Large-Scale Flexible Storage with SAN Technology

Phil Andrews, Tom Sherwin, and Bryan Banister
San Diego Supercomputer Center
University of California, San Diego
La Jolla, Ca 92093-0505
Andrews@sdsc.edu, sherwint@sdsc.edu, bryan@sdsc.edu
tel +1-858-534-5000
fax +1-858-534-5077

Abstract

The San Diego Supercomputer Center is installing a large (64-port) Fibre Channel Storage Area Network to form both an integral part of, and an alternative architecture to its existing traditional HSM system. Connected to the SAN will be a significant number (>10) Sun servers, a large amount of high-speed native FC disk (>7TB), high-end FC tape capability in the form of IBM 3590E and STK 9840 FC drives, and an IBM SP running the HPSS software.

A portion of the FC disk will be cache disk for the High Performance Storage System and some of the tape drives will also be utilized by HPSS. However, we will also be working with direct backups from the FC disk to the FC tape drives, administered by commercial file system packages such as Veritas and Legato. The initial configuration of the SAN will be as a FC arbitrated loop, but this will quickly become a full switch fabric.

We will be experimenting with several transfer options within this configuration:
1) Direct HPSS transfers between cache disk and tape storage.
2) Transfers between FC disk and FC tape drives mediated by third-party software.
3) Transfers between HPSS FC disk cache and non-HPSS FC disk and tape storage.
4) Transfers to and from the outside world via Gigabit Ethernet connections.
We will present transfer rates and other data on the success, or otherwise, of these operations.

Much of the data stored in this system will be accessible via the Web and several large-scale, high availability Web Servers will be running on the directly connected equipment. We will be attempting to establish this configuration as a workable model for making Petabyte-sized databases available for Web access in support of national-scale scientific experiments. In this respect we will provide latency and transfer rate results for Web access to the data. The SDSC already supports the world's largest HPSS archive (>200TB) and numerous servers including the Protein Data Bank national resource.

1 Introduction

We are presently in the middle of establishing our fibre-channel based Storage Area Network. We are awaiting delivery of four 16-port FC switches that will from the core of our SAN. Implementation has been slowed somewhat by the decision to bypass fibre-channel arbitrated loop technology and go directly to full switch fabric. This will allow our full bandwidth (~100MB/s) to be attained between any two devices, and will also allow the use of tape drives on communal SAN's. The FCAL architecture is prone to Loop Initialization Processes (LIP's) which cause the reset of any tape drives on that loop and terminate tape transactions in progress. In addition to the FC switches, we will be receiving 7TB of native FC disks, with about 1TB already here on test, together with native FC tape drives, also already under test.

Two "ends" of the SAN are already in use at the San Diego Supercomputer Center: the HPSS-based mass storage system, and the servers that can provide data to the WWW. As one of the aims of this work is to see just how fast data can be served up to the web, we have made extensive examination of the read transfer rates on our largest system, working to show just how quickly we can provide data to a network connection. In this paper we will now describe our mass storage system and show the results of our transfer rate experiments.

2 Mass Storage System

SDSC uses the High Performance Storage System running on a second IBM SP, consisting of eight 4-processor Silver nodes, 4 WinterHawk "wide" nodes and 8 WinterHawk "thin" nodes. Disk cache is approximately 1TB of SSA and a similar amount of HiPPI attached disk. Tapes are held in three STK "powderhorn" silos and tape drives are STK 9840's and IBM 3590E's. There are a total of 20 IBM 3590E's and 8 STK 9840's. A new silo and 8 new tape drives will be brought in support the SAN

.

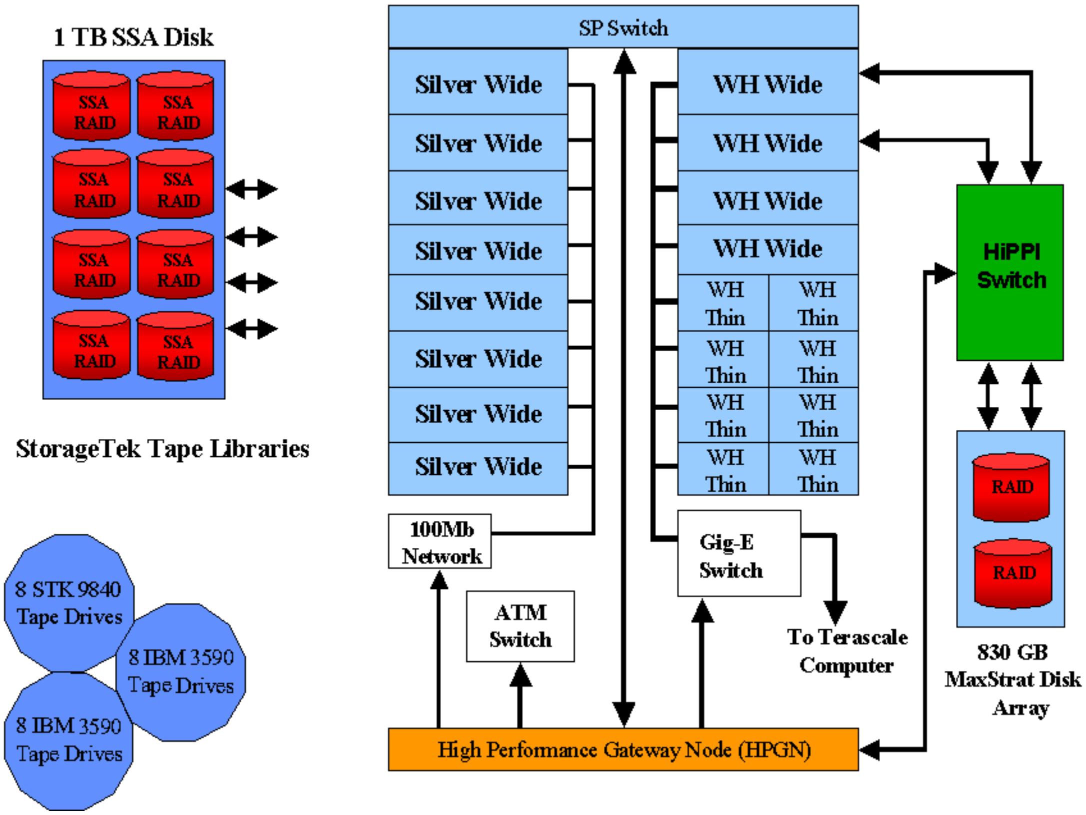

Figure 1. The HPSS mass storage system

3 Server File System and Performance

In order to reach the maximum possible server performance numbers we used our largest single computer, our BlueHorizon (BH) IBM SP.

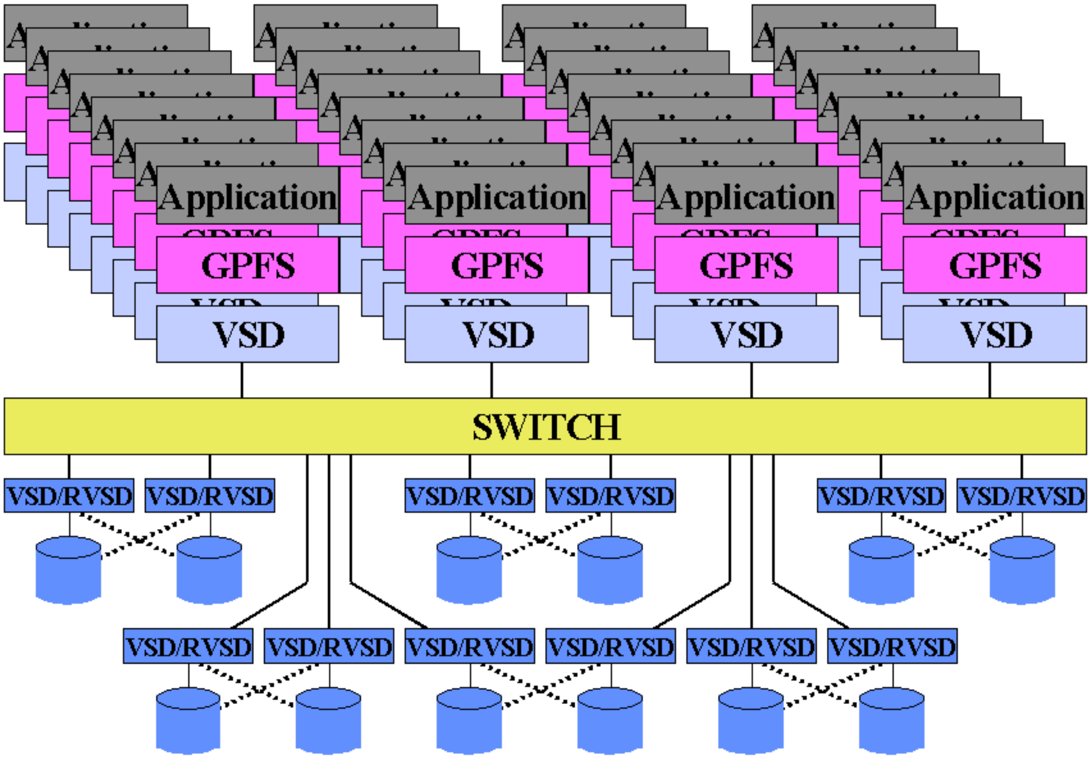

Figure 2. The GPFS file system layers for a 32-node application

Blue Horizon uses NightHawk2 nodes running at 375 MHz. There are 144 compute nodes, each with 8 processors for a combined floating point rating of 144 X 8 X 375 MHz X 4 flops/cycle = 1.728 Tflops. For performance, the local user accessible disk is arranged as a single General Purpose File System (GPFS). GPFS is a proprietary IBM file system which takes disk striping to its maximum by striping files across every disk in the file system. The GPFS client runs on all the compute nodes and accesses data thru a VSD client accessing Recoverable Virtual Shared Disk server software running on 12 RVSD servers. Each of these servers is a 2-processor, 222 MHz NightHawk1 node forming a recoverable SSA loop with another server and twelve 4+P RAID sets of 18 GB drives. This gives a total usable file system of 6 X 12 X 4 X 18GB = 5.2TB. The block size for the GPFS file system is 256KB. Communication between the nodes is over the IBM proprietary Trailblazer at approximately 150 MB/s theoretical maximum..

In table 1 we display the measured read rates for 1 to 128 nodes in powers of 2. The peak at 64 client nodes showed over 1GB/s of transfer rate. This is indicative of our capability for serving data. Currently our network connection from this machine to the outside world is quadruple Gigabit Ethernet, which this machine could easily saturate. We plan to increase connectivity significantly in the near future.

Table 1, read rates for numbers of nodes

1	119.72 [MB/sec]
2	202.15 [MB/sec]
4	358.96 [MB/sec]
8	480.77 [MB/sec]
16	614.65 [MB/sec]
32	728.28 [MB/sec]
64	1,029.95 [MB/sec]
128	913.66 [MB/sec]

The numbers represent 8 threads per node, working on 1,024 100MB files. Each file is accessed only once to avoid any effects of caching within the file system.

In Figure 3 we graph the data of table 1 for 4 to 128 nodes, together with a theoretical linear increase from 4 nodes.

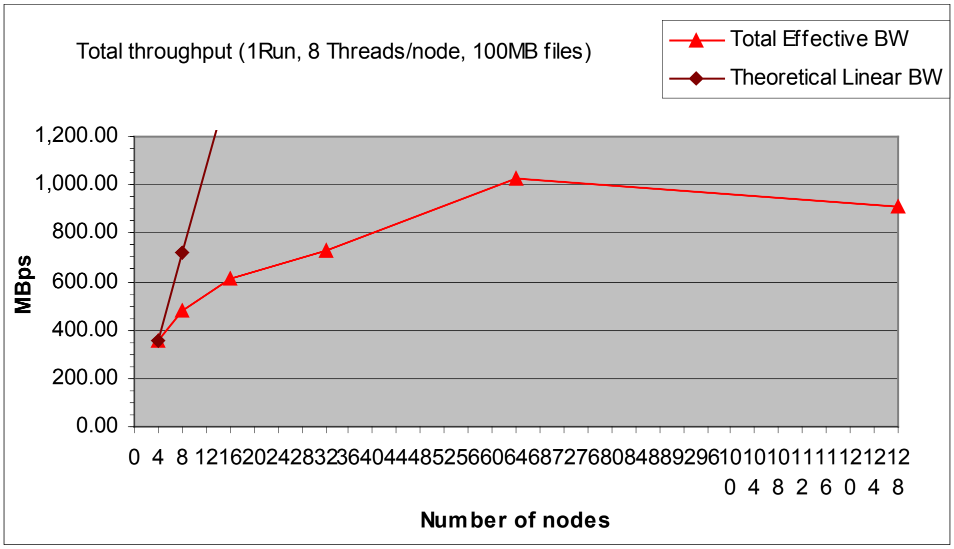

Figure 3. Read rates from GPFS with the Trailblazer switch

3 Improved Switch Performance

After upgrading to the IBM Colony switch with an estimated 450 MB/s transfer rate, we reran the previous tests and observed significantly greater performance. The average read rate increased to almost 1.3 GB/s with 128 client nodes while even at 16 client nodes it was in excess of 1 GB/s. Peak rates exceeded 2 GB/s.

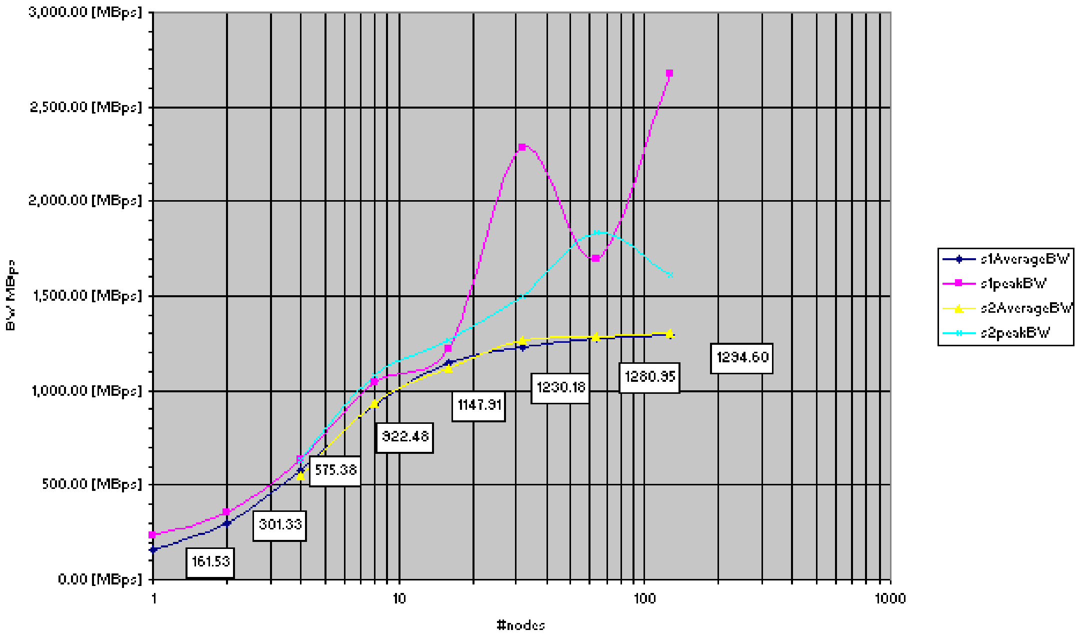

Figure 4. Read rates from GPFS with the Colony switch

4 Further work

We have taken delivery of four 16-port Fibre Channel switches, four Fibre Channel tape drives, and over 7 TB of Fibre Channel disk drives for integration into our SAN environment. These will be combined with a plethora of servers (both mass storage and web interfaces) to constitute an advanced storage area network. We expect to have the configuration in operation by the conference and will report on our performance measurements and experiences.