SAN and Data Transport Technology Evaluation at the NASA Goddard Space Flight Center (GSFC)

Hoot Thompson
Patuxent Technology Partners, LLC
11030 Clara Barton Drive
Fairfax Station, VA 22039-1410
Tel: +1-703-250-3754, Fax: +1-703-250-3742
e-mail: hoot@ptpnow.com

Abstract
Growing data stockpiles and storage consolidation continue to be the trend. So does the need to provide secure yet unconstrained, high bandwidth access to such repositories by geographically distributed users. Conventional data management approaches, both at the local and wide area level, are viewed as potentially inadequate to meet these challenges. This paper explores methods deploying a new breed of Fibre Channel (FC) technology that leverages Internet Protocol (IP) infrastructures as the data transport mechanism, a step towards creating a "storage area network (SAN) grid". These technologies include products using the FC Over IP (FCIP) and the Internet FC Protocol (iFCP) protocols. The effort draws upon earlier work that concentrated on standard FC and internet SCSI (iSCSI) technologies. In summary, the vendor offerings tested performed as expected and provided encouraging performance results. However, their operational readiness still needs to be understood and demonstrated. Installing and configuring the products was reminiscent of the early days of FC with driver and version compatibly issues surfacing once again. Maturity will take some time.

1. Introduction

GSFC, as part of a continuing technology evaluation effort, continues its interest in SAN products and related technologies by evaluating and demonstrating the operational viability of new vendor offerings. Under the auspices of the SAN Pilot, earlier testing has shown the advantages of high-speed transport mechanisms such as FC as well as the flexibility that iSCSI provides in deploying a SAN [1]. Subsequent testing is building upon this work, emphasizing higher speed campus backbones with a focus on manageability as well connectivity to geographically distributed sites. Standardized benchmarks provide measurement of inherent link throughput. In addition, the push is on to attract users with real applications that could benefit from these kinds of technologies

The vision is direct access to data regardless of geographical location, using IP based wide area networks (WAN) as the transport mechanism. Such distributed storage, whether for disaster preparedness or for logical proximity to a compute server, pushes the operational requirements normally associated with direct-attached storage onto the WAN. The storage will be expected to be both reliable and high performance, and to behave like direct attached and physically local. The vision promotes leaving data static and performing the necessary processing directly on a data store as opposed to moving large quantities of data between user facilities. Connections would be temporal in nature with a corresponding service, such as the Storage Resource Broker (SRB) [2], to assist users in

locating relevant data. The end result would be a SAN grid, analogous in many ways to more traditional grids currently gaining wide exposure. This paper explores a variety of topics seen as contributing to the vision.

2. SAN Pilot Infrastructure Description

The core of the SAN Pilot (figure 1) is the connectivity between multiple, on-campus buildings at GSFC. Traditional FC dominates the local GSFC infrastructure with a mix of 2 Gigabit/sec and 1 Gigabit/sec switches – Brocade 3800s and 2400s – providing ports for a variety of server and storage technologies. Linux, Solaris and Apple hosts are represented. RAID storage systems include a DataDirect Networks S2A6000, an Apple Xserve, an Adaptec/Eurologic SANbloc and a Nexsan ATABoy2. A pair of Nishan IPS 3000 Series Multiprotocol IP Storage Switches as well as a LightSand I-8100 augment the other switches by bridging the FC fabric to the IP network. A pair of legacy Cisco SN5420s used for iSCSI work completes the topology. The equipment is mostly GSFC owned. However, notable exceptions include the Nishan and LightSand IP switches. Cisco, Brocade and ADIC have also provided loaner equipment during the testing.

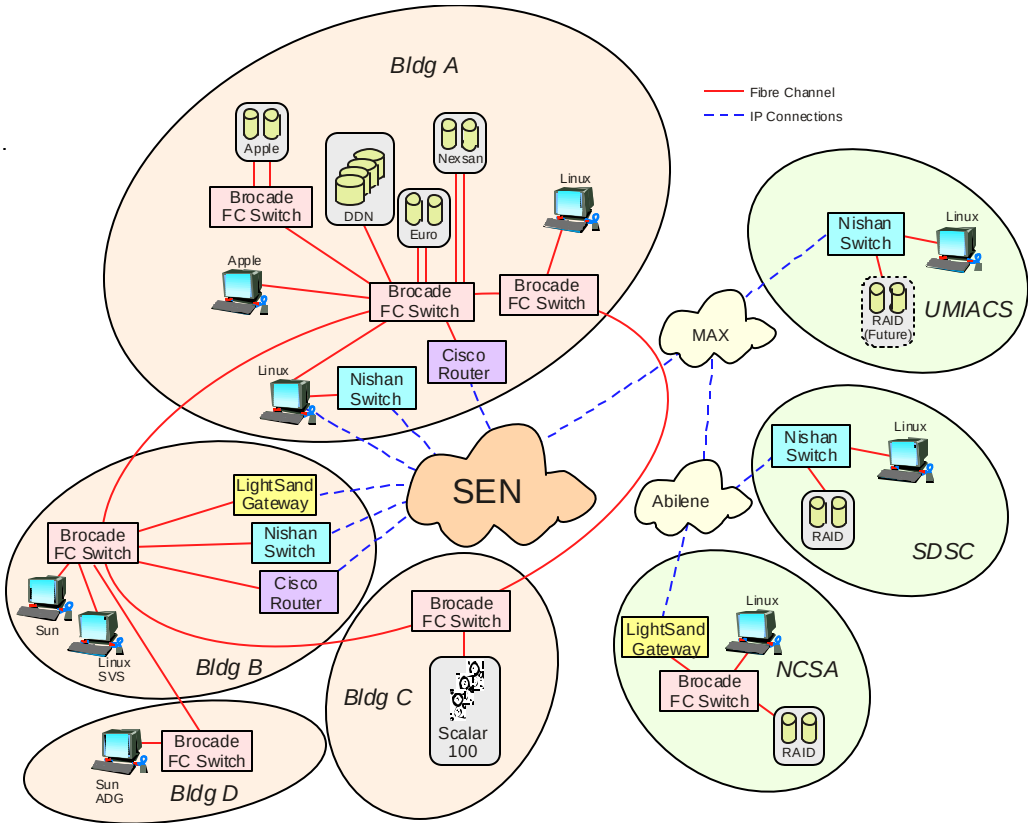

Figure 1 - SAN Pilot Infrastructure

The Nishan and LightSand equipment provide IP connections to similar boxes at the University of Maryland Institute for Advanced Computer Studies (UMIACS), the San Diego Supercomputer Center (SDSC) and the National Center for Supercomputing Applications (NCSA). The underlying networks have been key to the IP related testing. Local to GSFC, the primary backbone is the Science and Engineering Network (SEN)

[3]. Connection to UMIACS is attained by the Mid-Atlantic Crossroads (MAX) [4]. MAX is also the jump off point to the Abilene Network [5] that completes the circuit to both NCSA and SDSC. The result is full Gigabit Ethernet (GE) to all of the remote sites.

2.1. SEN Network

The SEN is a local, non-mission dedicated computer network with high-speed links to the Internet2's Abilene and other Next Generation Internet (NGI) networks. It serves GSFC projects/users who have computer network performance requirements greater than those allocated to the general-use, campus-wide Center Network Environment. The majority of the SEN's inter-building backbone links are 4 gigabits per second (Gbps), created using IEEE 802.3ad link aggregation standards with four separate GE connections between respective pairs of switches. For desktop workstations and servers, as well as for its other inter-building and intra-building links, the SEN minimally provides GE LAN connections. Only jumbo frame-capable GE switches are used in the SEN's infrastructure. The 9000-byte sized Ethernet jumbo frames (maximum transmission unit or MTU) generally provide individual users with approximately six times better throughput performance as compared to networks only supporting standard 1500 MTUs. The SEN presently supports a 2 Gbps jumbo frame-capable link with the MAX point-of-presence at the University of Maryland College Park.

2.2. MAX Network

The MAX is a multi-state metaPoP consortium founded by Georgetown University, George Washington University, the University of Maryland, and Virginia Polytechnic Institute and State University. The proximity of the MAX to Washington, D.C. places it in an advantageous location to partner with federal agencies as well as the business community and post-secondary institutions of DC, Maryland and Virginia. MAX represents a pioneering effort in advanced networking, with the potential to rapidly incorporate a broad cross-section of the not-for-profit community. The MAX, the regional gigapop for access to the Abilene network and the NGI-East Exchange, provides the SEN with excellent WAN connectivity.

2.3. Abilene Network

The Abilene Network is an Internet2 high-performance backbone network that enables the development of advanced Internet applications and the deployment of leading-edge network services to Internet2 universities and research labs across the country. The network supports the development of applications such as virtual laboratories, digital libraries, distance education and tele-immersion, as well as the advanced networking capabilities that are the focus of Internet2. Abilene complements and peers with other high-performance research networks in the U.S. and internationally. The current network is primarily an OC-192c (10 Gbps) backbone employing optical transport technology and advanced high-performance routers.

3. FCIP and iFCP Technology

Prior testing focused on standard FC and iSCSI technologies as it applied to on-campus connections and/or short distances. Interest shifted to assessing the feasibility of constructing a geographically distributed SAN system. This led to experimenting with

more suitable technologies, namely FCIP and iFCP. Several products are available that exploit these protocols. The two tested extensively were the IPS 3000 Series IP Storage Switch by Nishan Systems, now a part of the McData Corporation, and the i-8100 unit by LightSand Communications, Inc. The following paragraphs give a brief overview of each of the products and summarize the current evaluation status.

3.1. Nishan IPS 3000 Series IP Storage Switch

The IPS 3000 and 4000 Series IP Storage Switches use standards-based IP and GE for storage fabric connectivity. Nishan's Multiprotocol Switch supports iSCSI, iFCP, and E_Port for trunking to both IP backbones and legacy FC fabrics. The IPS 3000 Series connects to a wide variety of end systems, including FC, NAS, and iSCSI initiators and targets. The switch has a non-blocking architecture that supports Ethernet Layer 2 switching, IP Layer 3 switching and FC switching over extended distances at full Gigabit wire speed. The Series also supports standard IP routing protocols such as open shortest path first (OSPF) and distance-vector multicast routing protocol (DVMRP) and can be fully integrated into existing IP networks.

Three parameters assist in tuning the performance of the Nishan to a specific environment – Fast Write™ [6], compression [7] and MTU size. When servers and storage are interconnected via a WAN using a pair of Nishans, the normal SCSI exchange (figure 2) required for a 1MB file write will break the data into multiple transfers thereby compounding the "round trip time (rtt)" effect. In contrast, with Fast Write enabled, when the server sends the SCSI write command (figure 3) to set up the transfer, the local Nishan responds with a transfer ready specifying that the entire 1MB of data can be sent at once. At the same time, the sending Nishan forwards the SCSI write command across the WAN so that the target can be prepared to receive data. Having received the 1MB of data from the server, the sending Nishan streams the 1MB block across the WAN to the receiving Nishan. The receiving Nishan, in turn, mimics the normal command/response sequence for the transfers until all of the data is given to the target. The Nishans do not spoof write completion. Instead, the actual status generated by the storage target is passed back through the network to the server. This guarantees that all data was actually written to disk.

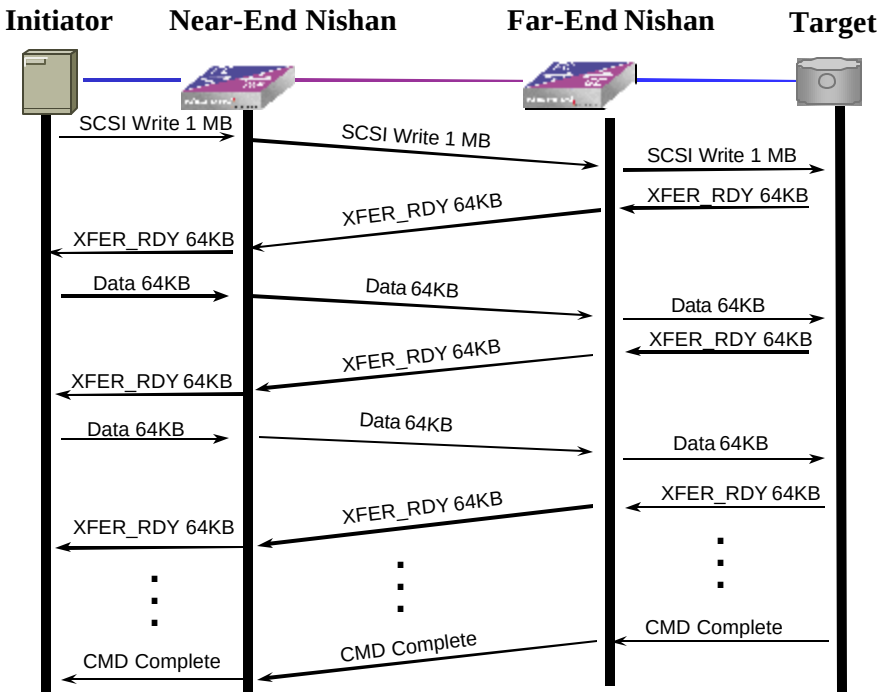

Figure 2 - Normal SCSI Exchange for a 1MB Write

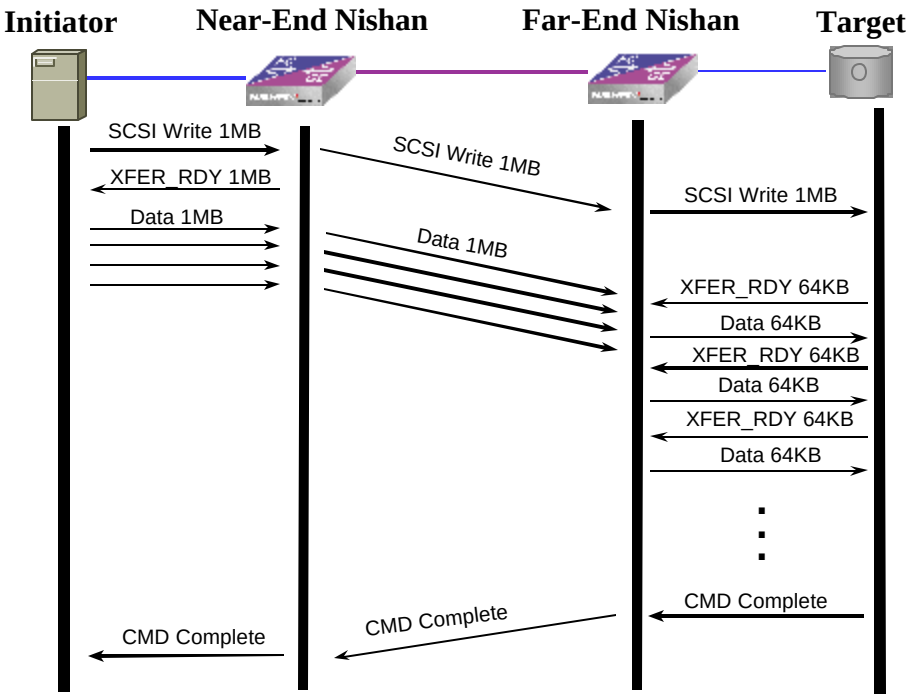

Figure 3 - Fast Write Modified SCSI Exchange

The Nishan switch also features software based *lossless* compression. The following options are available:

- **Off** - Data going out of the port is not compressed.
- **On** - Data going out of the port is always compressed using the appropriate algorithm to achieve maximum compression.
- **Auto** - Depending on the available bandwidth, the switch dynamically decides whether or not to compress the data, the level of compression to apply and the compression algorithm to use. With the Auto setting, the port keeps the data rate as close as possible to the Port Speed of the port.

The last key parameter is MTU. The Nishan switches can support packet sizes up to 4096 bytes, an increase of almost 3X over the nominal 1500. The larger data payload results in less header processing overhead and better link utilization. Packet sizes greater than 1500 bytes maximizes direct matching with FC originated frames. The full FC data payload of 2112 bytes can be delivered in a single jumbo, 4096 byte Ethernet frame. The "auto" option for MTU setting allows Nishan switches to negotiate the best possible rate.

Configuring the Nishan switch involves the interaction of two applications, the switch resident http GUI Element Manager and the host based (Linux or Solaris) SANvergence Manager application. Between the two, devices to be shared are placed in commonly seen, exported zones. The level of SAN merging is a cooperative effort between two or more switches. As a default, a CLI is also available.

3.2. LightSand i-8100

The LightSand i-8100A is an intelligent gateway that provides connectivity between FC fabrics across an IP WAN infrastructure. The i-8100A is an eight port, multi-protocol switch that provides isolation between FC SANs using Autonomous Region (AR) technology. Conventional FCIP bridging devices link two sites by merging the FC fabrics together. By maintaining Autonomous Regions, the i-8100A is able to share storage devices without merging fabrics. In the diagram (figure 4), two autonomous regions are joined. Each AR consists of four FC switches, the three original switches plus the gateway. If these two SANs had been bridged by a simple FCIP gateway (non-switching), the fabric would appear as six FC switches—all part of the same fabric. The storage arrays labeled Disk 1 and Disk 2 are shared. Once they have been imported into SAN 2, every initiator in SAN 2 can see the shared disks as if they were present in SAN 2. In reality, the I-8100A is performing Domain Address Translation (DAT) and the actual disks remain inside SAN 1. Because of this technology, each fabric is isolated from any disturbances that might occur in the other fabric.

The LightSand i-8100A employs the user datagram protocol (UDP) with an additional sequencing number to enable protection against packet-loss and mis-ordering. This protocol is referred to as UDP/SR (UDP with Selective Retransmission). Using UDP/SR, the i-8100A can be set for a desired WAN bandwidth. It will instantly jump to that bandwidth and execute appropriate backpressure against the FC fabric, if the WAN bandwidth is less than the native FC bandwidth. In the event that there is packet-loss on the WAN, the i-8100A will retransmit the lost data without throttling the bandwidth.

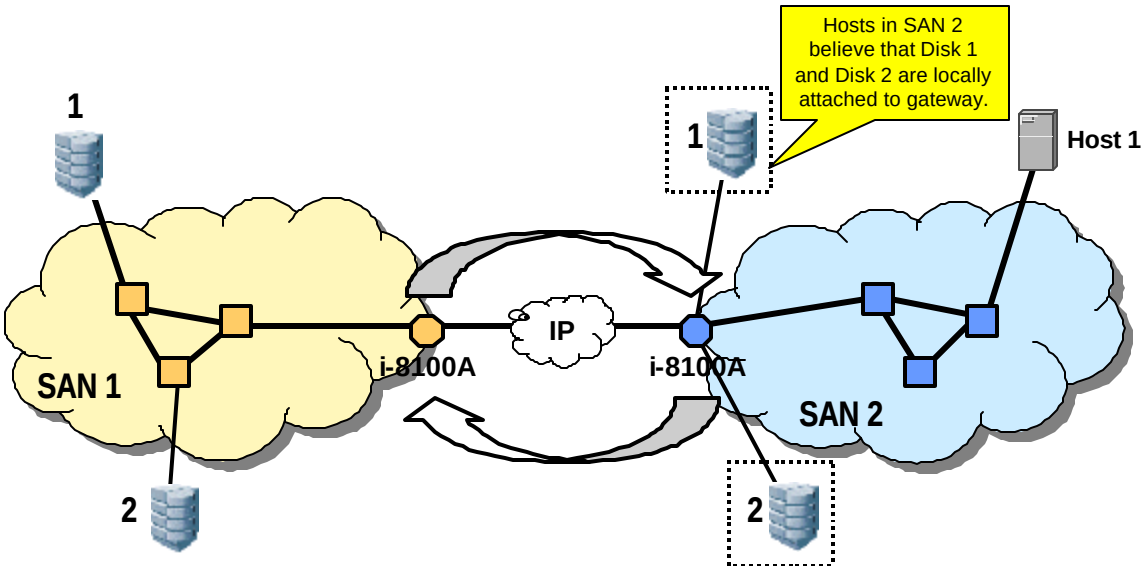

Figure 4 - LightSand Interconnect

Configuring the LightSand switch requires running the SANman GUI on each of switches or using the available CLI.

3.3. Evaluation Process and Results

As evidenced by the work done at SDSC for last year's Mass Storage conference [8], outstanding performance moving data over IP is achievable using a well-behaved, highly tuned network. The tact taken at GSFC has been more the "every day", out-of-the-box approach where nothing aggressive is done to enhance the performance of site-to-site WANS. In more real world networks, the effects of rtt, congestion and packet loss can render an application useless that requires high bandwidth. In the spirit of the SAN grid vision, laying a distributed file system, such as ADIC's StorNext File System (SNFS) or SGI's CXFS™, on the topology would further attenuate any irregularities.

FCIP and iFCP testing has been a multi-step process:

- Evaluate the technology on a local, campus basis under ideal network conditions.
- Artificially introduce non-zero rtts, packet loss and congestion into the circuit, and observe the impact on performance.
- Connect to a geographically distant center(s) and compare performance to predictions based on simulated distance testing.

Testing was performance centered using standard benchmarks such *lmdd* [9] and *IOzone* [10] as the primary tools. *lmdd* is good for quick, single threaded operations. *IOzone* permits a variety of IO operations including writes, reads, mixed writes and reads, multi-threaded operations, etc. all with options for setting attributes such as record and file size. The majority of the tests consisted of multiple *IOzone* operations described by the following script:

./iozone_mod -i 0 -i 1 [-+d] -r 1m -s 16g -b one_thread

./iozone_mod -t 2 -i 0 -i 1 [-+d] -r 1m -s 8g -b two_threads
./iozone_mod -t 4 -i 0 -i 1 [-+d] -r 1m -s 4g -b four_threads
./iozone_mod -t 8 -i 0 -i 1 [-+d] -r 1m -s 2g -b eight_threads

The scripts steps through 1, 2, 4 and 8 threaded write/read operations and in aggregate moves 16 Gbytes. *IOzone* was modified such that the [-+d] option would generate random data without doing the diagnostic byte-for-byte check of the data. This was done to evaluate the efficiency of the Nishan compression algorithm while not impacting performance with verification process. Tests were performed using mostly native file systems (ext2) with some minimal SNFS evaluation.

Network utilization was also monitored. Data traffic cannot be at the expense and disruption of existing communication traffic. At a minimum, the impact must be understood and anticipated. Nishan and LightSand use two different approaches to how the data is transported so the resulting network perturbation varies.

3.3.1. On-Campus Testing

Testing began at GSFC with a pair of Nishan switches. A Linux machine was FC connected to one of the Nishans co-located in the same building (figure 5). The other Nishan, in a different building provided tie-in to the SAN Pilot and its associated RAID. Initial results, with zero rtt, compared favorably with the same tests using directly connected RAID.

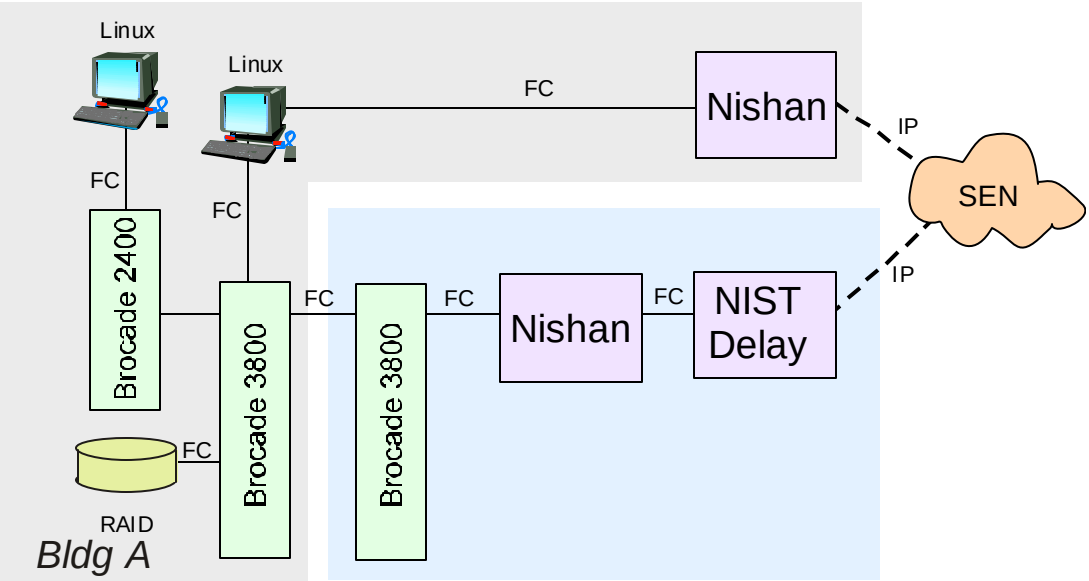

Figure 5 - Local GSFC Testing

The next step was to introduce set delays into the circuit using a NIST Net [11] network emulator to simulate the potential effects of geographically separating the two Nishan switches. The NIST Net network emulator is a general-purpose tool for emulating performance dynamics in IP networks. The tool is designed to allow controlled, reproducible experiments. By operating at the IP level, NIST Net can emulate the critical end-to-end performance characteristics imposed by various WAN situations (e.g.,

congestion loss) or by various underlying subnetwork technologies (e.g., asymmetric bandwidth situations of xDSL and cable modems).

Impressions

Installation and configuration of the Nishan units was relatively straightforward with the assistance of the product support engineers. Besides providing FC-IP translation, the Nishans are also full FC switches, an attribute that has different ramifications depending upon how the device is introduced into an existing SAN. As a standalone switch with directly connected devices, as was the case for one end of the GSFC circuit, operation was clear with only the usual zoning decisions to be made. The second switch was E-port connected, a more complicated configuration which requires choosing how the Nishan was to interoperate with the existing SAN Pilot Brocade infrastructure. Multiple options are available, so the ripple effect of zone changes, for example, need to be understood to avoid any unforeseen interruption of an operational SAN. Setting up the zones and mapping devices was easily accomplished using SANvergence and the Element Manager.

Large transfers (files) were required to overcome the buffering effects of the servers, the switches and the link. With *IOzone* modified accordingly, a variety of tests were executed varying rtt and MTU size while going through the permutations of the Fast Write and compression settings. Three observations were made:

- Fast Write seems to have an overall positive effect on write performance with this likely being the default setting. Nishan recommends setting to "on" for distances over 200km noting potential degradation if "on" for shorter distances.
- Compression can have a positive or negative effect depending upon rtt. Compression processing significantly reduces throughput when rtt is small. Conversely, for large rtt compression enhances performance. Nishan recommends the "auto" mode letting the switch dynamically determine the appropriate level of compression.
- The effect of increasing MTU size from 1500 to 4096 was somewhat inconclusive but an odd jump was noted when both FastWrite and compression were turned "off". Intuitively the larger frames should improve performance but the suspicion is that the effects of a large rtt on the SCSI exchange may mitigate the gain. This warrants further testing.

In summary, settings are situation dependent. This warrants exercising all the combinations before finalizing an installation. To illustrate the point, the following graphs (figure 6 and 7) depict bandwidth as a function of threads for rtt=35msec for different MTUs, Fast Write and compression settings. For MTU = 1500, the best write performance was for Fast Write, no compression while read was best for Fast Write with compression enabled. Bumping the MTU to 4096 resulted in both the write and read numbers being best with Fast Write and compression disabled. Incidentally, these parameters are changed using the Element Manager with each switch configured independently. The implication is that unpredictable results may occur if the switches are not configured the same. Overall, the write performance topped out at just slightly over

25 MB/sec while read approached 20MB/sec. For the most part, running multiple threads

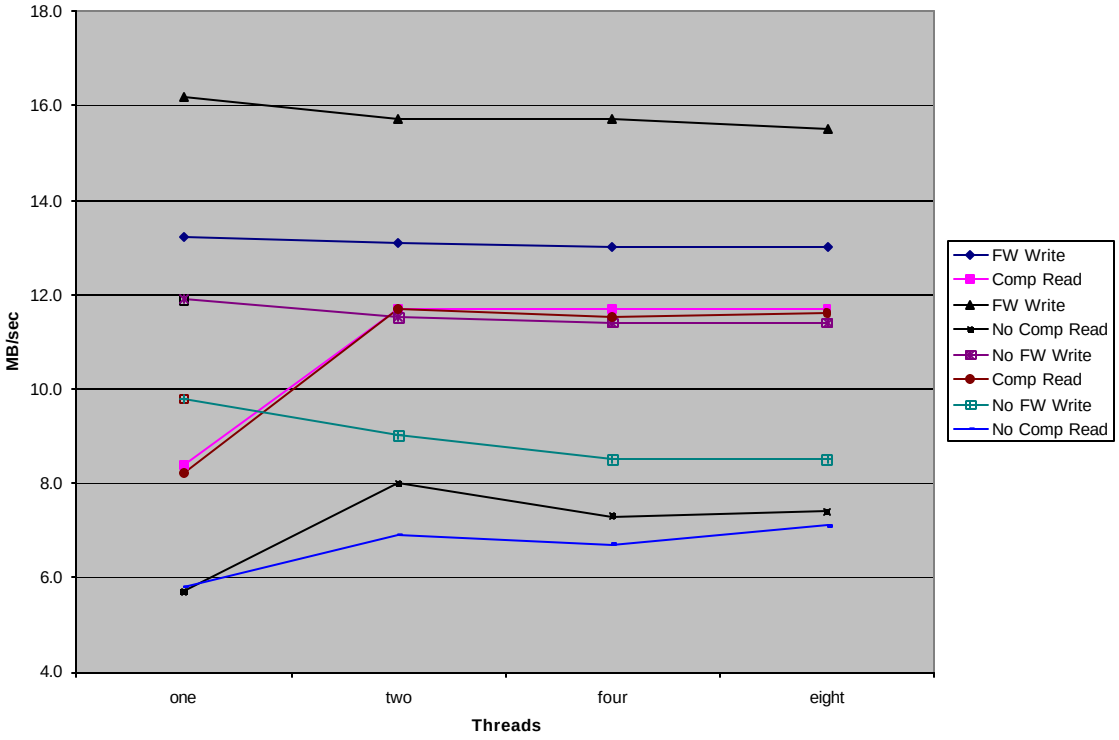

Figure 6 - Delay=35msec, MTU=1500

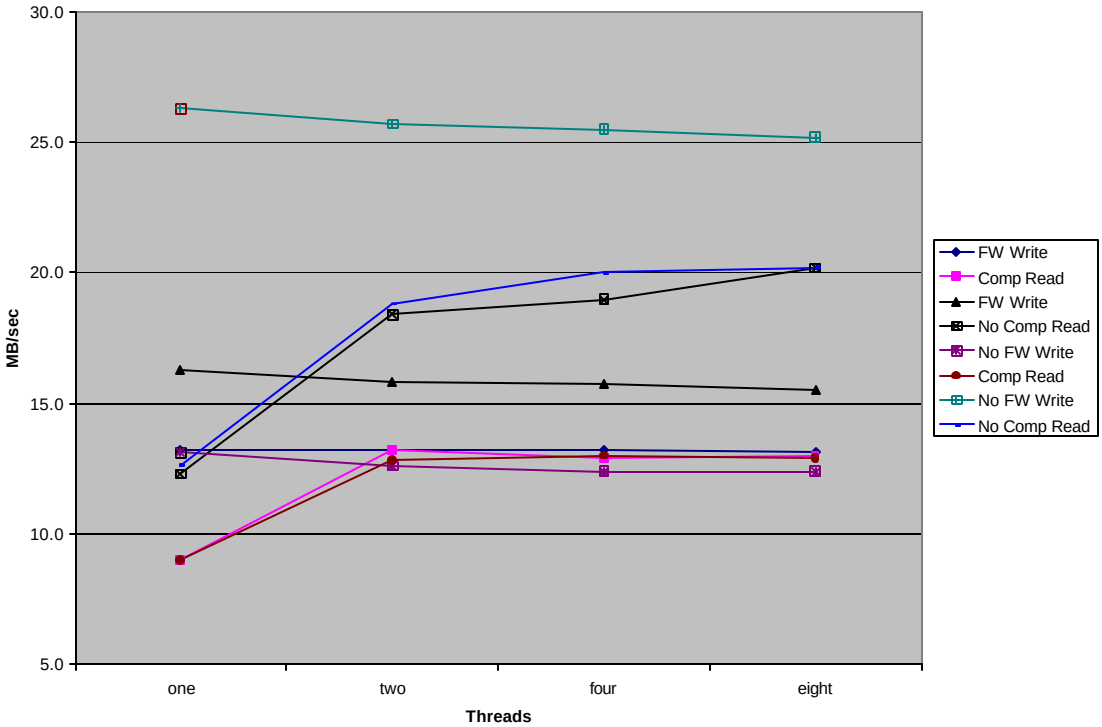

Figure 7 - Delay=35msec, MTU=4096

boosted aggregate throughput. These numbers are in contrast to 86 MB/sec writes and 78 MB/sec reads obtained running eight threads with rtt=0, MTU=1500 and both Fast Write and compression turned off.

Future Testing

Additional tests to be conducted include:

- Run tests with a broader range of rtt values while changing configuration of the Nishan units. This would give the full curve for bandwidth as a function of rtt.
- Test the compression "auto" setting in contrast to the "on/off" results.
- Induce deterministic packet loss and congestion, and measure the impact on write and read performance.

3.3.2. Multi-site Testing

The next series of tests involved different combinations of IP hardware and network connections to UMIACS, SDSC and NCSA. Experiments focused mainly on building and exercising native file systems (ext2) with server/host and storage at opposite ends of the WAN link. Some preliminary SNFS testing was also accomplished. In all cases, the assessment centered on:

- Gauging the impact of rtt or latency on performance in a real world setting where the network is potentially hostile.
- Comparing measured maximum network bandwidth, as determined using nuttcp, with file system oriented traffic.

3.3.2.1. UMIACS

Last year, UMIACS participated with GSFC in distance testing using iSCSI technology. That effort involved a Linux box at UMIACS routed through a Cisco SN5420 at GSFC to the associated storage assets. This time for comparison, one of the two loaner Nishan units was moved to UMIACS (figure 8). Nishan-to-Nishan communication was established using the MAX network. *IOzone* benchmarks were performed building a native ext2 file system on GSFC storage from an UMIACS resident Linux host.

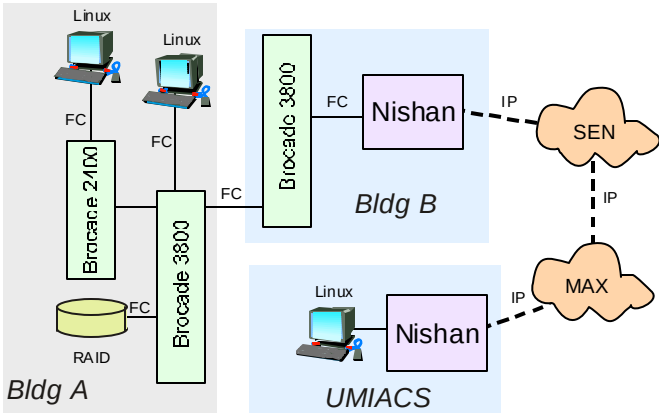

Figure 8 - GSFC - UMIACS Configuration

Impressions

Moving and establishing the Nishan to UMIACS connection was relatively simple. Network logistics provided the only significant obstacles. Getting the Nishan configuration tools functioning in a new environment posed a minor nuisance. Only certain browser/host combinations will run the Element Manager GUI. Secondly, UMD, except in specific instances, blocks SNMP which led to establishing a virtual private network (VPN) for remote access to both Element Manager and SANvergence.

Performing *IOzone* testing with random data yielded the following results (Table 1) for one, two, four and eight threaded operations. These results are for an MTU size of 1500 and a negligible rtt as registered by the Nishans.

Table 1 - Results

	FW, Comp		No FW, No Comp	
Threads	Write	Read	Write	Read
one	12.8	9.5	38.6	14.1
two	12.9	11.7	47.3	19.8
four	12.8	11.6	28.9	20.6
eight	12.8	11.6	59.8	25.8

Given the near zero rtt, the boxes ran best with both Fast Write and compression disabled. As noticed in other testing involving the Nishan, compression processing effectively halves the bandwidth in applications involving small rtts. The eight threaded write, 59.8 MB/sec, saturated the network given the available bandwidth, as measured by nuttcp [12], was 56.2MB/sec. Reads topped out at 25.8MB/sec. Single threaded *IOzone* tests saw 38.6MB/sec writes and 14.1MB/sec reads. As it turns out, the WAN connection at UMIACS end is not full GE but rather a fractional allocation of a full GE. By comparison to historical data, single threaded iSCSI operations using lmdd yielded 18MB writes and 12MB reads.

Future Testing

Additional tests to be conducted include:

- Increase network bandwidth between GSFC and UMIACS to a full GE and reevaluate Nishan performance. Given the almost negligible rtt, a significant performance jump is anticipated.
- Connect storage to the UMIACS Nishan then test reads and writes originating at GSFC.
- Exercise the UMIACS-to-SDSC connection and compare to the GSFC-to-SDSC results.

3.3.2.2. SDSC

Testing with SDSC (figure 9) leveraged the in-place, SDSC Series 4000 switch. WAN

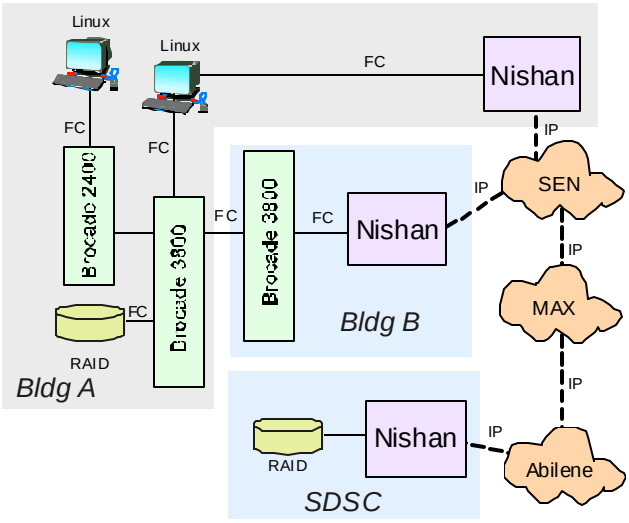

Figure 9 – SDSC Configuration

connection used the Abilene backbone with MAX as the local hopping off point for GSFC. *IOzone* benchmarks were performed building a native ext2 file system on SDSC Sun storage from a GSFC resident Linux host.

Impressions

Set-up was straightforward with only the expected configuration items to be dealt with, namely network routing and allocating the appropriate zones, resolving SAN IDs, etc. However, the switches could not be made to operate in the jumbo frame (MTU=4096) mode, although the network was theoretically configured for such operation. It was learned though trial and error that manually forcing the MTU setting to 4096 can result in very erratic behavior of the link including complete lock up. The next two graphs (figures 10 and 11) illustrate performance as a function of the various Nishan settings for random versus static data.

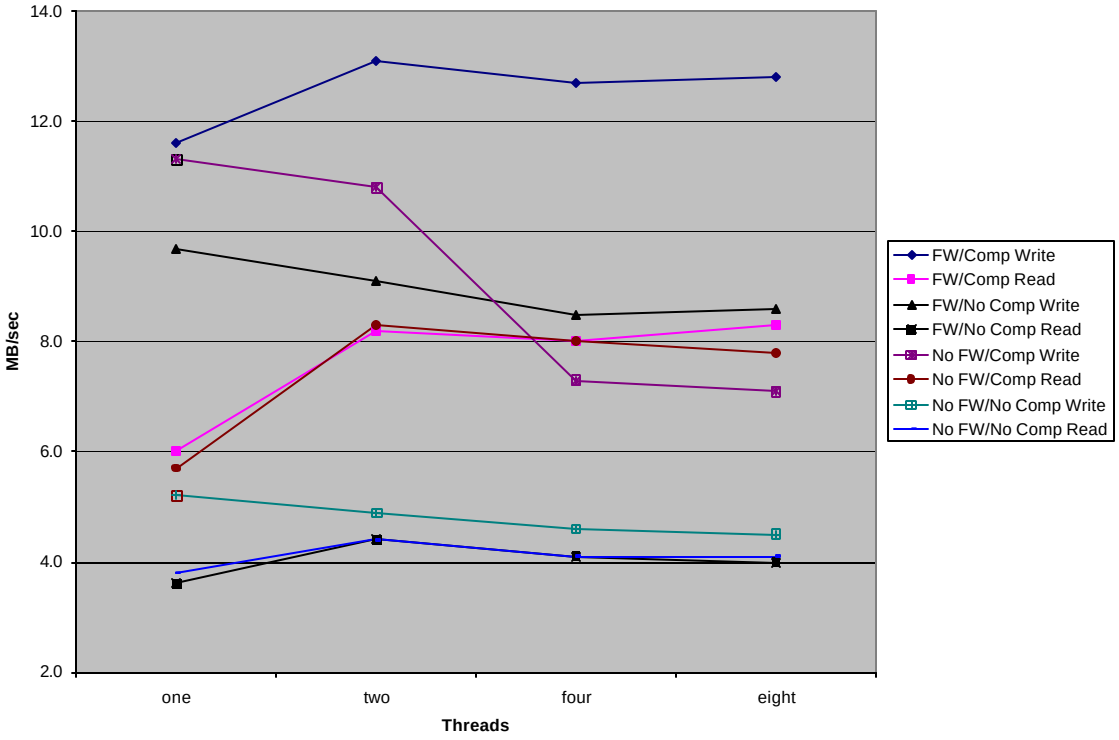

Figure 10 - Random Data, MTU=1500

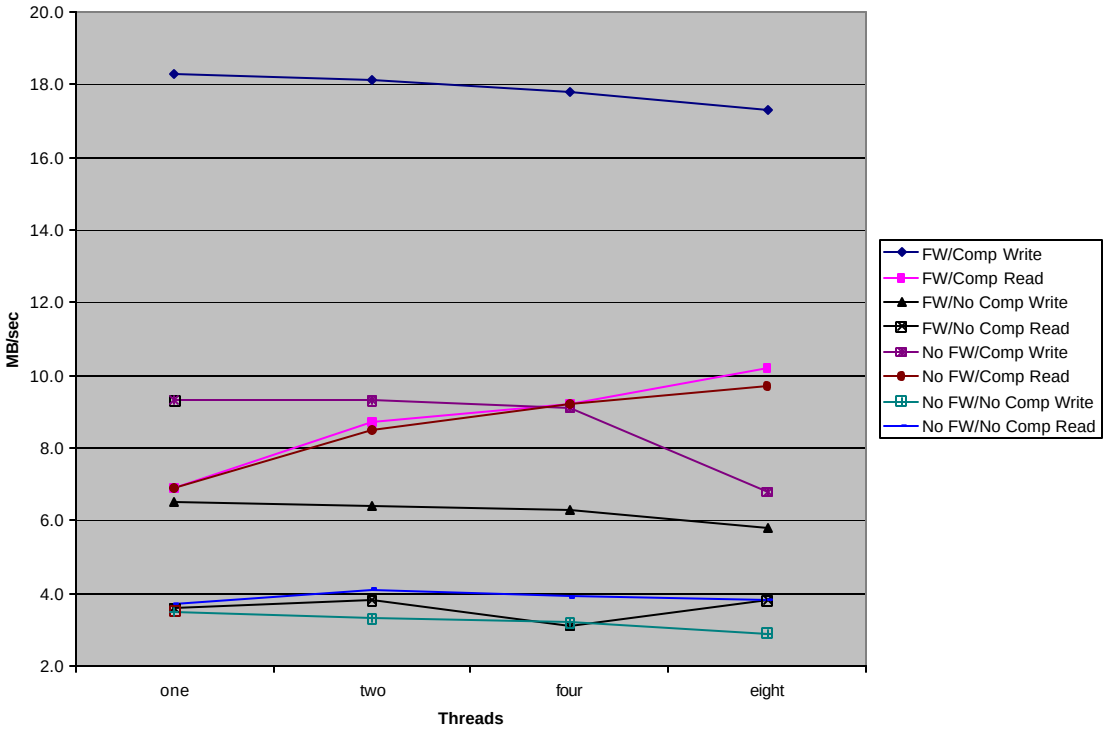

Figure 11 - Static Data, MTU=1500

The following data (Table 2) compares actual results of the GSFC-to-SDSC connection with test data using the NIST simulator with an equivalent rtt of 70msec. In both cases, Fast Write and compression are turned on. Note fair agreement in the data despite the difference in MTU sizes. The suspicion is that the rtt impact on the SCSI command interchange dilutes the performance gains of jumbo frames.

Table 2 - Results

	GSFC => GSFC rtt delay => 70msec MTU => 4096		GSFC => SDSC rtt actual => 70msec MTU => 1500	
Threads	Write	Read	Write	Read
one	13.1	5.6	11.6	6.0
two	13.1	11.5	13.1	8.2
four	13.1	12.5	12.7	8.0

Future Testing

Additional tests to be conducted include:

- Get jumbo frames (MTU=4096) working between GSFC and SDSC then reevaluate performance and compare to delay numbers. Determine if the jump in performance was an anomaly related to the NIST emulator.
- Exercise link in opposite direction – server/host at SDSC and storage at GSFC.
- Exercise the SDSC-to-UMIACS connection and compare to SDSC-to-GSFC results.

3.3.2.3. NCSA

The IP connection with NCSA (figure 12) was accomplished using a pair of LightSand i-8100s. As with SDSC, WAN connection used the Abilene backbone with MAX as the local hopping off point for GSFC. *IOzone* benchmarks were performed building a native ext2 file system on NCSA DataDirect storage from a GSFC resident Linux host.

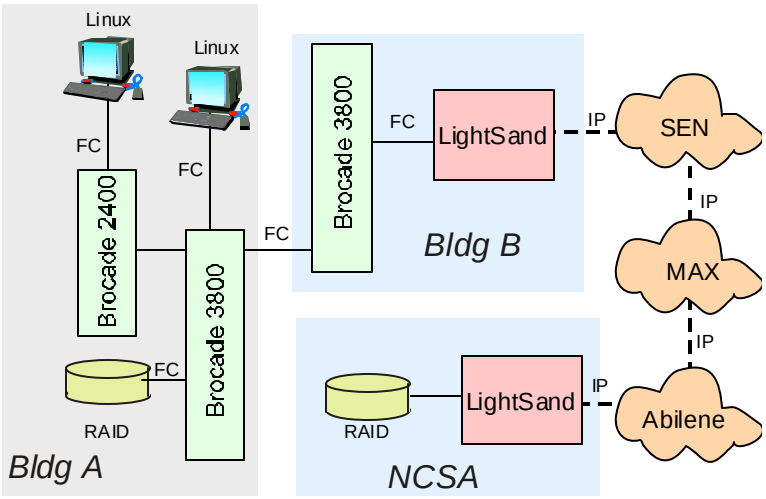

Figure 12 - NCSA Configuration

Impressions

Initial set-up was time consuming because of the learning curve of dealing with the LightSand equipment and establishing the network connection between GSFC and NCSA. The LightSands required that the Brocade 3800 switches be at the 3.1 firmware level. In addition, the command "portcfgislmode <port>,1" also had to be issued to the Brocades so that the switch ports connected to the 8100s would get the R_RDY set. An inordinate amount time was spent trying to determine why the SANman GUI would not execute properly from a remote workstation (off campus with respect to GSFC). As it turns out, NASA blocks external pings from open networks and the first thing the LightSand GUI requires is a successful ping to make sure the connection is in place. Once properly configured, the DataDirect Networks storage at NCSA was easily configured and accessed. Using the same *IOzone* script as before, the following results (Table 3) where obtained for native, ext2 file transfers.

Table 3 - Results

rtt => 30msec 1MB block

Linux Host 1		
Threads	Write	Read
one	37.0	12.1
two	37.5	28.9
four	37.3	35.6
eight	37.3	36.2

These numbers are consistent with the theoretical maximums as predicted by the TimeCalc utility provided with the SANman. An interesting although not perfect comparison is the 35msec rrt numbers obtained using the NIST Net network emulator and Nishan switches. The best results with Fast Write and Compression turned off, was 26MB/sec writes and 20 MB/sec reads. It seems fair to presume, that running the Nishans in the "auto" compression mode may have improved those results.

Future Testing

Additional tests to be conducted include:

- Exercise link in opposite direction – server/host at NCSA and storage at GSFC.
- Get raw bandwidth numbers for the GSFC to NCSA link using nuttcp.

4. Operational Users

As to what might seem like a sidebar to the major thrust of the evaluation, the search for a relevant application of this technology, a geographically distributed file system, continues. Two GSFC groups, the Scientific Visualization Studio (SVS) and the Advanced Data Grid (ADG) Project, are currently being pursued to provide on-campus operational proof of the various connectivity schemes. The plan is to also involve UMIACS, SDSC and NCSA in relevant application demonstrations.

4.1. Scientific Visualization Studio

The GSFC SVS has a need for approximately 1 TB of storage to use as an animation "scratch" area. The content/data to be stored will be scientific visualization animation frames in both HDTV and NTSC resolutions, and MPEG-1 and MPEG-2 movies in various resolutions from web to HDTV. Relatively fast (high bandwidth) access to such volumes is required, including constantly writing frames, various types of processing (read/write) of frames, and streaming frames from this volume to the local SVS workstations for animation preview. A Linux server in the SVS has an FC connection to the SAN Pilot.

4.2. Advanced Data GRID Prototype

In conjunction with NASA Ames, the ADG prototype is a new initiative that intends to leverage the availability of Landsat data The mechanism for making the data available is the SAN Pilot connected to a Sun 3800 located on the GSFC campus.

5. Supporting Technologies

Other technologies are being evaluated to ease the administrative burden of SANs as well improve the performance of the chosen data transport mechanism. The list includes SAN management software and a new generation of network interface (NIC) cards. Also, the evolution of network attached storage (NAS) is also being monitored.

5.1. SAN Management Software

With the emphasis on connecting operational users, part of the testing has focused on SAN management software and tools. The goal is to acquire a tool or suite of tools that enables efficient monitoring of the SAN health and utilization as well as providing for asset allocation and administration. A mechanism is needed that readily discovers SAN components and provides a topology view of the infrastructure.

Four such tools have been installed and evaluated:

- BrightStor™ SAN Manager by Computer Associates International, Inc.
- SANavigator® by SANavigator, Inc. a subsidiary of McData Corporation
- SANScreen by Onaro, Inc.
- Fabric Manager and WEB TOOLS by Brocade Communications Systems, Inc.

The shortcoming of all such products seems to be coverage of all the needed versions of operating systems, and storage and interface devices, something not usually supported. Recognizing the new breed of FC and FC related products, such as Nishan and LightSand boxes, is sporadic as well. No one product seems to do it all. Not tested but briefed was a StorageAuthority™ Suite from AppIQ, Inc. It possesses some very rich capabilities worthy of consideration. In the meantime, SANScreen was purchased and installed. It will be important to observe how the product deals with a heterogeneous, near operational environment with ever evolving security constraints.

5.2. NIC Evaluation

This testing is most relevant to iSCSI connected hosts. The plan is for parametric evaluation of generic NICs versus TCP Off-Load Engine (TOE) NICs and TOE iSCSI NICs. It will be key to measure end-to-end throughput performance and CPU utilization on hosts with different processor speeds. The intent is to include cards from multiple manufacturers such as Intel, Adaptec, and Alacritech. Testing is underway but not yet completed. So far, getting the basic set-up configured and operational is proving to be a challenge.

6. Summary

In retrospect, the testing permutations became formidable when the multiple locations, potential rtt, equipment configurations and settings are factored in. As a result, only a subset of possible hardware and software combinations were actually exercised. However, the size of the data sampling does not adversely impact the overall evaluation of the products. Evaluating IP devices has been an educational process punctuated by learning new jargon and redefining the concept of a SAN while dealing with the unavoidable reality of the hardware and software incompatibilities, typical of emerging technology. This class of product is mainly deployed in disaster recovery applications as opposed to file system applications. As a result, empirical data for comparison was not readily available, leaving conversations and paper exercises as the basis for determining the validity of the collected data. A better understanding of theoretical maximums as they relate to SCSI transfers as a function of rtt versus the selected FC-IP protocol (FCIP or iFCP) is needed.

The vendor products behaved admirably with one significant, non-performance concern. Security features were found to be lacking from a device management perspective – no secure login, clear text passwords, etc. To circumvent such shortfalls during the testing, network routing was altered and access lists were incorporated to minimize the perceived vulnerabilities. Also, a desirable feature available at the data level for iSCSI is host authentication by the IP interface. The following table (Table 4) presents a qualitative review of the Nishan and LightSand equipment:

Table 4 – Findings Summary

IP Device	Pros	Cons
General	• Perform as advertised. • Operationally fairly intuitive. • Both GUI and CLI management options. • Administrator defined level of SAN merging/isolation.	• Minimal security. • No ssh. • No CLI standard • Redundant, conflicting naming conventions. • Proprietary, same vendor product required at both ends of the WAN connection. • High skill level to configure, etc., multiple talents involved. • Incompatibilities, version issues, etc. reminiscent of the early days of FC.
Nishan 3000	• Built in performance graphs. • Good statistical info.	• Passwords in clear text.
LightSand i-8100	• Companion applications that provide data analysis.	• IP routes cleared by reboots. • Difficult to save and compare configurations.

A sidebar to the qualitative aspects of the testing is that the majority of configuration, benchmarking, etc. was done remotely from third party locations, not at any of the centers. Besides the obvious advantage of permitting geographic flexibility for the testers and vendors, it had the interesting side effect of revealing obstacles to deploying such a methodology for an operational IP based SAN. In place site security procedures and firewalls had to be acknowledged and understood. Blocked ports and disabled functionality had to be navigated. Such activity led to a greater understanding of the equipment and what changes would be welcomed in the products.

Certainly at one level the objective of the testing was met – to gain experience with data over IP devices. Understanding the requirements being levied against a proposed SAN has always been critical, but the extra layer of configuration encountered installing FC-IP devices makes such planning even more necessary. There is the usual FC zoning at the local SAN level but in addition, bridging disparate SANs requires designating which components – servers, storage, etc. – will be mutually shared by the co-joined SANs. This two-step mechanism, while adding to the rigor, ensures isolation and privacy of the local SAN while allowing the sharing of mutually agreed to assets. Plans fell short in terms of evaluating a geographically distributed file system (SNFS and/or CXFS) encompassing GSFC, UMIACS, SDSC and NCSA, an outcome planned to be rectified in the near future. These file systems have centralized agents that control their overall operation. It will be interesting to track data movement performance (throughput) as a function of where in the topology the agent is located and the latencies incurred in accessing it.

Acknowledgements
The author wishes to acknowledge the following individuals for their contributions: Bill Fink, Paul Lang, Wei-Li Liu and Aruna Muppalla at NASA GSFC; Bryan Bannister and Nathaniel Mendoza at SDSC; Chad Kerner at NCSA; and Fritz McCall at UMIACS. Gratitude is also extended to the vendor community for their rich support.

References
[1] Hoot Thompson, Curt Tilmes, Robert Cavey, Bill Fink, Paul Lang, Ben Kobler; Architectural Considerations and Performance Evaluations Of Shared Storage Area Networks at NASA Goddard Space Flight Center; Twentieth IEEE/Eleventh NASA Goddard Conference on Mass Storage Systems & Technologies; April 7-10, 2003.
[2] http://www.npaci.edu/DICE/SRB/
[3] J. P. Gary; Research and Development of High End Computer Networks at GSFC, Earth Science Technology Conference, College Park, MD; June 24-26, 2003.
[4] http://www.maxgigapop.net/
[5] http://abilene.internet2.edu/
[6] Maximizing Utilization of WAN Links with Nishan Fast Write; Nishan Systems.
[7] FAQ on Nishan Systems' Compression Technology; Nishan Systems.
[8] Phil Andrews, Tom Sherwin, Bryan Bannister; A Centralized Data Access Model for Grid Computing; Twentieth IEEE/Eleventh NASA Goddard Conference on Mass Storage Systems & Technologies; April 7-10, 2003.
[9] http://www.bitmover.com/lmbench/
[10] http://www.iozone.org
[11] http://snad.ncsl.nist.gov/itg/nistnet/
[12] ftp://ftp.lcp.nrl.navy.mil/pub/nuttcp/beta/nuttcp-v5.1.1.c