423.93	141.31
16K - 1M	512	gfs	42	15	33	730	17	500	83	230	15	252	16	248	16	730	17	460	21	270	18	139.76	3.33	423.93	10.09
16K - 1M	512	cvfs	204	92	5	730	3	500	5	230	2	252	2	248	2	730	3	460	27	270	2	139.76	702(K)	423.93	2.08
1M - 8M	4096	ext2	238	140	3	742	3	500	5	242	1	252	1	248	1	742	3	484	484	258	1	1210.73	5.09	3604.88	15.15
1M - 8M	4096	gfs	283	154	3	742	2	500	4	242	1	252	1	248	1	742	2	484	26	258	1	1210.73	4.28	3604.88	12.74
1M - 8M	4096	cvfs	1126	505	0	500	0	500	0	242	0	252	0	248	0	742	0	484	25	258	0	1210.73	1.08	3604.88	3.2

Figure 17 – UMD Client: postmark

The operational value of such connectivity to UMIACS is access to a shared data repository. Moderate Resolution Imaging Spectroradiometer (MODIS) data could be centrally stored with direct connectivity provided to geographically distributed researchers. Such access would save cumbersome FTP data movement as well as redundant storage. In that context, the UMIACS team evaluated the Maryland-to-GSFC link using the MOD44C application, code that takes daily MODIS data and produces 16-day composites.

The following times were noted for a typical MOD44C run using three different types of file access:

Local disk	105 min + 60 min for ftp
CVFS	195 min + 0
GFS	128 min + 0

In the first case, local disk, the preliminary ftp transfer of the required files took an addition hour. The difference between CVFS and GFS is likely attributable to metadata handling.

4.2.2   Gilmore Creek AK

GSFC is currently testing IP technology for the transfer of satellite data from Gilmore Creek Alaska Ground Station Facility to the Level Zero Processing Facility (LZPF) in Greenbelt MD (figure 18). The objective is to create a shared file system for moving data using CVFS, with shared storage in Alaska and file system clients in both Alaska and GSFC in Maryland. FCIP equipment will permit direct fibre channel connection to the storage from the two geographically distributed hosts. Referring to the figure, both SGI 2 and SGI 3 will mount the same file system resident on the shared storage. Subsequently, SGI 2 will write data into the shared storage and SGI 3 will read it out.

This long-distance connection is being accomplished in steps which include configurations completely local to GSFC using simulated circuits and delays and loop-back tests involving partners in California.
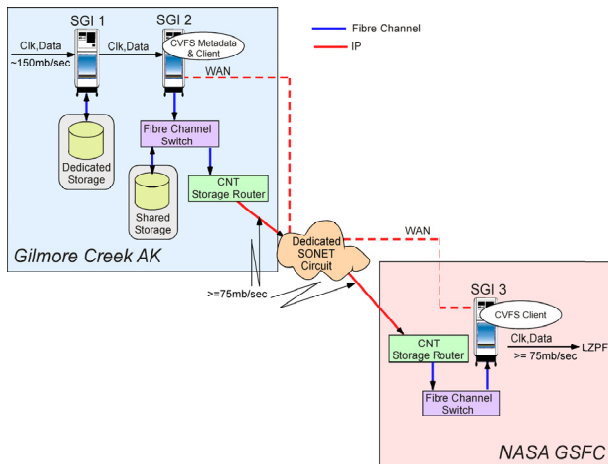

Figure 18 – Alaska-to-GSFC FCIP Configuration

Summary
The tests conducted indicate that iSCSI is a promising, lower-cost alternative to FC. The ability to use shared file systems, and thus operate in a heterogeneous environmen,t is a definite advantage; our tests, however, did not tune the file system parameters for optimum performance. Such tuning, and extending the reach of both FC and iSCSI to over 5000 km, is being pursued.

References

[1]http://networking.smsu.edu/general/info/CiscoGBIC.htm

[2] Tom Clark. IP SANs A Guide to ISCSI, iFCP, and FCIP Protocols for Storage Area Networks., Addison-Wesley, 2002.

[3] http://www.bitmover.com/lmbench/lmdd.8.html

[4] http://www.coker.com.au/bonnie++/

[5] http://www.netapp.com/tech_library/3022.html