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

Proceedings of the 22nd IEEE/13th NASA Goddard Conference on Mass Storage Systems and Technologies (MSST 2005)

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.

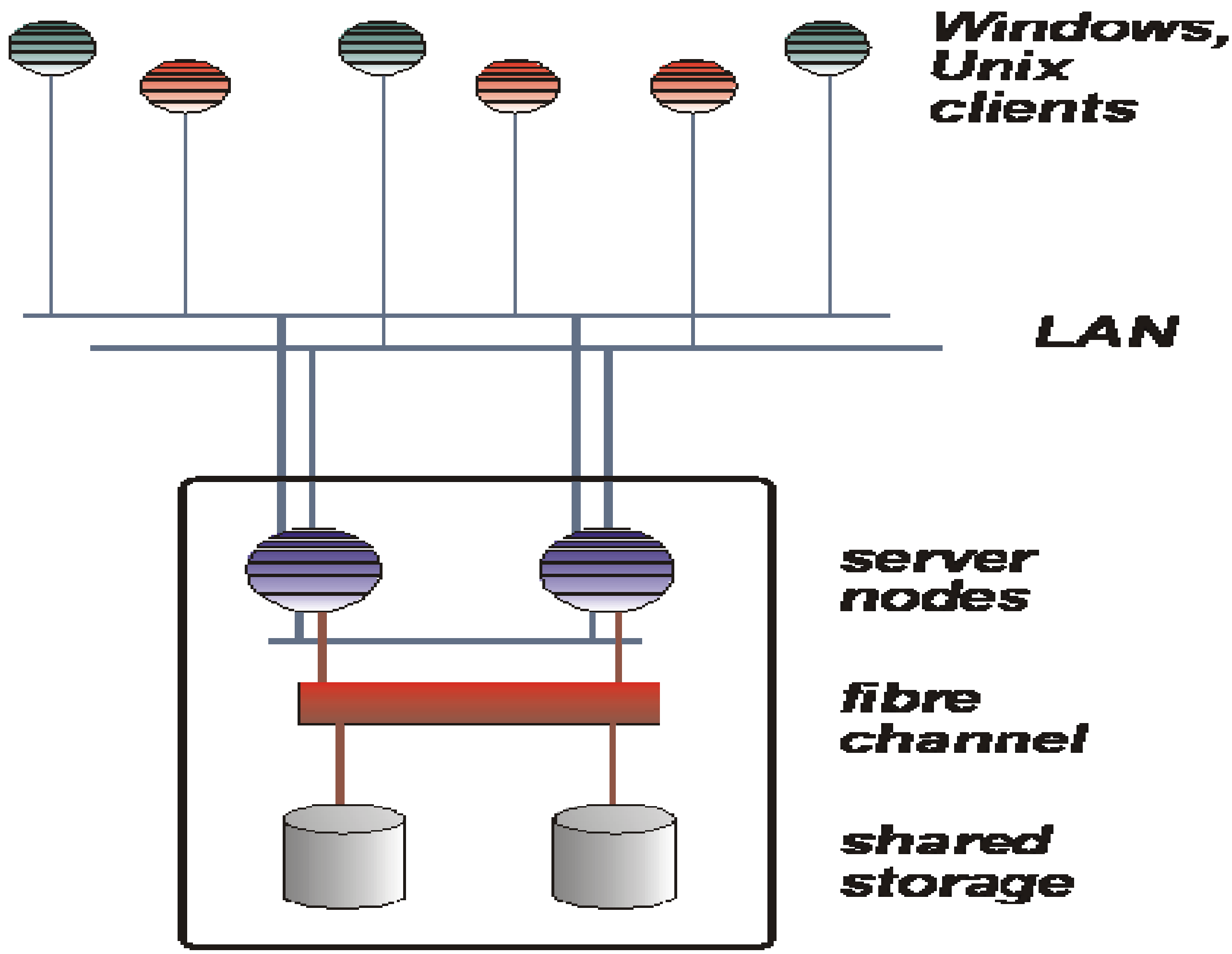

Figure 1. Normal GPFS architecture

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.

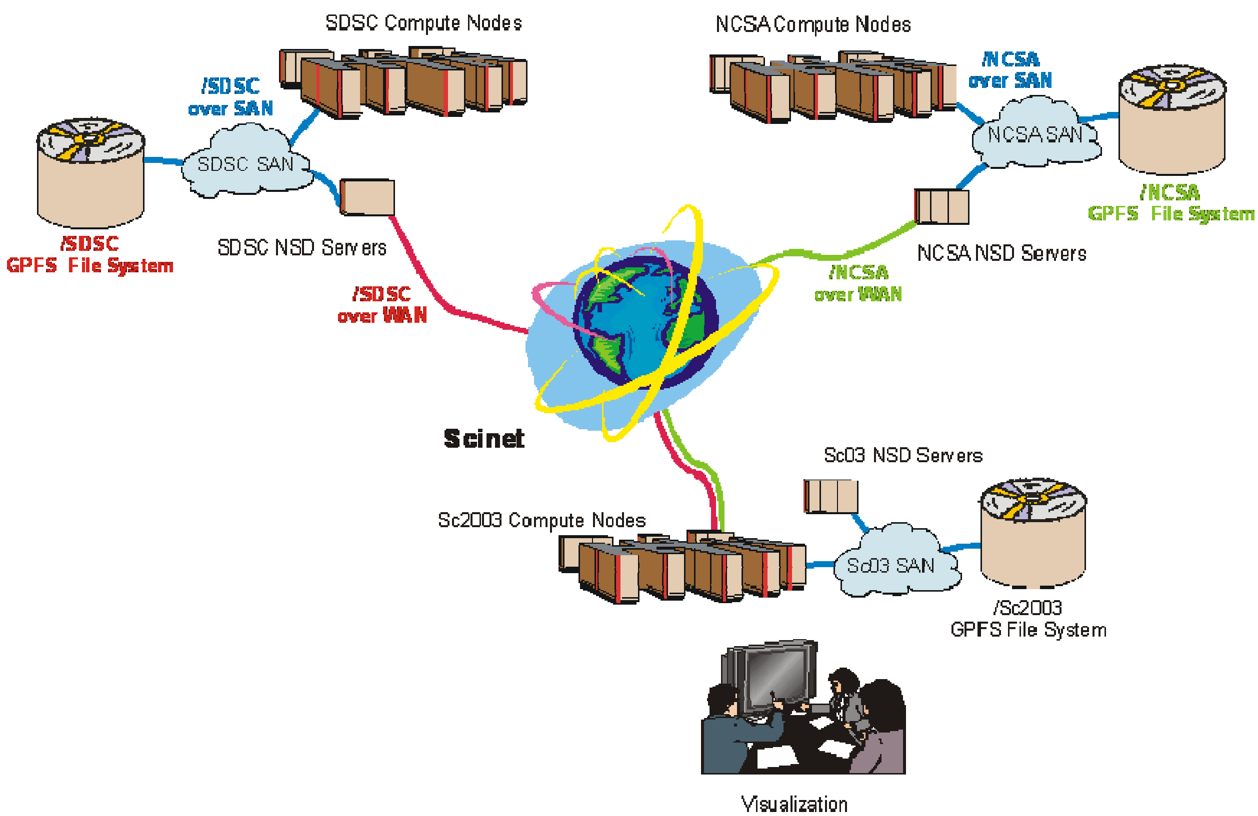

Figure 2. The SC'03 networking and computational infrastructure.

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.

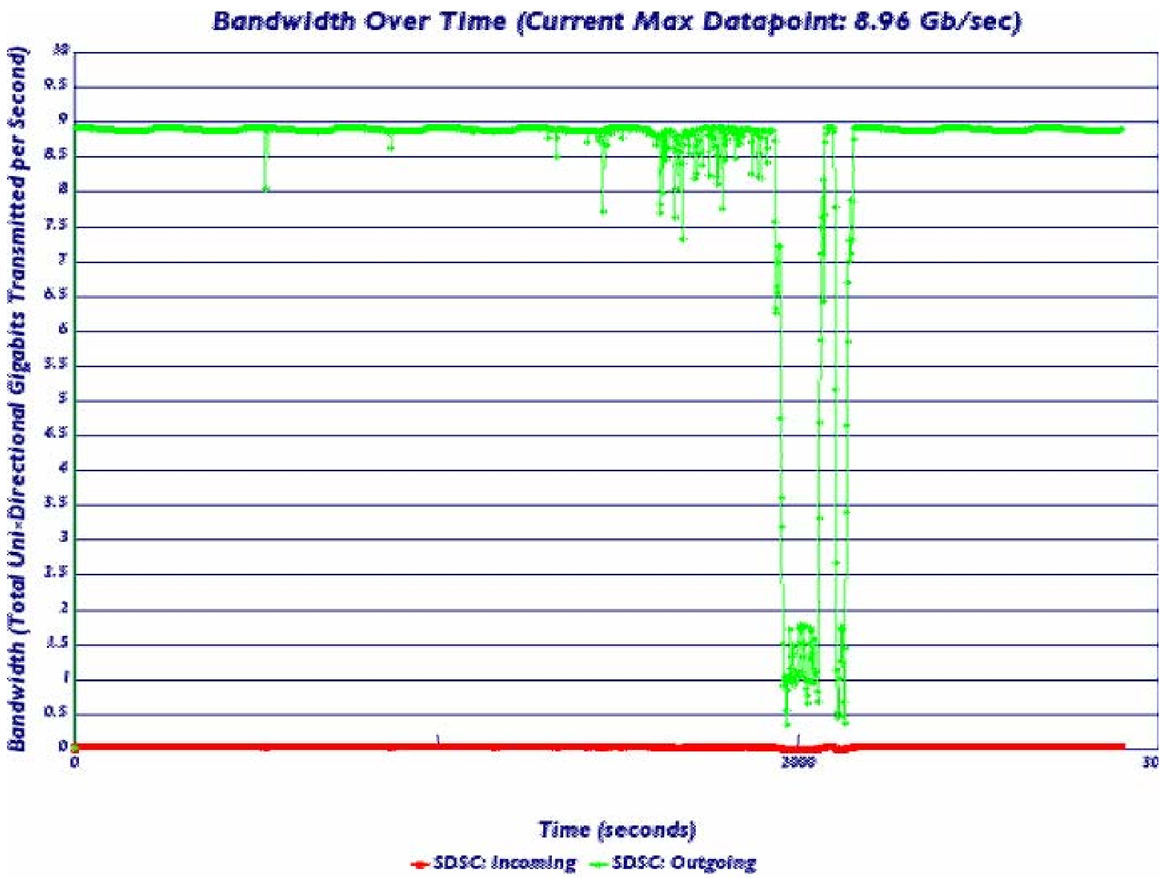

Figure 3. File transfer rates between San Diego and Phoenix

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.