Scaling a Global File System to the Greatest Possible Extent, Performance, Capacity, and Number of Users

Phil Andrews, Bryan Banister, Patricia Kovatch, Chris Jordan,
San Diego Supercomputer Center
University of California, San Diego
andrews@sdsc.edu, , bryan@sdsc.edu, pkovatch@sdsc.edu, ctjordan@sdsc.edu

Roger Haskin
IBM Almaden Research Center
roger@almaden.ibm.com

Abstract

We investigate here, both theoretically and by demonstration, scaling file storage to the very widest possible extents. We use IBM's GPFS file system, with extensions developed by the San Diego Supercomputer Center in collaboration with IBM. Geographically, the file system extends across the United States, including Pittsburgh, Illinois, and San Diego, California, with the TeraGrid 40 Gb/s backbone providing the Wide Area Network connectivity. We show the results from two demonstrations, at each of the past two Supercomputing conferences, SC03 in Phoenix, Arizona, and SC04 in Pittsburgh, Pennsylvania. The second demonstration was purposely designed to presage an intended production facility across the National Science Foundation's TeraGrid[1].

1. Introduction

At SC03, the work was performed over a single 10 Gb/s link and the performance characteristics came from a simple transfer between two sites: the SDSC machine room and the SDSC booth on the show floor in Phoenix. A Global File System (IBM's GPFS) was used for the transfers with an application simply opening the remote file issuing sequential read requests. The performance achieved was very encouraging: approximately 90% of the peak bandwidth (~9 Gb/s) was sustained over an extended period. This gave us the confidence to pursue this option as a viable high performance computing paradigm for Grid applications using the TeraGrid.

In the SC'04 demonstration, GPFS was agin used three sites were involved in high performance transfers: the SDSC booth in Pittsburgh, the SDSC machine room in San Diego, and the Nations Center for Supercomputing Applications (NCSA) machine room in Urbana, Illinois. Each site had 30 Gb/s connections to a 40 GB/s backbone and purposely mimicked the expected behavior of a large, distributed application using Grid computing to write many terabytes of data to a central repository, from where it is read by several sites. A new method of authentication, translating GSI certificates [2] to UIDs, as would be required in a true Grid computing environment, was developed by SDSC in collaboration with IBM and incorporated in the demonstration software. Performance was again very satisfying: sustaining approximately 24 Gb/s and peaking at over 27 Gb/s.

. Next year, we hope to connect to the DEISA computational Grid in Europe which is planning a similar approach to Grid computing, allowing us to unite the TeraGrid and DEISA Global File Systems in a multi-continent system.

.

2. The SC'03 Demonstration

0-7695-2318-8/05 $20.00 © 2005 IEEE

At Supercomputing 2002, we demonstrated a Wide Area Global File System [1] using FCIP encoding with specialized hardware. In this paper we consider the more general use of a Global File System across a standard TCP/IP network.

The Supercomputing 2003 meeting was held in Phoenix, Arizona, and was chosen
by SDSC and IBM for the first demonstration of a Wide Area Network implementation of IBM's General Purpose File System (GPFS). The normal GPFS[2] architecture is shown in Figure 1., where server nodes connected to the file system disks by Fibre Channel export the data to other nodes across a local area network, which may be running IP, or an IBM proprietary protocol. For the SC'03 demonstration, a pre-release version of GPFS was used which could export across a Wide Area Network, with enough latency tolerance to handle truly continental extent.

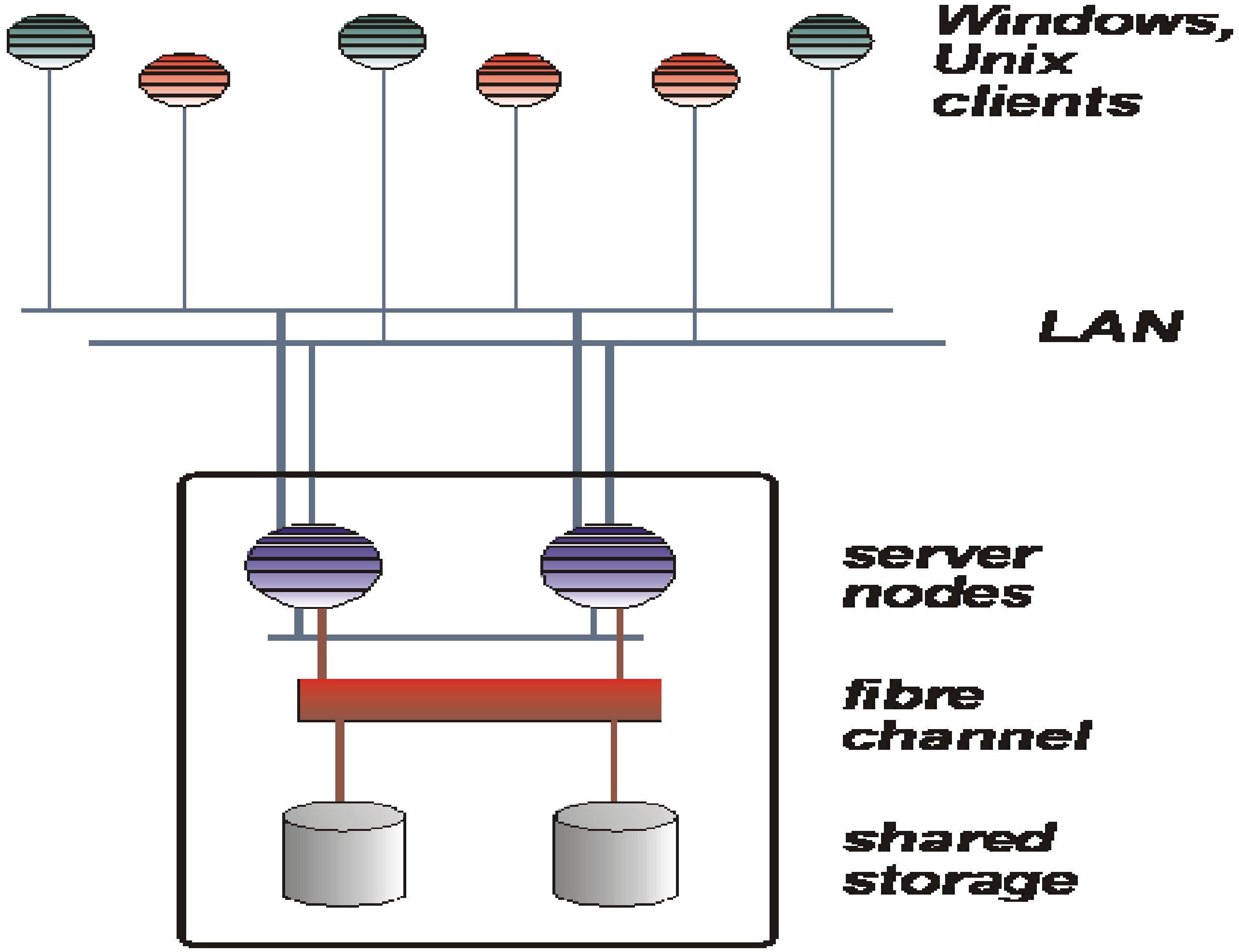

Figure 1. Normal GPFS architecture

For the SC'03 demonstration, a 10 Gb/s link was used from SDSC to the show floor in

Phoenix, and on to the NCSA machine room in Illinois. Data was served up from the high performance file system[3] in San Diego and used as input to a visualization application running in Phoenix and at NCSA. The file system was mounted directly at the two remote sites and straightforward I/O was performed by the visualization application. Figure 2 show diagrammatically the networking and computational infrastructure.

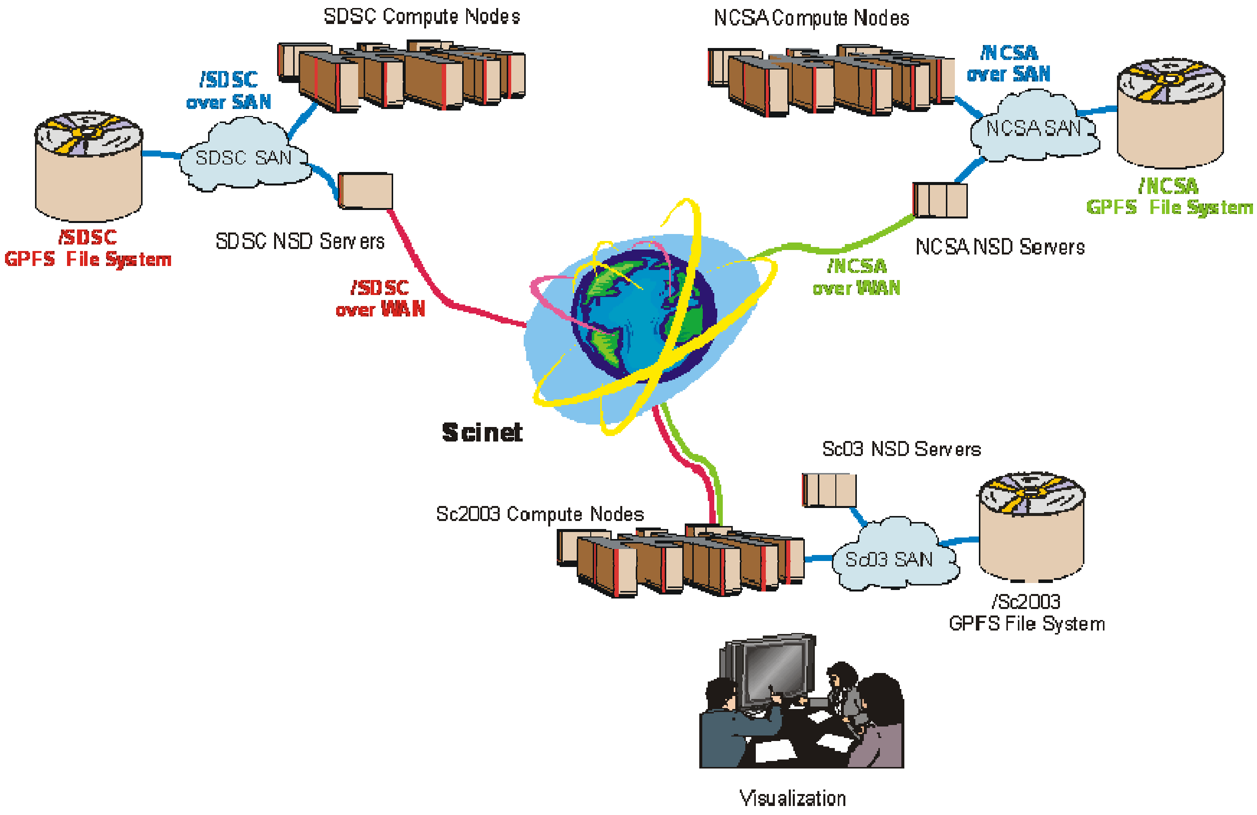

Figure 2. The SC'03 networking and computational infrastructure.

3. File System Performance.

The data used was output from a large scientific application, output to the GPFS file system at SDSC. It was then read in across the WAN by visualization applications on the SC'03 show floor and at SDSC. Figure 3 shows the transfer rate between San Diego and Phoenix: it was truly exceptional, sustaining approximately 90% of the 10 GB/s link using standard IP protocols. The dip in the middle is where the application terminated after completion and was restarted. As a first demonstration of the WAN GPFS file system, this was extremely encouraging and we looked for ways to implement this as a production system.

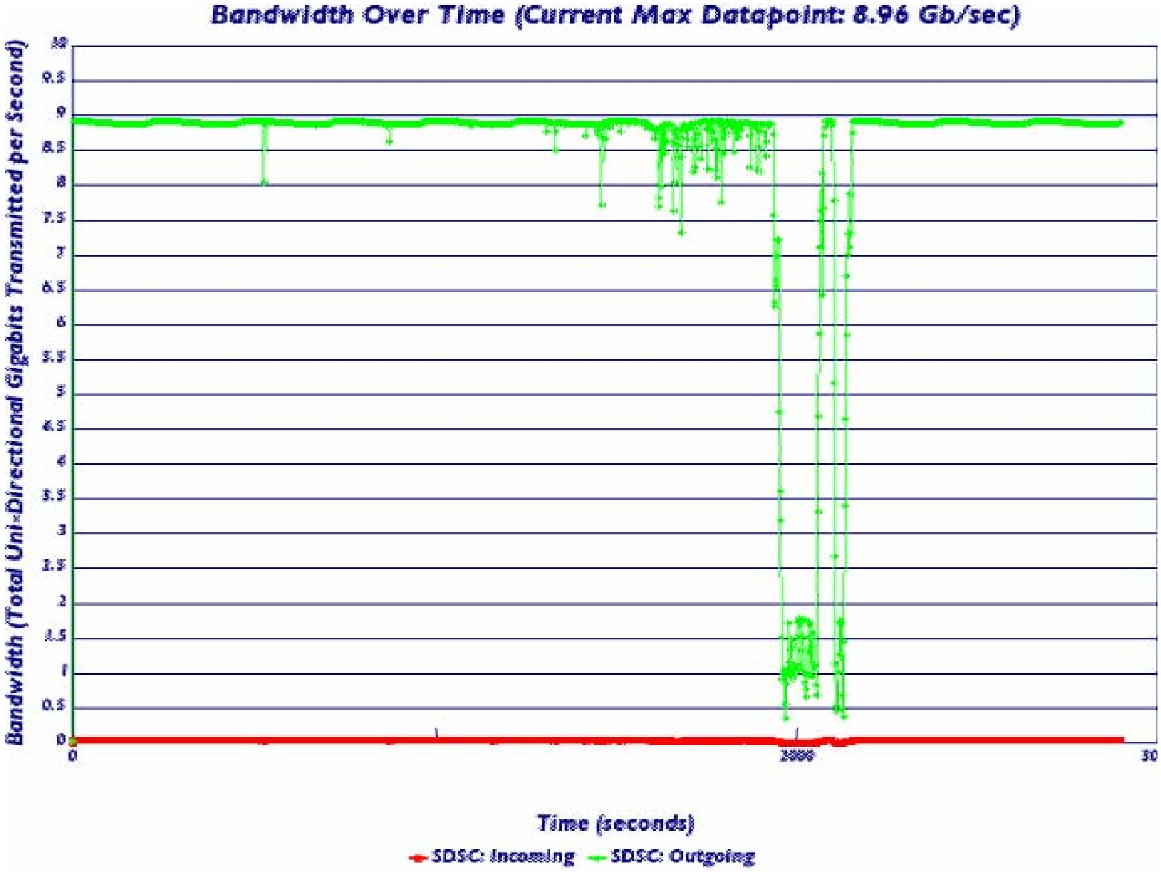

Figure 3. File transfer rates between San Diego and Phoenix

4. The SC'04 demonstration

To proceed towards a production implementation, several issues had to be considered. Firstly, we wanted to see the maximum performance that could be achieved across the TeraGrid WAN[4], where the backbone is at 40 Gb/s to which the major sites are connected at 30 Gb/s. Secondly, a large file system, preferably greater than 100 TB would be necessary for real production use in a Supercomputing environment. Thirdly, authentication must be extended from simple UIDs. We would also like to demonstrate extremely high local transfer rates for the file system, as these may be necessary for efficient data staging, backups, etc..

4.1 Infrastructure

Figure 4 shows the data and computational infrastructure for the SC'04 demonstration. We plan to make good use of the disk provided to the SC'04 StorCloud booth by IBM, using 160 TB of Fast T600 comprising 15 racks with 2 controllers per rack. The 40 servers are dual processor 1.3 GHz IA64 boxes, with each server having 2 FC connections. Local transfer rates should approach 20 GB/s. Each server has a single Gigabit interface connected to the SCinet LAN. This LAN is in turn connected to the TeraGrid backbone by a 30 Gb/s link, mimicking the normal TeraGrid major site connection. We hope to fill as much of the 30 Gb/s link as possible.

In order to ensure that the Wide Area Network capabilities of both the network hardware and file system software were being tested, it was essential to construct a very high performance local network on the show floor and to connect it to equivalently high performance systems at both SDSC and NCSA. The disks for data storage were provided by IBM and resided in the StorCloud booth as part of their vendor participation in the show. Fifteen racks of 4 FastT600 disk systems each were used, with 2 controllers per rack and four 2Gb/s Fibre Channel connections controller for a total of 120 FC connections. In the SDSC booth were 40 dual-processor, Itanium2 (1.3 GHz) systems, each with three 2Gb/s Fibre Channel Host Bus Adapters and a Gigabit Ethernet interface. The Itanium servers were running GPFS in SAN mode, i.e., using Storage Area Network connections, each of the servers could see all of the disks, and ran both GPFS server and client software on the nodes. The switching for the Storage Area Network was provided by Brocade, with 3 Brocade Silkworm 24000 director switches, each with 128 2 Gb/s FC interfaces.

Local disk performance was excellent, with approximately 15 GB/s being sustained and new world records were set in both the Terabyte sort (487 seconds) and the minute sort (120 GB).

Connectivity to the TeraGrid from the servers on the show floor was via a GbE connection from each of the 40 servers to the SCinet Local Area Network within the convention center, and then via three 10 GbE links to the TeraGrid backbone. At NCSA, 40 similar Itanium 2 servers, each with GbE connectivity, mounted the GPFS file system from the SDSC booth at SC'04. These nodes were used for the file transfers.

At SDSC in San Diego, the GPFS file system in the SDSC booth at SC'04 in Pittsburgh was mounted on two very different systems. The first was the IBM Power4 system, DataStar. This is a 10.4 Teraflop general purpose compute systems with highly parallel I/O capabilities running IBM's AIX operating system. Sixty-four of its 8-processor P655 nodes mounted the remote GPFS file system and these nodes (512 processors) ran the Enzo application; writing the output data directly to the remote disk. Connectivity was via a GbE adapter in each node to the local Force10 switch, then via a Juniper T640 router to the TeraGrid backbone.

In addition, the remote GPFS file system was mounted on 40 Itanium2 two-way nodes running Linux in the SDSC machine room. The version of GPFS used (2.3beta) allowed sharing of the file system between AIX and Linux clusters. These systems were used for visualization of the data, and it was transfers between the SC'04 show floor and these nodes at SDSC that were used for transfer rate measurements.

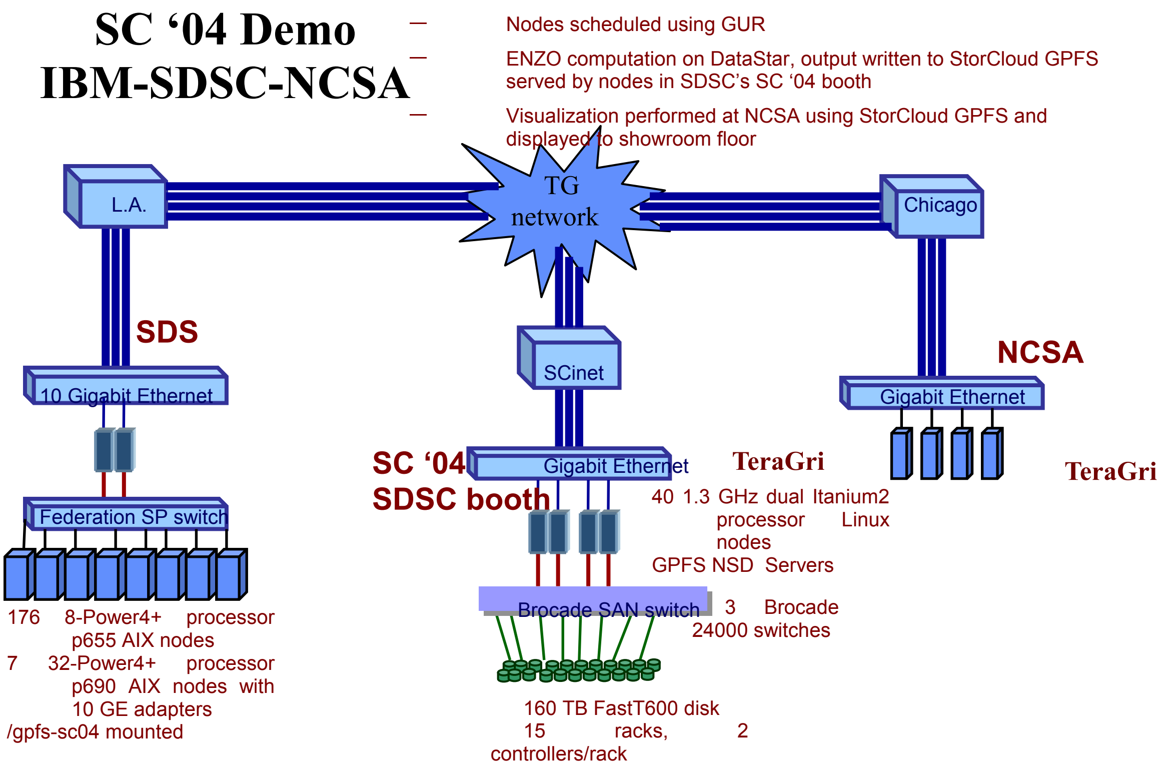
SC '04 Demo
IBM-SDSC-NCSA
Nodes scheduled using GUR
ENZO computation on DataStar, output written to StorCloud GPFS served by nodes in SDSC's SC '04 booth
Visualization performed at NCSA using StorCloud GPFS and displayed to showroom floor
L.A.
TG network
Chicago
SDS
10 Gigabit Ethernet
Federation SP switch
176 8-Power4+ processor p655 AIX nodes
7 32-Power4+ processor p690 AIX nodes with 10 GE adapters
/gpfs-sc04 mounted
SCinet
SC '04 SDSC booth
Gigabit Ethernet
40 1.3 GHz dual Itanium2 processor Linux nodes
GPFS NSD Servers
Brocade SAN switch
3 Brocade 24000 switches
160 TB FastT600 disk
15 racks, 2 controllers/rack
TeraGri
NCSA
Gigabit Ethernet
TeraGri

Figure 4. SC'04 networking and computational infrastructure

4.2 Authentication

As with basically all file systems, GPFS uses the User ID (UID) and Group ID (GID) for authentication. This is perfectly reasonable (and very efficient) for a single machine, or even (normally) within a single organization. However, in a true Grid environment, the same user is likely to have completely different UIDs at the different sites on the Grid. This obviously leads to problems in accessing his data from several sites. For simple datasets in a Supercomputing environment, it may be sufficient to mount them World Read-Only, but for the true facility of Grid computing another approach is necessary. For single sign-on across the TeraGrid, the Globus Security Infrastructure (GSI)[5] is used. IBM has enabled a "hook" into GPFS that has allowed us to write a remapping module that maps the remote user, authenticated by GSI, to a local UID/GID pair. The SC'04 demonstration was the first use of this outside of SDSC. This approach is essential to scaling the file system to a very large number of Grid users.

For a successful implementation, efficiency is essential, and we work closely with the assumptions of the standard GPFS software.

Every GPFS file system belongs to a particular cluster with a given ID space. We use this as the base ID space for GPFS and all Ids written to disk as GPFS metadata will be the IDs from the home cluster ID space, regardless of which node (home, or remote) initiates the operation. In order to map a numeric UID or GID from one cluster to that from another, we rely on an infrastructure of "globally unique names", or GUNs, that exist outside of GPFS. Any GUN will map to a single local ID, but the reverse operation may not be unique. The translation is performed using a set of user-supplied ID remapping helper functions (IRHFs). To avoid the problems of dealing with kernel-space operations, the IRHF is currently called in user space, via exec.

When initially mapping from a UID value to a GUN, the UID is first matched in the password file to extract a username. Once a username has been obtained, there are two ways to get a Globus DN that corresponds to the user - the first is to examine the Globus grid-mapfile on the node and find the DN that maps to that user as far as Globus is concerned. In addition, for situations where the username maps to more than one DN, there is a Globus command that can be used to extract the DN from a users actual GSI certificate on the machine. The first method is advisable since we are re-using an already existing mapping between a DN and a username, but it is not guaranteed that there is only one DN that maps to one username, i.e. a user can have several certificates that they use to access a single account. The second method is guaranteed to return only one DN.

When mapping back from a GUN to a UID, use of the grid-mapfile is unavoidable, but it is guaranteed that one DN will map to only one username in the grid-mapfile, i.e. any one certificate has to be used to access one and only one account. Once the username is obtained, the passwd file is used to find the UID, and the UID then provides a list of GID's from the groups file.

All of these operations are performed with various regular expressions, some of which can be complex.

4.3 Performance

Prior to the Bandwidth Challenge at SC'04, the Enzo application ran on the IBM Power4 DataStar system in San Diego. During the actual application run, significant amounts of data were written across the Wide Area Network to the GPFS disks on the show floor. As the Enzo application is not I/O bound, transfer rates during these operations are not indicative of the maximum possible; however, they are more typical of what might be required by an application writing output data to a centralized resource for future data mining and visualization.

For the actual measurements, we accessed the central GPFS data from both SDSC and NCSA and performed both prototypical data investigations, designed to be representative of real Grid work, and simplified data accesses intended to produce completely I/O bound operations. In this case the difference in transfer rated were inconsequential, but the artificially I/O bound runs produced very slightly greater numbers. It should be mentioned that the data produced by the Enzo application was real data, eventually stored back at SDSC, and became part of an ongoing scientific investigation.

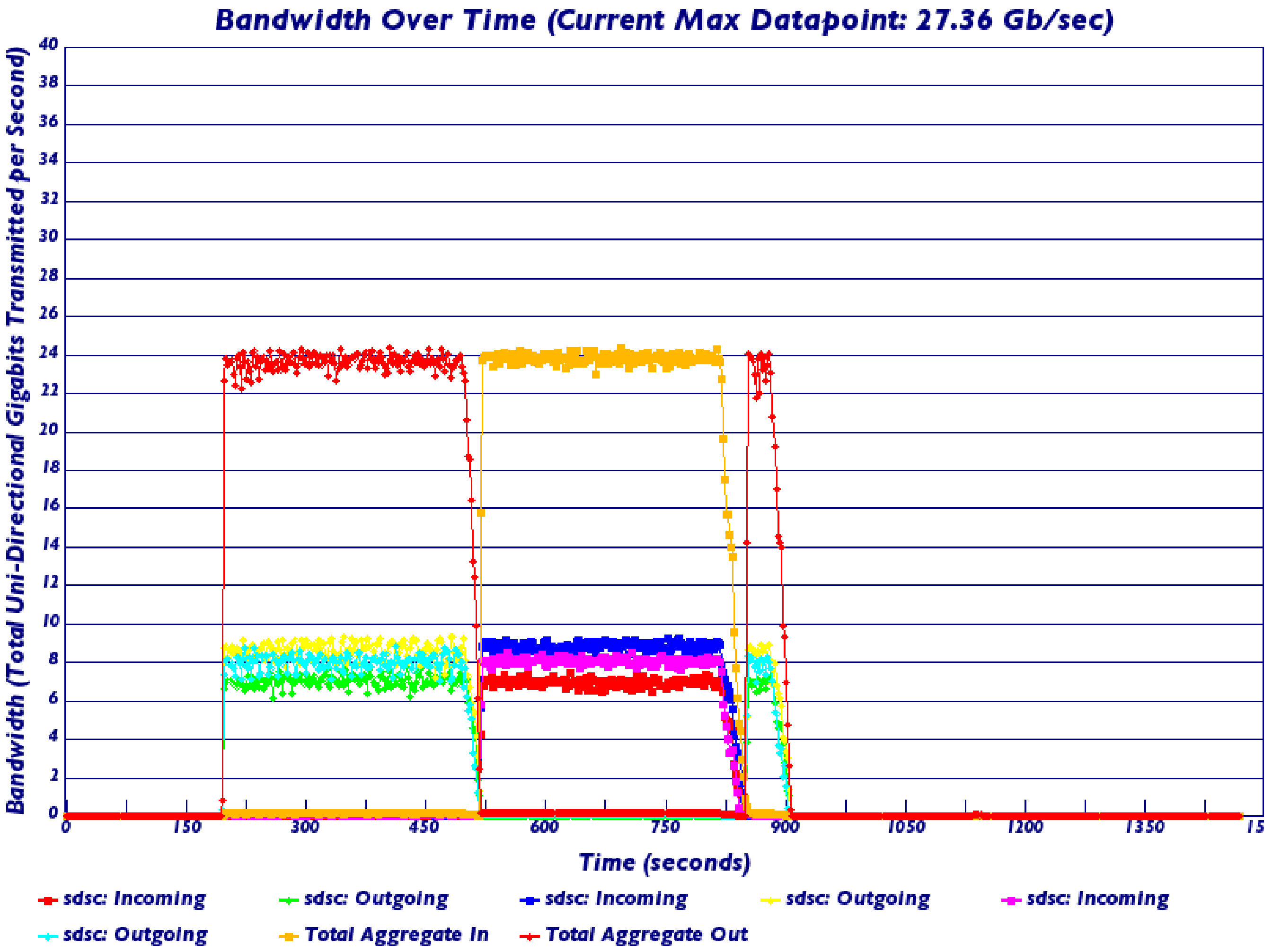

Figure 5. Measured WAN performance at SC'04 for sequential reads and writes

In Figure 5, we show the results of measuring the data transfer rates during the Bandwidth challenge of SC'04. Transfers on the three 10 Gb/s connections between the TeraGrid backbone and the show floor were monitored separately, and are shown both individually and in aggregate. For clarity, reads and writes were performed during distinct periods. Performance between the show floor and either SDSC or NCSA was essentially identical, but only transfers to and from SDSC are shown in the figure. For approximately 300 seconds, data was written from SDSC to the show floor, then read for another 300 seconds, then finally written for a short time to check reproducibility.

Performance for reads and writes was essentially identical, with each of the 10 Gb/s links varying between about 7 and 9 Gb/s transfer rates. The aggregate performance in each case was approximately 24 Gb/s, or 3

GB/s, with a short-term maximum of 27.37 Gb/s.

These were excellent results, as good or better than those obtained by specialist file transfer utilities such as GirdFTP.

5. Future work

Following the success of the SC'04 demonstration, we are implementing a prototype Global File System at SDSC for the TeraGrid, using the software prototyped for SC'04. We are initially mounting approximately 70 TB of FC disk as a WAN-available GPFS Global File System using the same servers that were used in the SDSC booth for the SC'04 demonstration. Later this year, we hope to move to a much larger disk farm, approaching 1PB in size with cheaper, GbE connected disks.

The DEISA grid organization in Europe[6] is also planning a significant shared Global File System and we hope to demonstrate the linkage of the TeraGrid GFS with theirs, creating a multi-continental, high performance, shared file system.

We also plan to explore the possibilities of automatic data archiving[7].

We would like to thank the whole Enzo group[8], especially Robert Harkness and Mike Norman.

6. References

[1] Catlett, C. The **TeraGrid**: A Primer, 2002. www.**teragrid**.org

[2] Foster, I., Kesselman, C., Nick, J.M. and Tuecke, S. Grid Services for Distributed Systems Integration. *IEEE Computer*, *35* (6). 37-46. 2002

[3] A Centralized Data Access Model for Grid Computing, Phil Andrews, Tom Sherwin, and Bryan Banister, Twentieth IEEE Symposium on Mass Storage Systems, (April 2003)

[4]GPFS: A Shared-Disk File System for Large Computing Clusters , Frank Schmuck and Roger Haskin, Conference Proceedings, FAST (Usenix) 2002

[5] State of the Art Linux Parallel File Systems: The 5th Linux Clusters Institute International Conference on Linux Clusters: The HPC Revolution 2004, Austin, TX, May 2004, P. Kovatch, M. Margo, P. Andrews and B. Banister.

[6] http://www.deisa.org

[7] Large-Scale Flexible Storage with SAN Technology, Phil Andrews, Tom Sherwin, Bryan Banister, Eighteenth IEEE Symposium on Mass Storage Systems (April 2001)

[8] http://cosmos.ucsd.edu/enzo/