High-Speed Data Transfer via HPSS using Striped Gigabit Ethernet Communications

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

Abstract
The San Diego Supercomputer Center (SDSC) hosts what is currently the world's largest High Performance Storage System (HPSS) with over 200TB of data. The major compute facility at the center is a Teraflop-sized IBM SP called "BlueHorizon". Although both the compute facility and the HPSS utilize IBM SP computers, they are physically distinct, not sharing a switch fabric. Presently communications between the two systems are via a High Performance Gateway Node (HPGN) but when the BlueHorizon switch is upgraded in the next two months to "Colony" communications, the HPGN will no longer be able to connect the two switches.

To prepare for this we have already established a GbE network oriented towards the highest possible speed communications between the two systems. The IP addresses of the IBM SP nodes on BlueHorizon have been arranged in quadrants so that is possible to stripe GbE communications across 4 parallel GbE connections. This will be used in collaboration with the inherent striping capability of HPSS to attain the highest possible speeds in data transfers between BlueHorizon and HPSS.

We intend to perform four-way, eight-way, and possibly twelve-way striped data transfers between the two systems, primarily in a direct to tape mode. In early experiments that have not yet utilized the GbE connection, we have managed to obtain a transfer rate of 75 MB/sec to eight IBM 3590 tape drives. These drives have a native transfer rate of approximately 9MB/sec (perceived rate is greater because of hardware compression) and have since been upgraded to 14 MB/sec. In addition we intend to purchase 20 MB/sec STK 9840 drives.

The intention is to demonstrate 150 MB/sec transfer rates between BlueHorizon and HPSS via an eight-way software stripe across a four-way GbE hardware stripe. Simultaneously we will be running experiments on BlueHorizon to examine it's capability as an extremely high-speed Web data server, using its GPFS file system for local store and rapid data transfer.

Experiments with IBM's General Purpose File System (GPFS) have shown data transfer rates of up to 1.3B/sec on a 5TB file system. Network connectivity into the SDSC is presently dual OC12 connections with OC48 planned and OC192 threatened. Combining the large storage capacity of HPSS, rapid data transfers to GPFS-controlled local disk on BlueHorizon, extremely rapid parallel transfers from GPFS and very large network pipes

gives us the opportunity to experiment with a truly terascale dataserver for Large Scale Storage in the Web.

1 Introduction

Supercomputing Centers are no longer temples to monolithic compute engines; they must now be an amalgam of resources, carefully tuned so that no one system is starved or throttled by the others. Ordinarily this means that the I/O systems must be adapted to match both the capabilities of the compute system (since it is generally the least flexible) and of each other. In a modern center there is normally a hierarchy of storage, moving from fastest and most expensive to largest and cheapest. The normal progression for today's supercomputer is from the main memory of the compute engine (typically on the order of a TB), to the local disk (~10TB), to mass storage (~100TB), although this is a moving target, seemingly enormous by yesterday's standards and (presumably) piddling by tomorrow's.

Transfer rates at each interface must be reasonably commensurate, though that will include a degree of degradation as data moves down the performance hierarchy. After all, not all bits produced in the compute engine will require archiving. Normally the input and output rates for any particular subsystems are reasonably, though not completely, symmetric. In this paper we will concentrate on the action of *reading* data from the local disk of the compute engine, and *writing* it to mass storage. Just as modern MPP supercomputers are complex beasts, with inherently great parallelism, so the associated I/O systems must also be complicated by multiple streams to adequately fulfill their requirements.

2.1 The Local File System

The major compute system at SDSC is an IBM SP using 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 GPFS file system. 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 Virtual Shared Disk (VSD) client accessing Recoverable Virtual Shared Disk (RVSD) server software running on 12 server nodes. Each of these servers is a 2-processor, 222 MHz NightHawk1 node forming a recoverable IBM Storage System Architecture (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. A graphical representation of the software and hardware hierarchy is given in figure 1.

Performance is an intricate combination of server and client configurations, together with the details of the application. The server configuration is normally relatively stable, while the application details (number of threads, etc.) and client node count can be varied to reach optimum results.

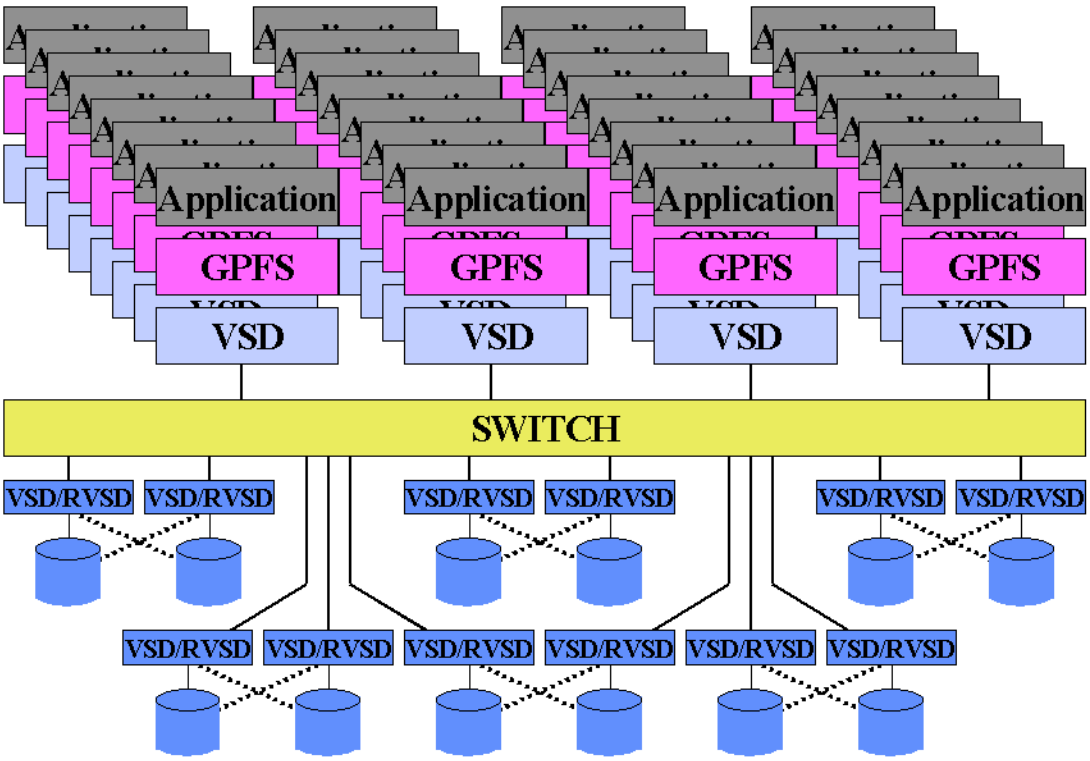

Figure 1. The GPFS file system layers for a 32-node application

2.2 The 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. Figure 2 shows a simplified representation of the HPSS system. Four of the IBM 3590E tape drives are Fibre Channel (FC) connected and are awaiting the installation of a Storage Area Network (SAN) switch.

The normal HPSS operation involves initial writes to cache disk, allowing for rapid movement of small files and caching in case of relatively short term reads. In this case we made all writes directly to tape in order to find the maximum sustainable write rate. While superior number may have been found, they would not have been representative of what the system could sustain over long periods. The file sizes used were in excess of 1GB to reduce the effect of tape mount latencies, and tests were repeated for 2GB files, with virtually the same results.

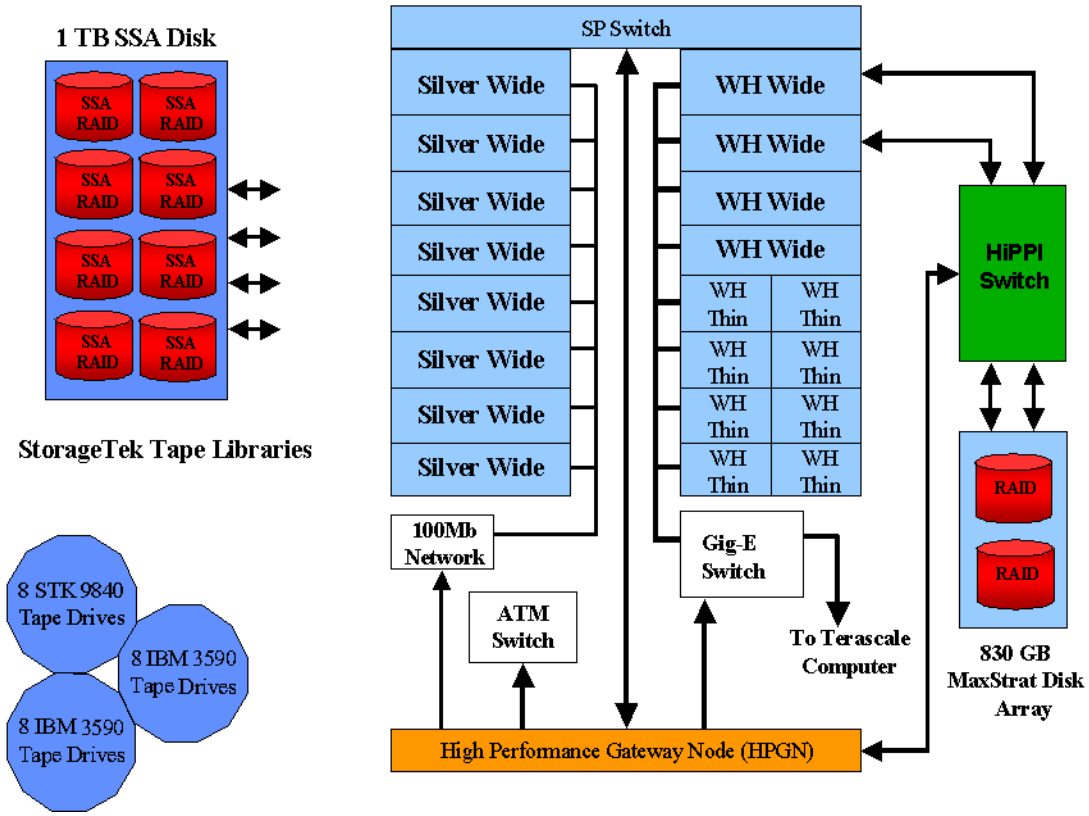

Figure 2. The HPSS mass storage system

2.3 Network Communications

The original communication between the two SP's was by the HPGN, in preparation for the loss of PGN communications when the Colony switch is installed, and to provide higher throughput, we substituted four Gigabit Ethernet connections, via a GbE switch. The IP addresses of the 144 client SP nodes are divided into quadrants, with each quadrant routed over a different GbE connection, each reaching a separate WinterHawk node in HPSS. This allows loadlevelling for applications, while a backup connection from each NightHawk "router" gives redundancy. Figure 3 gives a "birds' eye view" of the operation. We are presently using both standard 1500 byte packets and "jumbo" (~9K) packets. At our present network usage, limited by tape speeds, there is little difference, but we expect this to change when we install a large amount of FC disk.

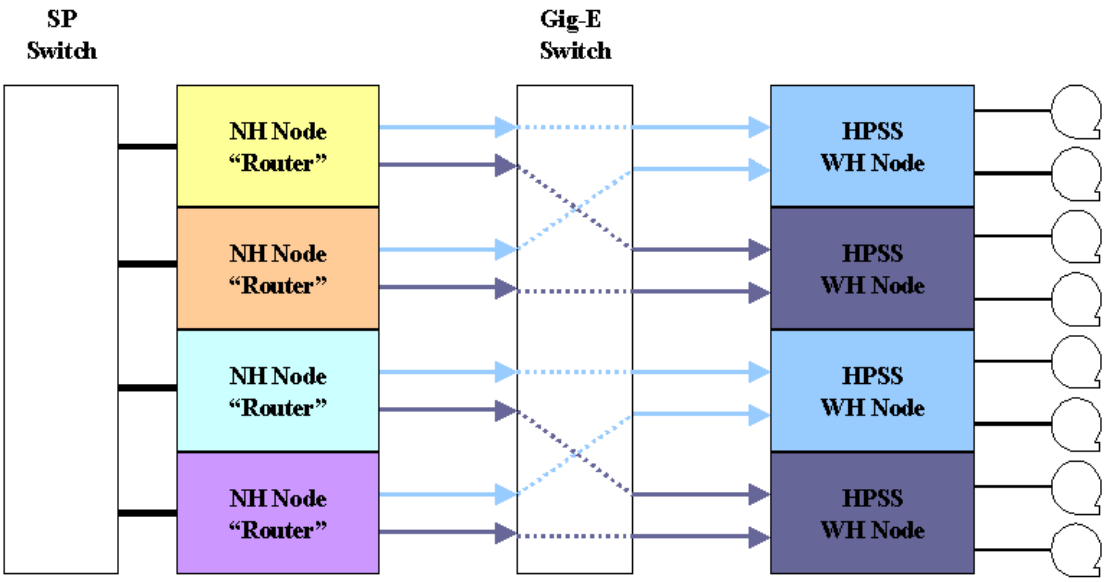

Figure 3. Communications via GbE

3 Performance measurement procedures

For convenience we used the IBM 3590E tape drives for the HPSS writes, in all cases going directly to tape rather than utilizing disk cache. The individual performance of these drives is nominally 14 MB/s, while hardware compression will lead to some variance in the observed rate.

3.1 GPFS Read performance with the IBM "Trailblazer" switch

The first set of experiments were done using IBM's Trailblazer switch to connect the nodes within the compute SP. Nominal transfer rate for this switch is 150 MB/s. We covered a wide area of parameter space, considering different file sizes, different numbers of client nodes, and different numbers of threads per node. In figure 4 we display the results for 8 threads/node for 4 to 128 nodes. The maximum performance was achieved with 64 nodes at just over a gigabyte per second. We plot both the measured performance (series 5) and the theoretical linear scaling (series 6) to show that the achieved performance scales roughly logarithmically with the number of client nodes.

These tests were done for a very wide array of file sizes, threads per node, etc., for space reasons we show only one graph here.

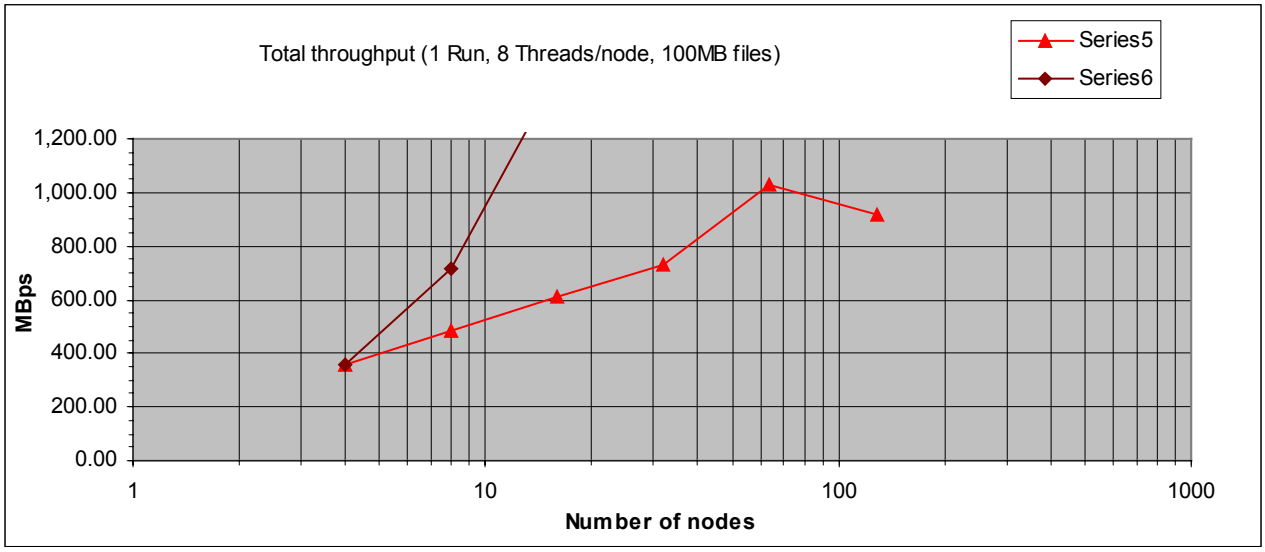

Figure 4. GPFS read performance with the trailblazer switch

3.2 GPFS Read performance with the IBM "Colony" switch

In January of this year we converted the compute SP's switch from Trailblazer to Colony. The new switch has a nominal transfer rate of roughly triple (~450 MB/s) the old one's. The previous tests were rerun, with both the peak and average rates being retained over many runs. Significantly better performance was observed, with an average read rate that peaked at almost 1.3 GB/s with 128 client nodes reading at once. Even more impressively, the average rate had risen to approximately 1.15 GB/s with only 16 client nodes. In figure 5 we display the results for 8 threads/node for 1 to 128 nodes

3.3 HPSS write performance

We show one table, looking at the aggregate performance of various numbers of stripes, originating from one, two, or four nodes.

Transfer type	Sparse file	Uncompressible file	Scientific data
One one-way	16.4 MB/s	11.4 MB/s	16.3 MB/s
One two-way	29 MB/s	23.7 MB/s	25.5 MB/s
Two two-ways	52.1 MB/s	45.5 MB/s	50.4 MB/s
Four two-ways	108 MB/s	89.6 MB/s	106.3 MB/s
One eight-way	36.6 MB/s	30.8 MB/s	31.7 MB/s

Table 1. HPSS write performance

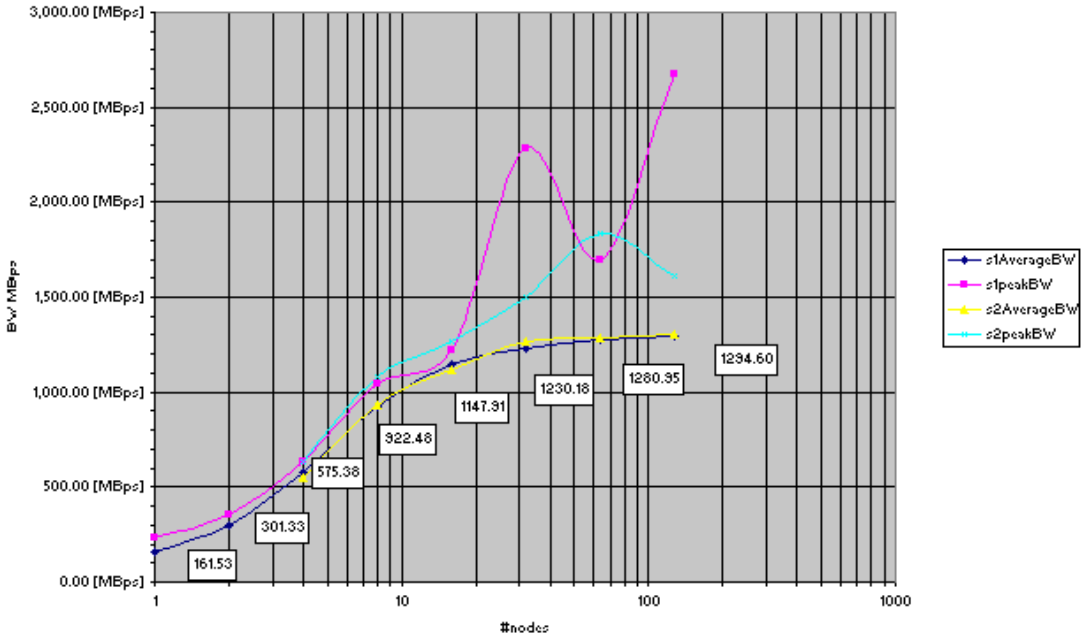

Figure 5. GPFS read performance with the colony switch

4 Future Work

We have upgraded communications recently to use jumbo (~9K) packets over Gigabit Ethernet. This provided only a small improvement in HPSS write rates to tape, but as we install a significant amount of Fibre Channel based disk into HPSS we will approach the nominal capabilities of the network connectivity. This should be done in the next few months. We hope to exceed 2 Gb/s from the compute SP to the HPSS disk cache. We also plan to add more tape drives and explore direct SAN transfers.

5 Conclusions

We have shown that a scientific supercomputer can also be thought of as a super I/O engine, capable of reading locally stored data a 1GB/s and shipping it to tape at speeds in excess of 1Gb/s. As the direction of High Performance Computing moves towards WWW capabilities, these figures of merit will rank with the floating point performance.