A Centralized Data Access Model for Grid Computing

Phil Andrews
San Diego Supercomputer Center
University of California, San Diego
La Jolla, Ca 92093-0505
andrews@sdsc.edu

Tom Sherwin
San Diego Supercomputer Center
University of California, San Diego
La Jolla, Ca 92093-0505
sherwint@sdsc.edu

Bryan Banister
San Diego Supercomputer Center
University of California, San Diego
La Jolla, Ca 92093-0505
bryan@sdsc.edu

Abstract

Global access to storage is a common theme of Grid Computing, with access mechanisms often enforcing a major restriction on the distribution of significant applications across a computational grid. The established approach is to distribute the data with the jobs, sometimes requiring lengthy delays on job completion and the necessity for significant resource discovery to establish local data capabilities. For applications that are truly data intensive, this may render them either highly inefficient, or even incapable of using grid computing environments.

In this paper we describe a different approach, where the opportunity to design a grid environment from scratch is used to build a tightly-coupled, data-oriented infrastructure that leverages deep investment in leading edge technology to provide very high-speed, widespread access to large data storage. Results from a geographically distributed Grid established for the Supercomputing 2002 conference, using preliminary TeraGrid infrastructure, are included and show encouraging performance including data transfer rates of over 700 MB/s using eight 1 Gb/s links from a Storage Area Network to a 10 Gb/s Wide Area Network

.

1.1 Introduction

The San Diego Supercomputer Center (SDSC) is, along with the National Center for Supercomputing Applications (NCSA), California Institute of Technology (CIT), Argonne National Laboratory, and Pittsburgh Supercomputing Center (PSC), a member of the NSF-funded TeraGrid. Basic infrastructure and networking connections, together with specialties of the sites are shown in Fig. 1. The networking backbone is 40 Gb/s with each of the five sites connected at 30 GB/s. The principal computational resources are at NCSA, PSC and SDSC, and the main data repository is at SDSC.

In addition to the approximately 500 TB of rotating storage planned for installation at SDSC by the end of 2003, there is already an archival capacity of 6 PB (uncompressed) provided by 5 STK Powderhorn silos and 24 STK 9940B tape drives. Additional tape drives are 20 IBM 3590E and 8 STK 9840 systems. Presently 30 TB of Fibre Channel disks (Sun T3B RAID sets) are installed in a Storage Area Network using 3 Brocade 12000 FC switches. The tape drives are also on the SAN, and both tape and disk drives are managed by a 64 processor, 256 GB memory, Sun F15K.

It is anticipated that SDSC will host several large datasets (10-50 TB or more) for read-only use by the TeraGrid Computational resources. Several access mechanisms have been proposed for the TeraGrid that would allow the various sites to use this data in either distributed or single-site jobs, including FTP, GridFTP, NFS and SRB. In this paper we discuss the data management details of the SDSC site and a further access mechanism based on extending the SAN across the Wide Area Network.

1.2 Motivation

The conventional approach (conventional in the sense of assumed, rather than actually in widespread use) to handling the data requirements of grid-specific jobs is that before the job starts to run on whichever system it has landed, it first retrieves all the data it requires to run the job from some central location, moving it to local disk resources, and then proceeds to compute. While this is probably viable for small jobs performing cycle scavenging on, e.g., campus networks, it faces major

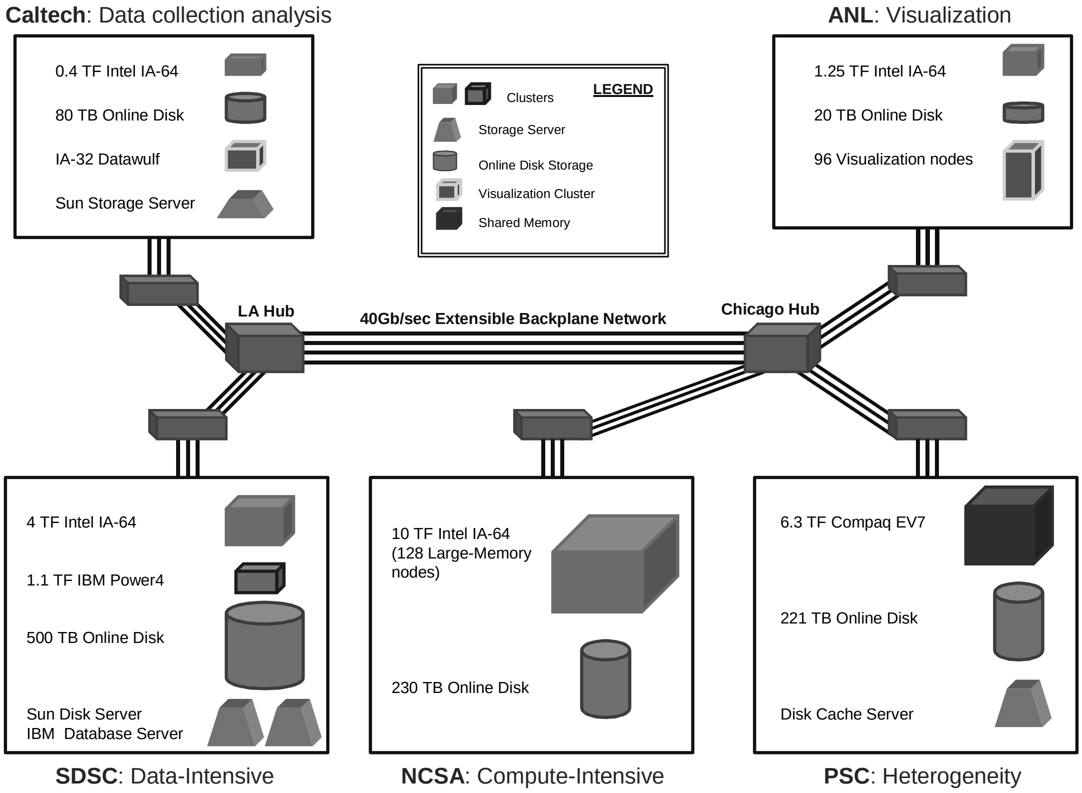

Figure 1. The TeraGrid networking and computational infrastructure.

problems for large jobs of the supercomputing ilk. Many of these applications perform work on very large datasets, of the order of 10 TB or more, and the time required to transfer that much data would likely be large, wasting many resources at the local site. In addition, the local storage may be incapable of absorbing such a large amount of data, eliminating even the possibility of running the job. Even if these problems can be overcome, job submission and setup becomes considerably more complicated and may well deter users from Grid Computing.

In this alternate approach, we envision the data as never being moved en masse, with a single central site being the data repository for the whole grid. The disk is Fibre Channel attached to a Storage Area Network, and the SAN is exported across the Wide Area Network by encoding the Fibre Channel frames within another protocol. The Grid job then uses the same access mechanism (simple file opens and reads) wherever on the Grid it is running and only the pieces of data actually used are shipped across the network on demand. Data access is then transparent for the user, and local data resource discovery is not required. WAN transfer rates are now comparable to SAN rates, and within Supercomputing large sequential accesses are the norm, so that unavoidable latencies should not be crippling to user applications. There remain concerns about latency effects on system-level software, and we explored this in some of the experiments described in this paper.

2. Local File Storage and Performance.

2.1 Disk to Disk performance.

Although the Sun F15K is the data manager for the SAN disk, we expect that the most important performance criteria will be read rates to other computers accessing the data across the Storage Area Network. Accordingly, we benchmarked the transfer rates to a second Sun system using the Tivoli SANergy software to access the data

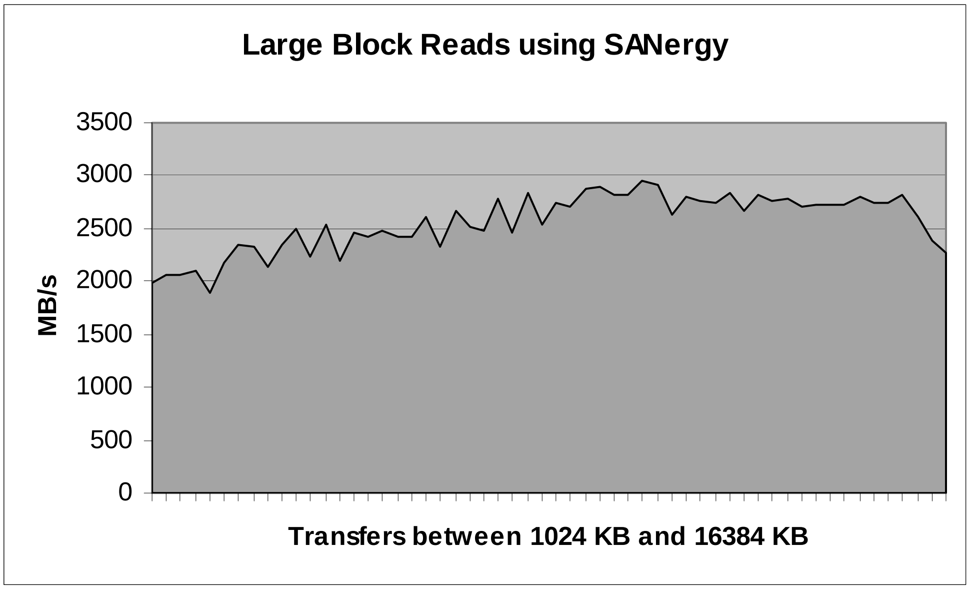

Figure 2. File system performance.

directly via the SAN. For this we used sixteen 2 Gb Fibre Channel Host Bus Adapters in the Sun 6800 reading the data, and thirty-two 1 Gb Sun T3B FC RAID sets, connected across the 2 Gb Brocade 12000 FC switches. Sun's QFS was used. This is a highly parallel, high performance file system

The FC standard actually allows 1.07 Gb/s to be transferred on a 1 Gb port, but the FC frames add 2 bits for every 8 bits of payload, so that the maximum data transfer rate per port is 107 MB/s. The theoretical maximum for the configuration tested was thus 3,424 MB/s. In the event, a peak performance of 3,200 MB/s was recorded and as Fig 2 shows, sustained performance of around 3 GB/s was achieved.

2.2 Disk to Tape performance

Our intention is to present the SAN disk to users as an apparently limitless disk cache, using Sun's SAM-FS software to manage transfers between disk and tape, automatically migrating files from disk to tape as needed to free up space for new files being written to disk or recalled from tape. All Inodes are retained on disk and the QFS file system appears as a normal disk repository to users. For this approach to work, excellent file transfer rates between disk and tape are essential and we tested aggregate rates using 25 STK 9940B tape drives backing up 7 TB of SAN disk. The individual parameters of the 9940B tape drives are 30 MB/s transfer rates and 200 GB per cartridges, both numbers representing uncompressed data.

Only preliminary results are available at the moment, but we were able achieve a peak performance of 828 MB/s, and as Fig 3 shows, sustained numbers around 800 MB/s. With more disks available, and some tuning of parameters, we hope to reach 1 GB/s in the near future with representative scientific data files.

2.3 Presentation to Users

With approximately 1 GB/s transfer rate available between tape and disk media, we believe it should be possible to provide acceptable performance for all users without the necessity of requiring disk allocations. In return for no guarantee of disk residence, the users will be presented with an apparently inexhaustible supply of storage space. What is in fact a very large investment in storage infrastructure should appear to be an essentially unlimited pool of online storage. Compared to earlier attempts, the sheer size of this system allows statistics, especially the laws of large numbers, to help us in covering most eventualities.

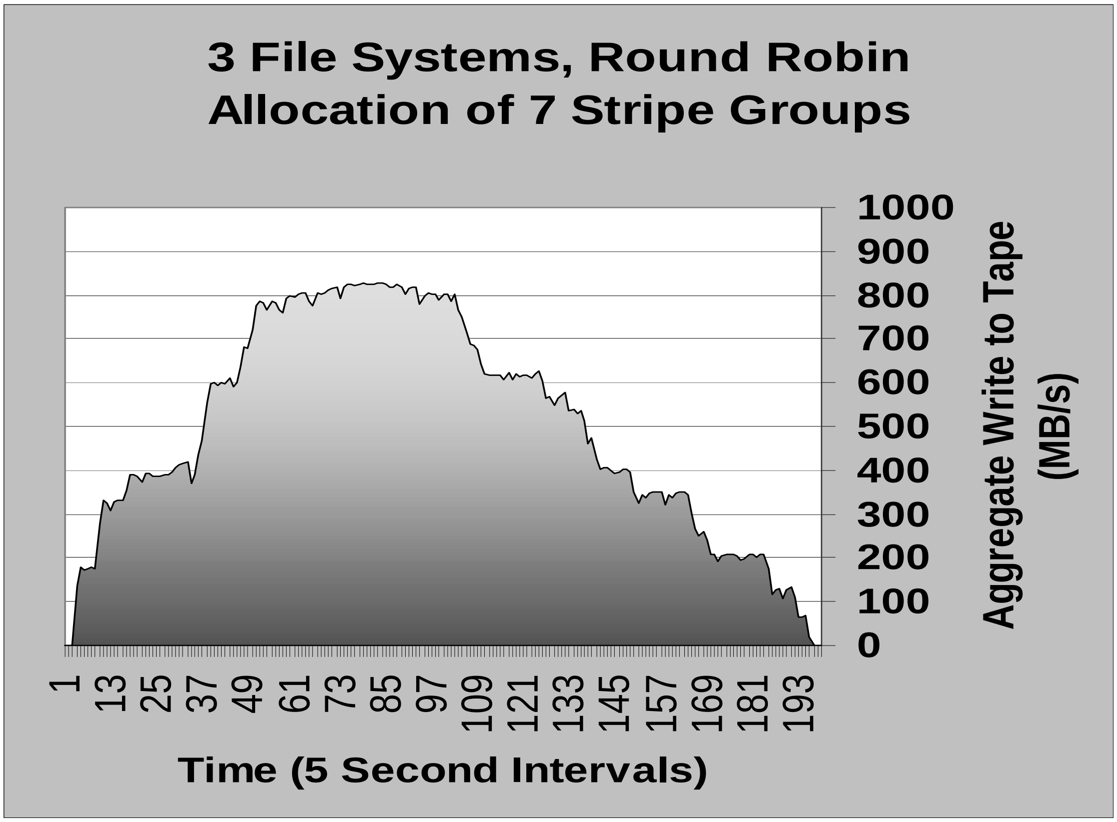

Figure 3. Disk to Tape performance, using SAM-FS.

3. Extending the Storage Area Network

3.1 Communication across the Wide Area Network

While we are establishing an effectively “bottomless pit” of online storage, the question remains as to how users without direct access to the SDSC storage area network will use it? The proposal for the original Distributed Terascale Facility (DTF) assumed that Wide Area Network access would be the only option for non-local users, and that will always be true for sufficiently remote sites. However, we have the unique opportunity of designing an extremely high-speed network from scratch and have sufficient bandwidth available to contemplate extending the SDSC SAN across the TeraGrid backbone using FCIP or other methods. In FCIP, the Fibre Channel frames are encoded in IP packets, which can then be shipped across the WAN to geographically remote sites. They are then decoded and passed to the local FC environment, allowing the two SANs to become part of the same fabric. Whereas ISCSI (encoding SCSI commands within IP) allows a remote server to access storage, FCIP allows the actual extension of the SAN fabric. FCIP has already been used for remote mirroring operations, but at relatively low speeds (e.g., OC3), while we plan on approximate 1 GB/s transfer rates.

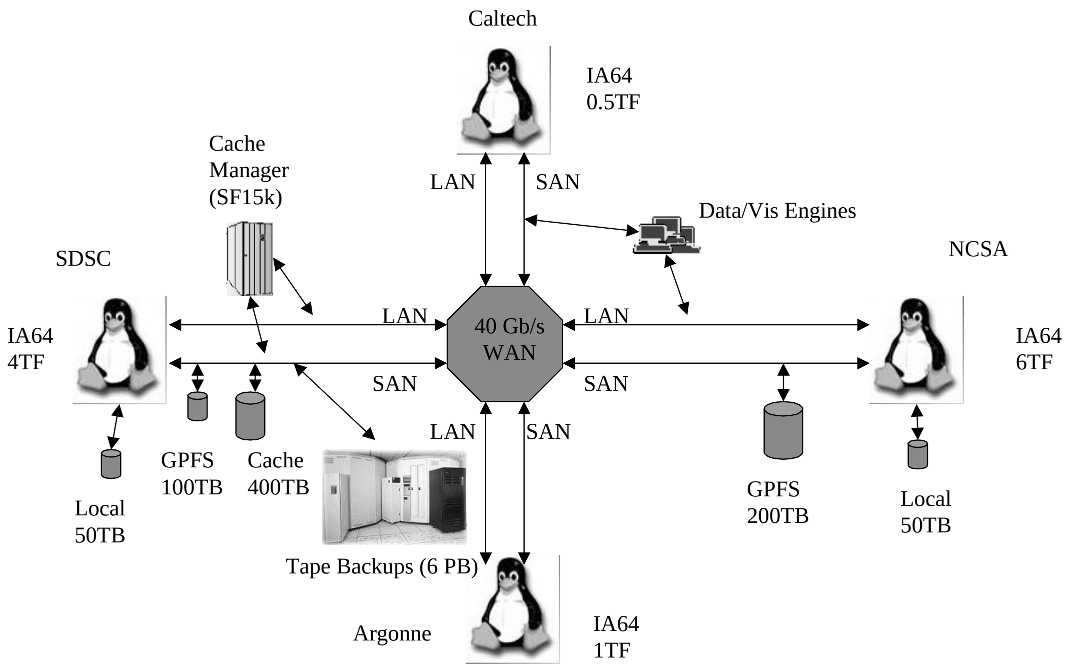

Figure 4. Schematic of Data Access for the Distributed Terascale Facility

3.1.1 Fibre Channel over IP experiment

At the recent Supercomputing ’02 meeting in Baltimore (Nov 17-22, 2002) we were able to perform a proof of principle experiment between the SDSC machine room in San Diego, California, and the SDSC booth on the show room floor in Baltimore, Maryland. In this case the existing TeraGrid network was used to provide a 10 Gb/s connection from San Diego to Chicago, Il, and then extended to Baltimore.

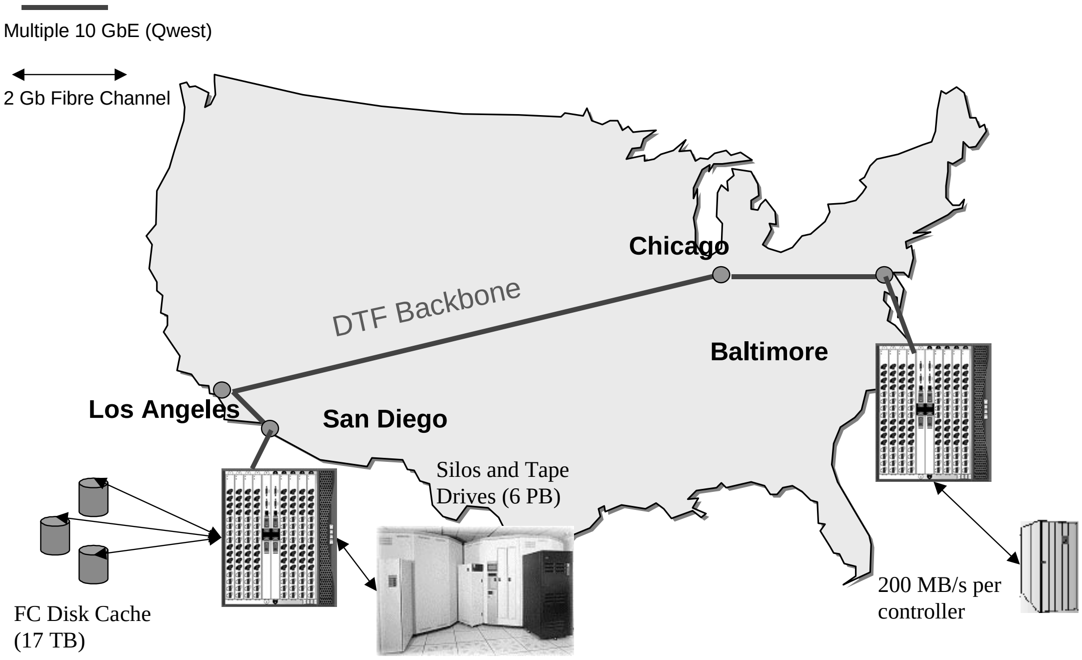

Figure 5. Schematic of Network Connectivity for the SC'02 demonstration

In the SDSC machine room two Nishan IPS 4000 boxes were used to take eight 1 Gb/s FC connections and multiplex them into the 10 Gb/s IP connections from SDSC to the Baltimore show floor. Connectivity from the IPS 4000 systems to the WAN was via a Force10 12000 Gigabit Ethernet switch and a Juniper T640 router. In the SDSC booth at SC'02, two more IPS 4000 systems were used to connect to the FC SAN fabric locally. Various configurations were tried, including connections to Brocade 12000 FC SAN switches at each end and direct connections to disks and servers. The server in the SDSC booth was a Sun SF 6800.

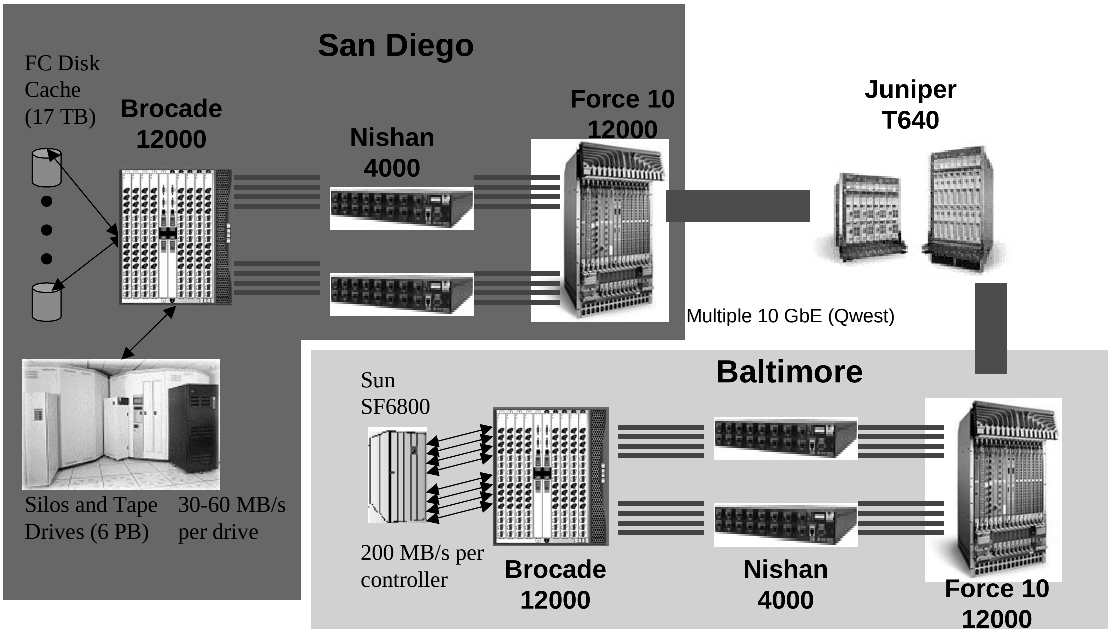

Figure 6. Schematic of Data Access for the SC'02 demonstration

3.1.2 Performance results

One of the main motivations behind this experiment was to see what effect the latency would have on performance. San Diego to Baltimore is in excess of 2,600 miles, or 43,000 Kilometers, more than the physical extent of the TeraGrid. Measured round-trip latency was between 70 and 90 milliseconds, but relatively constant at about 80 milliseconds. Initially, this led to some problems, but these were resolved with some adjustments to the FC switches.

Maximum possible throughput would have been about 800 MB/s and recorded transfer rates improved over the 4 days of the conference. As expected, read performance was slightly better than writes and a graph of transfer rates from 8 individual RAID sets across the WAN, but using FC SAN access mechanisms, are shown in Fig. 6. Individual channels were reliably in the 95 MB/s range while the aggregate performance was relatively constant at 717 MB/s. These were disk to memory transfers, looking for the greatest possible performance.

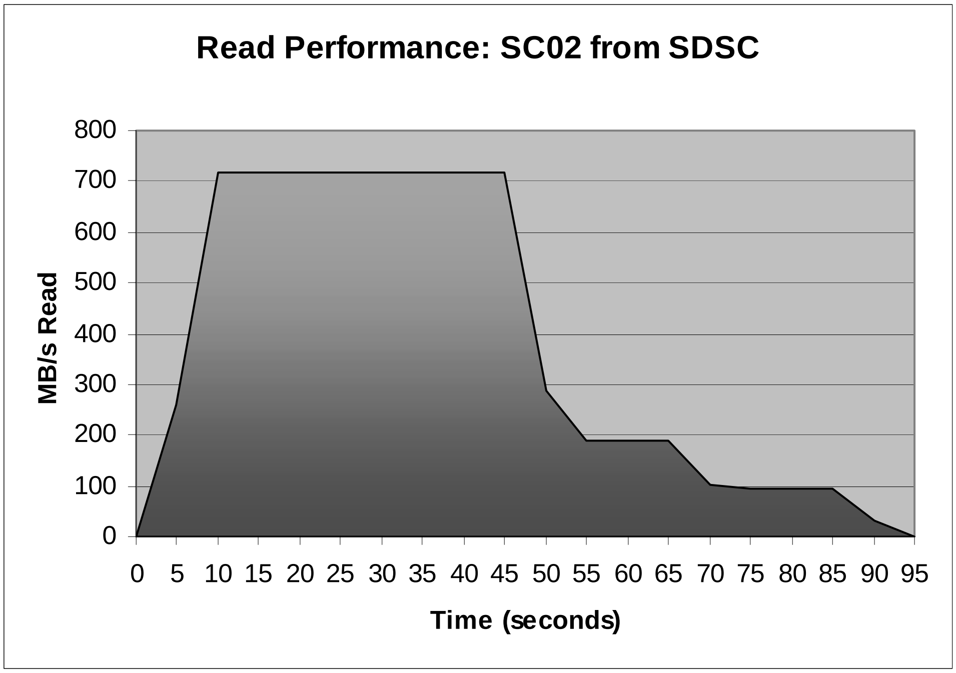

Figure 7. Read performance from San Diego to Baltimore

In addition to reads from San Diego, the same configuration was used for writes, and the corresponding performance chart is shown in Fig. 7. Gratifyingly, the read performance is very close to the writes, with a maximum value of 691 MB/s.

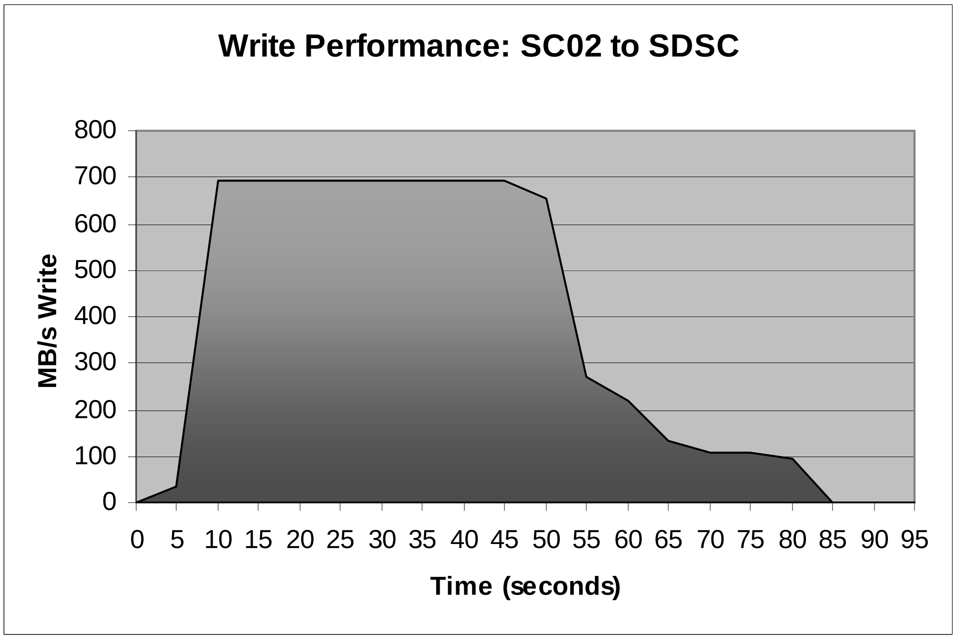

Figure 8. Write performance from Baltimore to San Diego

3.1.3 Fibre Channel over SONET experiment

Another approach would be to use Fibre Channel over SONET, which has both advantages and disadvantages. The protocol stack for FC over SONET is much simpler, with only two layers, offering the promise of greater performance and lower latencies. However, it is more difficult to share the network, and since we had to support conventional IP traffic between the SDSC machine room and SC'02 in addition to encoded FC frames, it was judged too difficult to explore FC over SONET for that route. Instead we arranged for a dedicated fibre link on the SC'02 show floor between the SDSC booth and the Pittsburgh Supercomputing booth. We ran SONET across the Fibre and used Akara systems to encode the FC frames. In the SDSC booth, the FC frames were connected to the local SAN fabric, which was globally connected to the SDSC machine room via FC over IP. Thus, communication between the PSC booth and the SDSC machine room (which all appeared on the same Storage Area Network) was via two encoding mechanisms, FC over SONET between the show booths and FC over IP from Baltimore to San Diego.

In this case, there was insufficient equipment in the PSC booth to fully test out transfer rate compatibility, so we concentrated on establishing capability. The Sun workstation in the PSC booth saw the QFS file system in the SDSC machine room, backed up by SAM-FS, as a local file system and successfully transferred data to that file system and ultimately to the tape library. Thus we had an apparently limitless data source/sink appearing purely local across 2,500 miles.

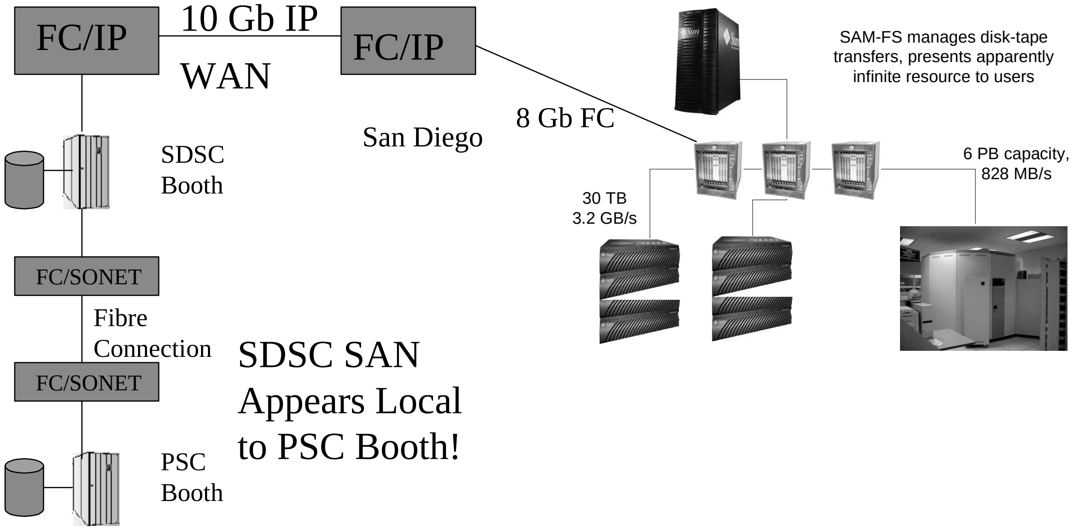

Figure 9. FC/IP and FC/SONET connections for the SC'02 demonstration

3.2 File System and Other Issues

One difficulty to be overcome is exporting the data across Operating Systems. We are dealing with three in this case: Solaris (Sun), AIX (IBM), and Linux (64 bit). Solaris to Solaris is just a matter of having QFS installed on each system, for Solaris to AIX we are using the Tivoli SANergy product to export the data, and for Solaris to Linux we are involved in a joint research project with Sun to export the QFS client to Linux. In addition we are exploring the possibility of exporting IBM’s GPFS file system across a WAN-San and examining the Lustre file system from Hewlett-Packard. As well as moving from one operating system to another, there can be problems of byte ordering within the data. These can all be overcome, but must be watched carefully to avoid adversely impacting performance.

Another problematic area is in authentication and access control. In a very tightly coupled grid partnership, such as NPACI (http://www.npaci.edu) it is possible to require that userids be consistent across all machines. The TeraGrid (like most grids) is formed from pre-existing systems and uniform userids is not an option. Luckily, the dominant means of access to large data sets is in a read-only mode: it is much more likely that a user will read to an existing dataset of the night sky than write to it! Initially, we plan to sidestep the access problems by offering up large datasets in a read-only, globally accessible mode. For a few users that need the capability, it should be possible to synchronize userids. Meanwhile, we are working on certificate-based access systems that we expect to be the future of Grid access controls.

3.3 Future work

In the near future we hope to dedicate one or more lambdas of the TeraGrid network from SDSC to Chicago to examining this approach. This will allow us to use FC over SONET as our transport mechanism and investigate claims of lower latencies and higher transfer rates than FC/IP.

4. Conclusions

We have described a novel approach to Grid data architecture, where, rather than moving data to be near the running jobs, a single stationary site is chosen with sufficient networking bandwidth to make the data appear local to the remote sites. This greatly simplifies resource discovery, job distribution, etc., and may make the difference between viable grid computing and an approach that is too cumbersome to be attractive to the majority of users.

The experiments performed showed that latencies across Wide Area Networks, though unavoidable, do not seem to be crippling, and we used the opportunity of the Supercomputing ’02 meeting in Baltimore to demonstrate this with cross-continental communications between there and San Diego. We also demonstrated excellent transfer rates in excess of 700 MB/s and used multiple encodings to extend the SAN.

.

.