A Data Intensive Distributed Computing Architecture for "Grid" Applications

Brian Tierney, William Johnston, Jason Lee, Mary Thompson
Lawrence Berkeley National Laboratory
Berkeley, CA 94720

Abstract. Modern scientific computing involves organizing, moving, visualizing, and analyzing massive amounts of data from around the world, as well as employing large-scale computation. The distributed systems that solve large-scale problems will always involve aggregating and scheduling many resources. Data must be located and staged, cache and network capacity must be available at the same time as computing capacity, etc. Every aspect of such a system is dynamic: locating and scheduling resources, adapting running application systems to availability and congestion in the middleware and infrastructure, responding to human interaction, etc. The technologies, the middleware services, and the architectures that are used to build useful high-speed, wide area distributed systems, constitute the field of data intensive computing. This paper explores some of the history and future directions of that field, and describes some specific application examples.

1.0 Introduction

High-speed data streams resulting from the operation of on-line instruments and imaging systems are a staple of modern scientific, health care, and intelligence environments. The advent of high-speed networks is providing the potential for new approaches to the collection, organization, storage, analysis, visualization, and distribution of the large-data-objects that result from such data streams. The result will be to make both the data and its analysis much more readily available.

For example, health care imaging systems illustrate the need for both high data rates and real-time cataloging. Medical video and image data used for diagnostic purposes — e.g., X-ray CT, MRI, and cardio-angiography — are collected at centralized facilities and may be accessed at locations other than the point of collection (e.g., the hospitals of the referring physicians). A second example is high energy physics experiments, which generate high rates and massive volumes of data that must be processed and archived in real time. This data must also be accessible to large scientific collaborations — typically hundreds of investigators at dozens of institutions around the world.

In this paper we will describe how "Computational Grid" environments can be used to help with these types of applications, and give specific examples of health care and high energy physics applications in this environment. We describe how a high-speed application-level network cache is a particularly important component in a data intensive grid architecture, and describe our implementation of such a cache.

2.0 Data Intensive Grids

The integration of the various technological approaches being used to address the problem of integrated use of dispersed resources is frequently called a "grid," or a computational grid — a name arising by analogy with the grid that supplies ubiquitous access to electric power. See, e.g., [7]. Basic grid services are those that locate, allocate, coordinate, utilize, and provide for human interaction with the various resources that actually perform useful functions.

Grids are built from collections of primarily independent services. The essential aspect of grid services is that they are uniformly available throughout the distributed environment of the grid. Services may be grouped into integrated sets of services, sometimes called "middleware." Current grid tools include Globus [6], Legion [12], SRB [2], and workbench systems like Habanero [8] and WebFlow [1].

From the application's point of view, the Grid is a collection of middleware services that provide applications with a uniform view of distributed resource components and the mechanisms for assembling them into systems. From the middleware systems points of view, the Grid is a standardized set of basic services providing scheduling, resource discovery, global data directories, security, communication services, etc. However, from the Grid implementor's point of view, these services result from and must interact with a heterogeneous set of capabilities, and frequently involve "drilling" down through the various layers of the computing and communications infrastructure.

2.1 Architecture for Data Intensive Environments

Our model is to use a high-speed data storage cache as a common element for all of the sources and sinks of data involved in high-performance data systems. We use the term "cache" to mean storage that is faster than typical local disk, and temporary in nature. This cache-based approach provides standard interfaces to a large, application-oriented, distributed, on-line, transient storage system.

Each data source deposits its data in the cache, and each data consumer takes data from the cache, often writing the processed data back to the cache. A tertiary storage system manager migrates data to and from the cache at various stages of processing. (See Figure 1.) We have used this model for data handling systems for high energy physics data and for medical imaging data. These applications are described in some detail in [11] and [10].

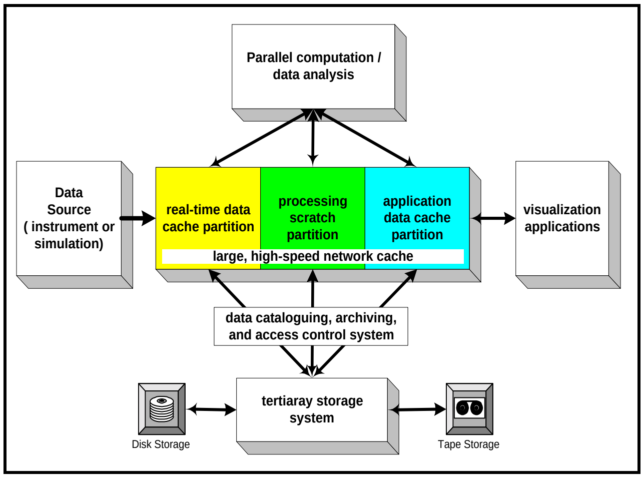

Figure 1 The Data Handling Model

The high-speed application-level cache serves several roles in this environment. It provides a standard high data rate interface for high-speed access by data sources, processing resources, mass storage systems (MSS), and user interface / data visualization elements. It provides the functionality of a single very large, random access, block-oriented I/O device (i.e., a "virtual disk"). This cache also serves to isolate the application from tertiary storage systems and instrument data sources, helping eliminate contention for those resources.

This cache can be used as a large buffer, able to absorb data from a high rate data source and then to forward it to a slower tertiary storage system. The cache also provides an "impedance matching" function between a small number of high throughput streams to a larger number of lower speed streams, e.g. between fine-grained accesses by many applications and the coarse-grained nature of a few parallel tape drives in the tertiary storage system.

Depending on the size of the cache relative to the objects of interest, the tertiary storage system management may only involve moving partial objects to the cache. In other words, the cache may contain a moving window for an extremely large off-line object/data set. Generally, the cache storage configuration is large (e.g., 100s of gigabytes) compared to the available disks of a typical computing environment (e.g., 10s of gigabytes), and very large compared to any single disk (e.g. hundreds of ~10 gigabytes).

In this type of environment, a client typically will copy large portions of a data set from the remote archive to local disk before visualizing or processing the data. However, if the network is fast enough and if the cache is tuned for remote access and uses parallel disks, the cache can provide data access to remote clients that is even faster than local disk. Additionally many applications actually only need a small portion of the total data set. By leaving the data on a remote cache, a much smaller amount of data may actually be moved over the network.

3.0 The Distributed-Parallel Storage System

Our implementation of this high-speed, distributed cache is called the Distributed-Parallel Storage System (DPSS) [16]. LBNL designed and implemented the DPSS as part of the DARPA MAGIC project [4], and as part of the U.S. Department of Energy's high-speed distributed computing program. This technology has been successful in providing an economical, high-performance, widely distributed, and highly scalable architecture for caching large amounts of data that can potentially be used by many different users.

Typical DPSS implementations consist of several low-cost workstations as DPSS block servers, each with several disk controllers, and several disks on each controller. A four-server DPSS with a capacity of one Terabyte (costing about $50K-$80K in mid-1999) can thus produce throughputs of over 50 MBytes/sec by providing parallel access to 20-30 disks. The overall architecture of the DPSS is illustrated in Figure 2.

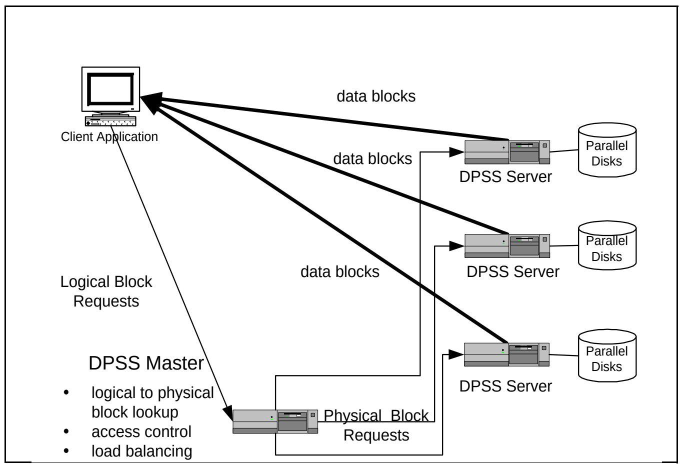

Figure 2 Overall DPSS Architecture

Other papers describing the DPSS in more detail include [16], which describes how the DPSS was used to provide high-speed access to remote data for a terrain visualization application, [17], which describes the basic architecture and implementation, and [18], which describes how the instrumentation abilities in the DPSS were used to help track down a wide area network problem.

The application interface to the DPSS cache supports a variety of I/O semantics, including Unix-like I/O semantics, through an easy to use client API library (e.g. dpssOpen(), dpssRead(), dpssWrite(), dpssLSeek(), dpssClose()). The data layout on the disks is completely up to the application, and the usual strategy for sequential reading applications is to write the data "round-robin," striping blocks of data across the servers. The client library also includes a flexible data replication ability, allowing for multiple levels of fault tolerance. The DPSS client library is multi-threaded, where the number of client threads is equal to the number of DPSS servers. Therefore the speed of the client scales with the speed of the server, assuming the client host is powerful enough.

The internal architecture of the DPSS is illustrated in Figure 3. Requests for blocks of data are sent from the client to the "DPSS master" process, which determines which "DPSS block servers" the blocks are located on, and forwards the requests to the appropriate servers. The server then sends the block directly back to the client. Servers may be anywhere in the network: there is no assumption that they are all at the same location, or even the same city.

DPSS performance, as measured by total throughput, is optimized for a relatively smaller number (a few thousand) of relatively large files (greater than 50 MB). Performance is the same for any file sizes greater than 50 MB. We have also shown that performance scales well with the number of clients, up to at least 64 clients. For example, if the DPSS system is configured to provide 50 MB/sec to 1 client, it can provide 1 MB/sec to each of 50 simultaneous clients. The DPSS master host starts to run out of resources with more than 64 clients.

Because of the threaded nature of the DPSS server, a server scales linearly with the number of disks, up to the network limit of the host (possibly limited by the network card or the CPU). The total DPSS system throughput scales linearly with the number of servers, up to at least 10 servers.

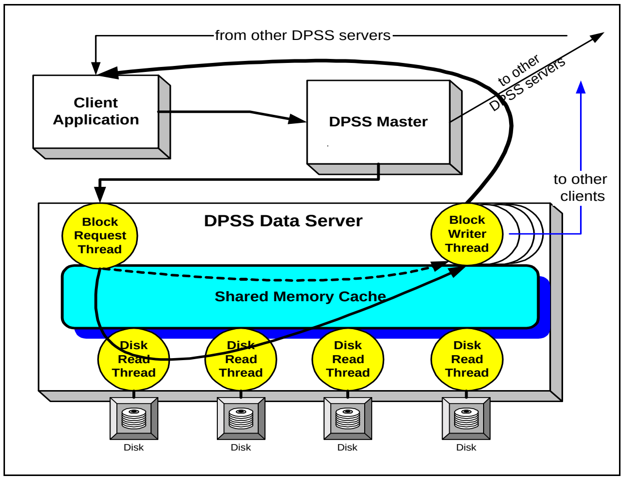

Figure 3 DPSS Server Architecture

The DPSS provides several important and unique capabilities for data intensive distributed computing environments. It provides application-specific interfaces to an extremely large space of logical blocks; it offers the ability to build large, high-performance storage systems from inexpensive commodity components; and it offers the ability to increase performance by increasing the number of parallel disk servers.

DPSS data blocks are available to clients immediately as they are placed into the cache. It is not necessary to wait until the entire file has been transferred before requesting data. This is particularly useful to clients requesting data from a tape archive. As the file moves from tape to the DPSS cache, the blocks in the cache are immediately available to the client. If a block is not available, the application can either block, waiting for the data to arrive, or continue to request other blocks of data which may be ready to read.

The DPSS is dynamically reconfigurable, allowing one to add or remove servers or disks on the fly. This is done by storing the DPSS hardware resource information in a Globus Metacomputing Directory Service (MDS)[3] formatted LDAP database, which may be updated dynamically. Software agents are used to monitor network, host, and disk availability and load, storing this information into the LDAP database as well. This information can then be used for fault tolerance and load balancing.

4.0 A Medical Application Example

BAGNet was an IP over OC-3 (155 Mbit/sec) ATM, metropolitan area network testbed that operated in the San Francisco Bay Area (California) for two years starting in early 1994. The participants included government, academic, and industry computer science and telecommunications R&D groups from fifteen Bay Area organizations. The goal was to develop and deploy the infrastructure needed to support a diverse set of distributed applications in a large-scale, IP-over-ATM network environment.

In BAGNet, there were several specific projects involving subsets of the connected sites. In particular, LBNL, the Kaiser Permanente health care organization, and Philips Palo Alto Research Center collaborated to produce a prototype production, on-line, distributed, high data rate medical imaging system.

The Kaiser project [9] focused on using high data rate, on-line instrument systems as remote data sources.

When data is generated in large volumes and with high throughput, and especially in a distributed environment where the people generating the data are geographically separated from the people cataloguing or using the data, there are several important considerations for managing instrument generated data:

- automatic generation of at least minimal metadata;
- automatic cataloguing of the data and the metadata as the data is received (or as close to real time as possible);
- transparent management of tertiary storage systems where the original data is archived;
- facilitation of cooperative research by providing specified users at local and remote sites immediate as well as long-term access to the data;

♦ mechanisms to incorporate the data into other databases or documents.

The WALDO (Wide-area Large-Data-Object) system was developed to provide these capabilities, especially when the data is gathered in real time from a high data rate instrument [9]. WALDO is a digital data archive that is optimized to handle real-time data. It federates textual and URL linked metadata to represent the characteristics of large data sets. Automatic cataloguing of incoming real-time data is accomplished by extracting associated metadata and converting it into text records; by generating auxiliary metadata and derived data; and by combining these into Web-based objects that include persistent references to the original data components (called large data objects, or LDOs) [15]. Tertiary storage management for the data components (i.e., the original datasets) is accomplished by using the remote program execution capability of Web servers to manage the data on mass storage systems. For subsequent use, the data components may be staged to a local disk and then returned as usual via the Web browser, or, as is the case for several of our applications, moved to a high-speed cache for access by specialized applications (e.g., the high-speed video player illustrated in the right-hand part of the right-hand panel in Figure 4). The location of the data components on tertiary storage, how to access them, and other descriptive material are all part of the LDO definition. The creation of object definitions, the inclusion of “standardized” derived-data-objects as part of the metadata, and the use of typed links in the object definition, are intended to provide a general framework for dealing with many different types of data, including, for example, abstract instrument data and multi-component multimedia programs. WALDO was used in the Kaiser project to build a medical application that automatically manages the collection, storage, cataloguing, and playback of video-angiography data[1] collected at a hospital remote from the referring physician.

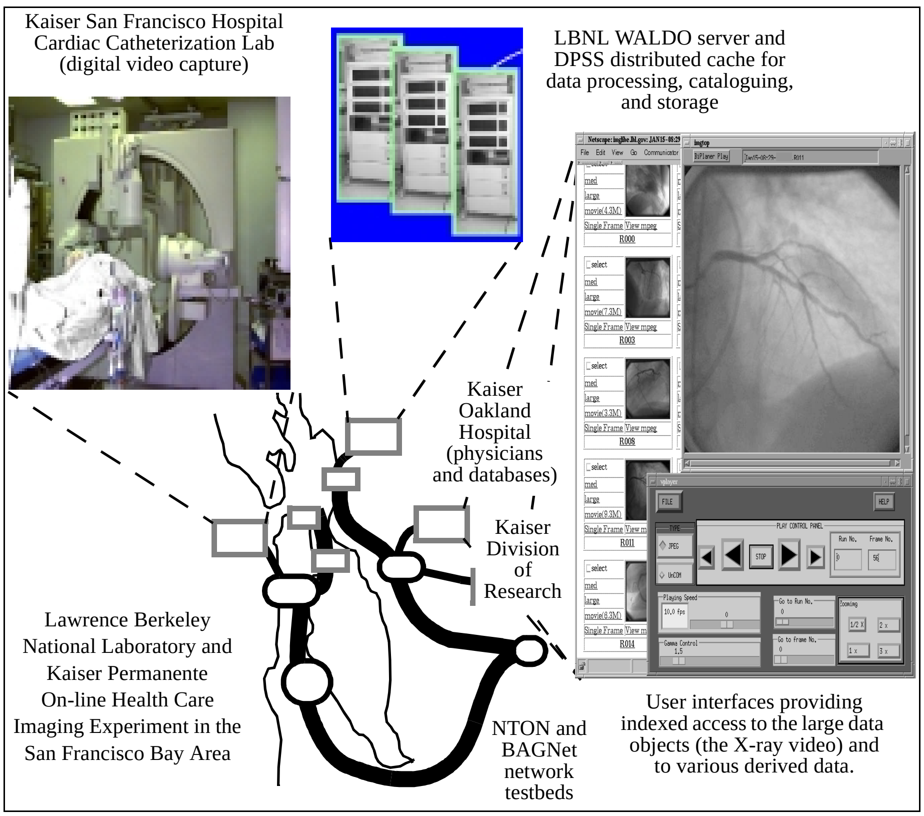

Figure 4 Capture of Digitized Angiograms

Using a shared, metropolitan area ATM network and a high-speed distributed data handling system, video sequences are collected from the video-angiography imaging system, then processed, catalogued, stored, and made available to remote users. This permits the data to be made available in near-real time to remote clinics (see Figure 4). The LDO becomes available as soon as the catalogue entry is generated — derived data is added as the processing required to produce it

1. Cardio-angiography imaging involves a two plane, X-ray video imaging system that produces from several to tens of minutes of digital video sequences for each patient study for each patient session. The digital video is organized as tens of data-objects, each of which are of the order of 100 megabytes.

completes. Whether the storage systems are local or distributed around the network is entirely a function of optimizing logistics.

In the Kaiser project, cardio-angiography data was collected directly from a Philips scanner by a computer system in the San Francisco Kaiser hospital Cardiac Catheterization Laboratory. This system is, in turn, attached to an ATM network provided by the NTON and BAGNet testbeds. When the data collection for a patient is complete (about once every 20–40 minutes), 500–1000 megabytes of digital video data is sent across the ATM network to LBNL (in Berkeley) and stored first on the DPSS distributed cache (described above), and then the WALDO object definitions are generated and made available to physicians in other Kaiser hospitals via BAGNet. Auxiliary processing and archiving to one or more mass storage systems proceeds independently. This process goes on 8–10 hours a day.

5.0 A High Energy Nuclear Physics Application

We have conducted a set of high-speed, network based, data intensive computing experiments between Lawrence Berkeley National Laboratory (LBNL) in Berkeley, Calif., and the Stanford Linear Accelerator (SLAC) in Palo Alto, Calif. The results of this experiment were that a sustained 57 megabytes/sec of data were delivered from datasets in the distributed cache to the remote application memory, ready for analysis algorithms to commence operation. This experiment represents an example of our data intensive computing model in operation.

The prototype application was the STAR analysis system that analyzes data from high energy physics experiments. (See [5].) A four-server DPSS located at LBNL was used as a prototype front end for a high-speed mass storage system. A 4-CPU Sun E-4000 located at SLAC was a prototype for a physics data analysis computing cluster, as shown in Figure 5.

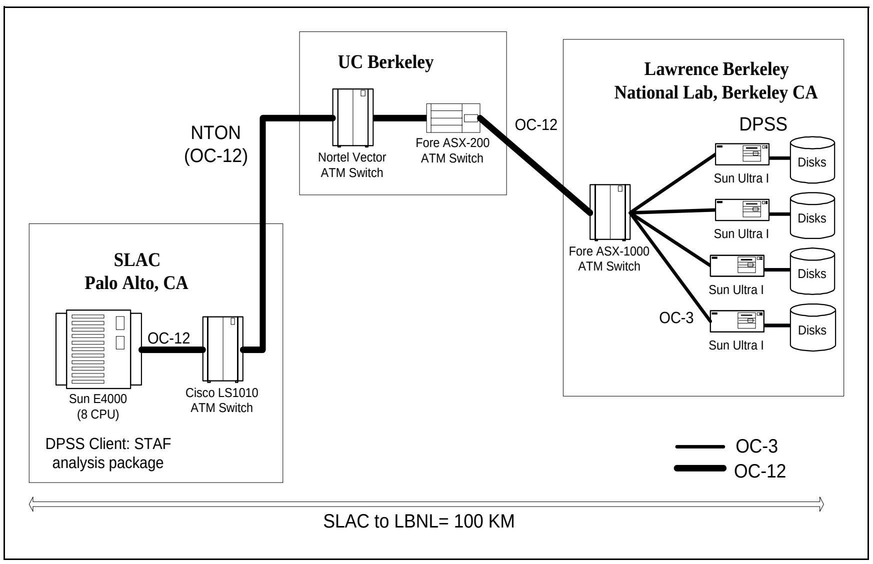

Figure 5 HENP Experiment Configuration

The National Transparent Optical Network testbed (NTON - see [13]) connects LBNL and SLAC and provided a five-switch, 100-km, OC-12 ATM path. All experiments were application-to-application, using TCP transport.

Multiple instances of the STAR analysis code read data from the DPSS at LBNL and moved that data into the memory of the STAF application where it was available to the analysis algorithms. This experiment resulted in a sustained data transfer rate of 57 MBytes/sec from DPSS cache to application memory. This is the equivalent of about 4.5 TeraBytes / day. The goal of the experiment was to demonstrate that high-speed mass storage systems could use distributed application-level caches to make data available to the systems running the analysis codes. The experiment was successful, and the next steps will involve completing the mechanisms for optimizing the MSS staging patterns and completing the DPSS interface to the bit file movers that interface to the MSS tape drives.

6.0 Conclusions

We believe this architecture, and its integration with systems like Globus, will enable the next generation of configurable, distributed, high-performance, data-intensive systems; computational steering; and integrated instrument and computational simulation. We also believe a high performance network application-level cache system such as the DPSS will be an important component to these "computational grid" environments.

7.0 Acknowledgments

The work described in this paper is supported by the U. S. Dept. of Energy, Office of Energy Research, Office of Computational and Technology Research, Mathematical, Information, and Computational Sciences and ERLTT Divisions under contract DE-AC03-76SF00098 with the University of California, and by DARPA, Information Technology Office.

8.0 References

[1] Erol Akarsu, Geoffrey C. Fox, Wojtek Furmanski, Tomasz Haupt, "WebFlow - High-Level Programming Environment and Visual Authoring Toolkit for High Performance Distributed Computing," Proceedings of IEEE Supercomputing '98, Nov. 1998.

[2] Chaitanya Baru, Reagan Moore, Arcot Rajasekar, Michael Wan, "The SDSC Storage Resource Broker," Proc. CASCON'98 Conference, Nov.30-Dec.3, 1998, Toronto, Canada. (http://www.npaci.edu/DICE/SRB/)

[3] K. Czajkowski, I. Foster, C., Kesselman, N. Karonis, S. Martin, W. Smith, S. Tuecke. "A Resource Management Architecture for Metacomputing Systems," Proc. IPPS/SPDP '98 Workshop on Job Scheduling Strategies for Parallel Processing,1998.

[4] B. Fuller and I. Richer "The MAGIC Project: From Vision to Reality," IEEE Network, May, 1996, Vol. 10, no. 3. http://www.magic.net/

[5] W. Greiman, W. E. Johnston, C. McParland, D. Olson, B. Tierney, C. Tull, "High-Speed Distributed Data Handling for HENP," Computing in High Energy Physics, April, 1997. Berlin, Germany. http://www-itg.lbl.gov/STAR/

[6] Globus: See http://www.globus.org

[7] Grid: *The Grid: Blueprint for a New Computing Infrastructure*, edited by Ian Foster and Carl Kesselman. Morgan Kaufmann, Pub. August 1998. ISBN 1-55860-475-8. http://www.mkp.com/books_catalog/1-55860-475-8.asp

[8] Habanero: http://www.ncsa.uiuc.edu/SDG/Software/ Habanero/

[9] W. Johnston, G. Jin, C. Larsen, J. Lee, G. Hoo, M. Thompson, B. Tierney, J. Terdiman, "Real-Time Generation and Cataloguing of Large Data-Objects in Widely Distributed Environments," International Journal of Digital Libraries - Special Issue on "Digital Libraries in Medicine". November, 1997. (Available at http://www-itg.lbl.gov/WALDO/)

[10] William E. Johnston. "Real-Time Widely Distributed Instrumentation Systems," In *The Grid: Blueprint for a New Computing Infrastructure*. Edited by Ian Foster and Carl Kesselman. Morgan Kaufmann, Pubs. August 1998.

[11] William E. Johnston, W. Greiman, G. Hoo, J. Lee, B. Tierney, C. Tull, and D. Olson. "High-Speed Distributed Data Handling for On-Line Instrumentation Systems," Proceedings of ACM/IEEE SC97: High Performance Networking and Computing. Nov., 1997. http://www-itg.lbl.gov/ ~johnston/papers.html

[12] Legion: See http://www.cs.virginia.edu/~legion/

[13] NTON, "National Transparent Optical Network Consortium." See http://www.ntonc.org/.

[14] J. Semke, J. Mahdavi, M. Mathis, "Automatic TCP Buffer Tuning," Computer Communication Review, ACM SIGCOMM, volume 28, number 4, Oct. 1998.

[15] M. Thompson, W. Johnston, J. Guojun, J. Lee, B. Tierney, and J. F. Terdiman, "Distributed health care imaging information systems," PACS Design and Evaluation: Engineering and Clinical Issues, SPIE Medical Imaging 1997. (Available at http://www-itg.lbl.gov/Kaiser.IMG)

[16] Brian Tierney, William E. Johnston, Hanan Herzog, Gary Hoo, Guojun Jin, Jason Lee, Ling Tony Chen, Doron Rotem. "Distributed Parallel Data Storage Systems: A Scalable Approach to High Speed Image Servers," ACM Multimedia '94 (San Francisco, October 1994). http://www-itg.lbl.gov/DPSS/papers/

[17] Brian Tierney, W. Johnston, H. Herzog, G. Hoo, G Jin, and J. Lee, "System Issues in Implementing High Speed Distributed Parallel Storage Systems," Proceedings of the USENIX Symposium on High Speed Networking, Aug. 1994, LBL-35775. http://www-itg.lbl.gov/DPSS/ papers.html.

[18] Brian Tierney, W. Johnston, G. Hoo, J. Lee, "Performance Analysis in High-Speed Wide-Area ATM Networks: Top-to-Bottom End-to-End Monitoring," IEEE Network, May, 1996, Vol. 10, no. 3. LBL Report 38246, 1996. http://www-itg.lbl.gov/DPSS/papers.html