

HIGH-SPEED DISTRIBUTED DATA HANDLING FOR HIGH-ENERGY AND NUCLEAR PHYSICS¹

*William E. Johnston*², *William Greiman, Brian Tierney, Arie Shoshani, Craig Tull*
Information and Computing Sciences Division
Douglas Olson, Nuclear Science Division
Ernest Orlando Lawrence Berkeley National Laboratory, University of California

1 INTRODUCTION

The advent (and promise) of shared, widely available, high-speed networks provides the potential for new approaches to the collection, organization, storage, and analysis of high-speed and high-volume data streams from on-line instruments. Such data streams originate from many types of on-line instruments and imaging systems, and are a “staple” of modern scientific, health care, and intelligence environments. We are defining and implementing an approach that provides for real-time analysis, cataloguing, and archiving of the data streams through the integration of data management techniques, a high-speed distributed application-oriented cache, distributed high performance applications, and transparent management of tertiary storage systems.

In the “Data Access and Analysis of Massive Datasets for High-Energy and Nuclear Physics” (a Grand Challenge project of DOE, Energy Research, Mathematical, Information, and Computational Sciences Division) we are addressing the issues of organizing and querying massive data sets.

In our data-intensive computing projects we are addressing issues associated with capture, processing, cataloguing/indexing, and tertiary storage management.

In our high-speed, widely distributed computing project we are addressing the technology and architectures needed to support widely dispersed resources, users, and data sources all having location transparent access to the data and resources through the use of parallel-distributed computing and high-speed wide area networks.

This type of problem - dealing with high volume, high rate data streams from instruments, the associated data management problems for the resulting massive data sets, and widely distributed user communities - is a key issue for modern, large-scale science.

2 THE HENP GRAND CHALLENGE³

Advances in computational capabilities, information management, and multi-user data access are essential if the next generation of experiments in both high energy and nuclear physics are to be able to fully address the forefront scientific issues for which they are designed. Among these forefront issues are two most fundamental questions facing high energy and nuclear physics today, namely characterization of the transition to the Quark-Gluon Plasma (QGP) phase of matter and the discovery of the mechanism responsible for electro-weak symmetry breaking.

These experiments will record and analyze data from physics events of unprecedented complexity. The resulting data streams of up to tens of megabytes per second and the requirements for “data mining” in huge (tens of terabytes) data sets by multiple, geographically distributed teams of scientists set the scale of this Grand Challenge proposal. Simple extrapolations of existing techniques will not be sufficient; new

1. This work is supported by the U. S. Dept. of Energy, Energy Research Division, Mathematical, Information, and Computational Sciences and the High Energy Physics and Nuclear Science Division, under contract DE-AC03-76SF00098 with the University of California. This document is report LBNL- 40459.

2. W. E. Johnston: mail address: 50B-2239, Lawrence Berkeley National Laboratory, Berkeley, CA 94720. Tel: +1-510-486-5014, fax: +1-510-486-6363, wejohnston@lbl.gov, <http://www-itg.lbl.gov/~johnston>.

3. Taken from <http://www-rnc.lbl.gov/GC>

approaches and solutions are required. The Energy Research Supercomputer Centers have a major role in both addressing this intellectual challenge and subsequently in the implementation of solutions that complement the onsite computational capabilities envisioned for the host laboratories of the experiments.

This Grand Challenge Application (GCA) proposal is the result of an unprecedented collective effort of physicists who are participating in the next generation of major experiments supported by DOE's HENP Program (STAR and PHENIX for RHIC, ATLAS for LHC, BABAR for the SLAC B-factory and CLAS for CEBAF) and by computer scientists who will contribute to the solutions needed to address this challenge. The resulting multi-institutional team seeks a common solution because of the similarity of the problems faced by these experiments. Solutions to this "Grand Challenge" will permit major advances in capability for both the high energy and nuclear physics programs of DOE's Office of Energy Research.

The coming generation of HENP experiments will produce orders of magnitude more data than their predecessors; tens to hundreds of terabytes (TB) of raw data per year for each experiment and equivalent amounts of Monte Carlo simulated data to model detector response, detector acceptance and provide a baseline for looking for "new physics." Even after the first level of analysis, each experiment will be left with many tens of TB of data per year. The data must be available to a hundred or more collaborators per experiment, spread across the United States and the world. Effective analysis of this reduced data requires that it be accessed multiple times. The total cpu power required is several 100 GFLOPS for the larger experiments. Then, the tens of TB of processed data must be sorted and further analyzed by many researchers in order to extract publishable scientific results on a wide variety of topics.

Three challenging aspects of data management and access must be solved: 1) efficient organization of the data to be stored, managed, and accessed to allow timely selection of interesting events (avoid full data set reads); 2) providing the cpu cycles and massive parallelism required for analysis and simulations; and 3) development of the software and remote access environment that permits many (100) physicists to individually select the data sets of interest and implement particular physics analysis algorithms.

(From <http://www-rnc.lbl.gov/GC>)

3 APPROACH

In this paper we describe a distributed, wide area network based architecture intended to deal with many aspects of the data that originates from on-line instruments. This architecture is centered on a network-based, distributed high-speed cache that provides:

- Transient storage for collection, processing, and organization of instrument data streams
- A common high-performance data interface for:
 - instruments
 - parallel-distributed processes (simulation, reconstruction, and analysis applications in the case of HENP)
- Support for large numbers of distributed applications (e.g. many remote analysts)
- Mass storage system independent, tertiary storage management and access
- Possible query refinement mechanism through in-line data filtering

One motivation for providing and managing *distributed* access to tertiary storage is that laboratory instrumentation environments, hospitals, etc., are frequently not the best place to maintain a large-scale digital storage system. Such systems can have considerable economy of scale in operational aspects, and an affordable, easily accessible, high-bandwidth network can provide location independence for such systems.

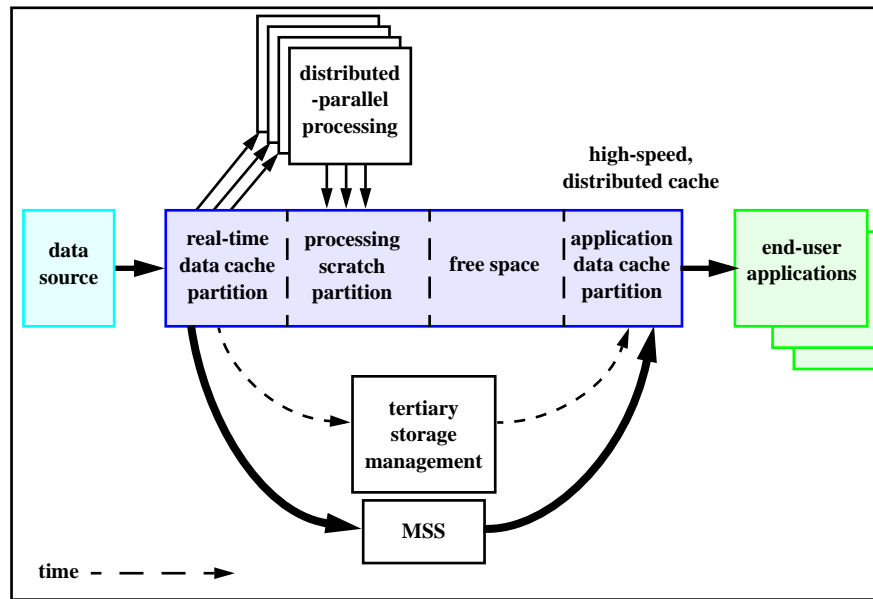


Figure 1 Data-Intensive Computing Model

Our model (Figure 1) is that data from on-line systems is collected in a cache, from which it is processed and archived (the processing could be initial analysis for tertiary storage optimization, for deriving alternate data representations, etc.). The same cache (which may be centralized or as distributed as the user community) is used for staging data (or providing a window on very large data sets) from tertiary storage in order to provide application access to the data.

In one current prototype of this architecture, high-volume health care video data used for diagnostic purposes (a cardio-angiography system) are collected at centralized imaging facilities and, through the use of the architecture described here, are stored, accessed, managed, and used at hospitals of the referring physicians. This instrument generates about 3 megabytes/sec of network traffic. In health care imaging systems the importance of remote end-user access is that the

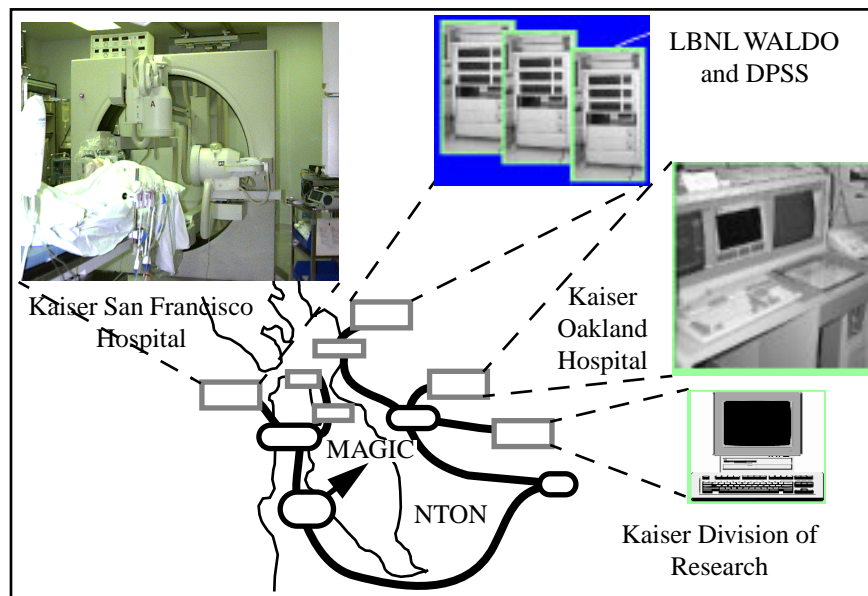


Figure 2 Kaiser / LBNL Distributed Health Care Imaging System

health care professionals at the referring facility (hospitals or clinics frequently remote from the tertiary imaging facility) will have ready access to not only the image analyst's reports, but the original image data as well.

Similarly with data intensive, scientific collaborations, researchers that are at sites remote from the data generation and storage require ready access to the data objects. See, e.g., [Johnston95V] and [Greiman97H].

In this paper we specifically discuss projects whose goals are to demonstrate a new and scalable approach to the problem of high-bandwidth data handling for analysis of high-energy and nuclear physics data, when the source of data (20-40 megabytes/sec) is remote from the computational and storage facilities. The STAR experiment at RHIC ([STAR1], [STAR2]) is used as the basis for a realistic example. The STAR data characteristics are illustrated in Figure 3.

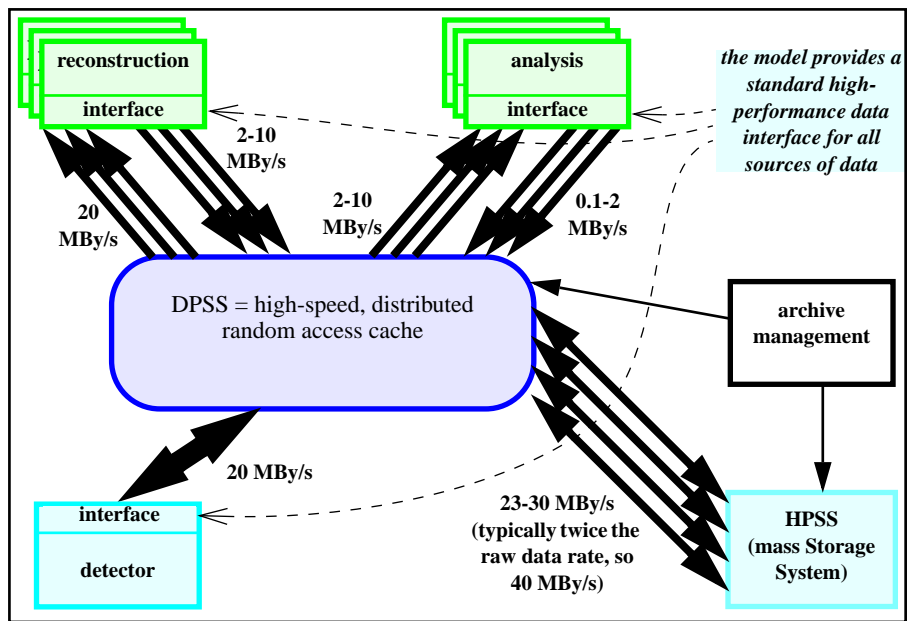


Figure 3 STAR Data Flow Characteristics

One objective of this work include demonstrating the use of distributed computational systems to do the first level of data analysis in real-time. The results of this first level analysis will support two capabilities. First it provides auxiliary information to assist in the organization of data as it is transferred to tertiary storage (the STAR experiment is expected to generate about 1.7 terabytes/day), and second it can

potentially provide feedback to the instrument operators about the functioning of the accelerator - detector system and the progress of the experiment, in order that changes and corrections could be made for this objective. This data model is illustrated in Figure 4.

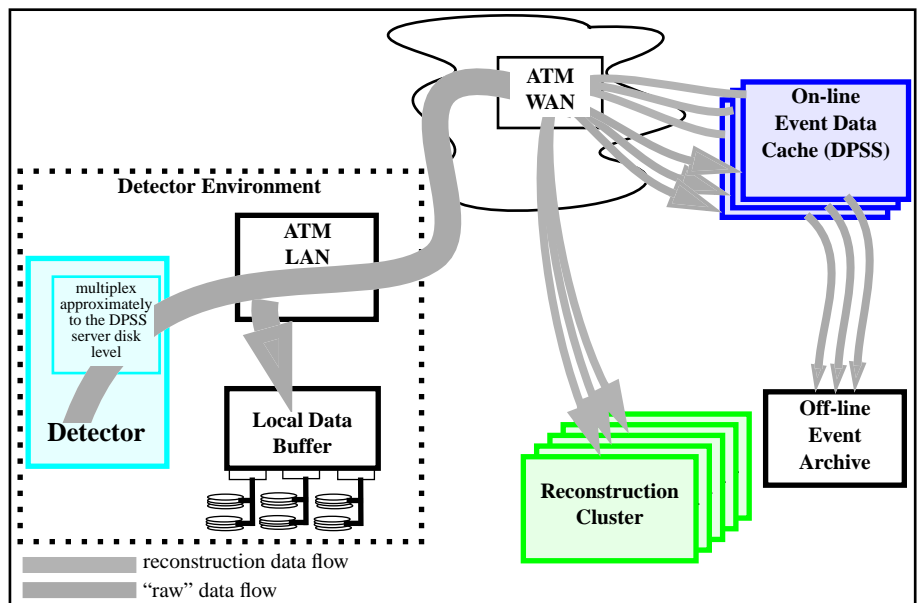


Figure 4 HENP Data Handling Model for Event Reconstruction

A second objective is to support distributed analysis. In this situation many (of order 100) users who are scattered throughout the US, as well as other HENP science locations require access to collections of second-level data in order to perform the physics analysis. In this case, the cache will serve as a “window” on the data sets that reside on tertiary storage. The cache will provide many different analysis processes with a

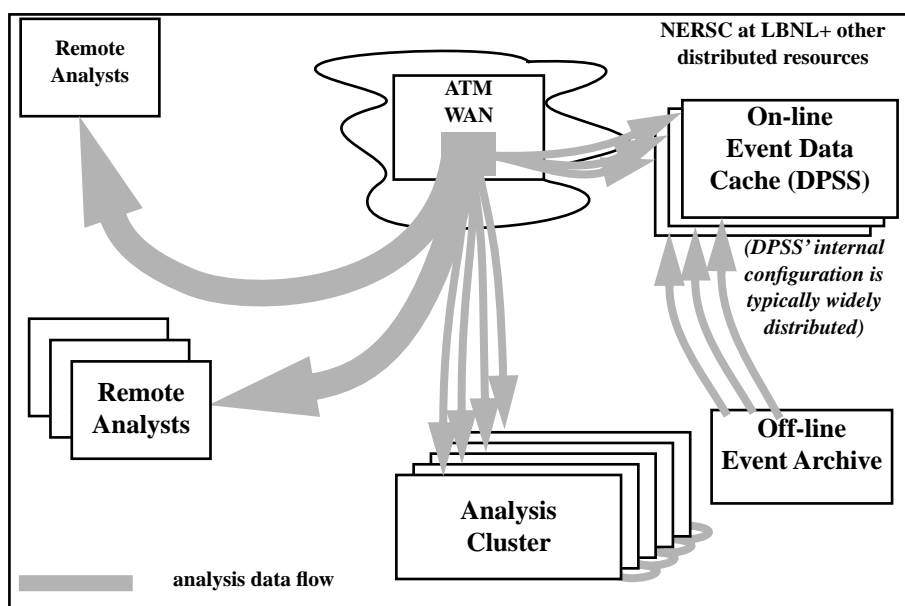


Figure 5 The HENP Analysis Data Handling Model

uniform access to data as a result of queries identifying data sets on tertiary storage, and then migrating those data sets (or the portions of the data for which on-line storage is available) into the cache. The cache can be centrally located, or distributed around the major analysis sites. In any event, all analysis codes have access to all of the data. This data model is illustrated in Figure 5.

The architectural issues include the use of a distributed-parallel storage system as the cache for all stages of data manipulation, the organization of the cache, the configuration of the supporting networks, the various interfaces to the cache, and the management of the movement of data to and from the tertiary storage systems, all in a wide area network.

The primary requirement is to be able to deal with the steady state (for weeks at a time) operation of the distributed system so that data is not lost because of failure or congestion in network congestion or computing systems in the processing / storage pipeline.

This paper also describes an experiment that is designed to validate and demonstrate the approach, and some early results using an OC-12 (622 megabits/sec) ATM network to connect the components that implement the architecture.

4 IMPORTANCE OF THE NEXT GENERATION INTERNET

A range of data-intensive applications will potentially be enabled by networks that routinely provide a thousand times the bandwidth available today.

For example, if we posit that each major instrument / experiment at the National labs had access to between 50 and 100 megabytes/sec of data bandwidth (500 - 1000 megabits/sec of ATM/SONET level bandwidth), then major experiments with highly distributed user communities could operate in a truly distributed fashion and provide new capabilities through real-time data analysis using distributed resources.

For example in the case of a particle accelerator/detector such as RHIC/STAR, the 20-40 megabytes/sec data stream out of the detector could be distributed directly to the experimenter's sites. Each “experiment” (several weeks worth of data collection) could be distributed to the, say 5 - 10 sites directly involved in the experiment. This mode of operation would permit the use of many distributed storage and computing resources, and would allow the collaborators to immediately start on data analysis. The original data could be archived at one, or several, “central” archives (not necessarily at the instrument site, but at a large archival storage operation such as that provided at NERSC), or could be stored in many smaller archives at the collaborators' sites. Upon completion of the first level of analysis, the processed data would immediately be available for second-level analysis. The second-

level analysis might be performed on a different set of resources than the first level analysis, so data would once again be sent into the network to a different set of distributed processing and storage resources, and so on, until the analysis is complete.

This is a continuous process, and so the aggregate traffic is the sum of all of the components, with the network links closest to the source of data (and perhaps the central archive systems) having to sustain the highest bandwidth “single” flows.

This model of operation will:

- enable much more rapid analysis of data, with the ultimate goal being real-time analysis so that the experiment may be validated, modified, or controlled based on the immediate feedback provided by real-time analysis;
- enable alternatives to the expensive, single-use central computing and storage systems sited with the instrument, as is the current mode of operation;
- allow for the aggregation of distributed resources located at the collaborator’s sites for computation, cache storage, and archival storage systems, in order to produce systems that are larger than those found at any one site.

This model is applicable to any high-data rate scientific instrument, including, for example, synchrotron light source micro-spectrometers, scanning confocal microscopes, electron microscopes, etc. In the health care field, applications like cardio-angiography (multi-plane X-ray video) present similar issues.

In both of these cases LBNL is developing experimental systems that are designed to demonstrate the practicality and validity of this use of high speed networks.

5 THE OVERALL MODEL

The high-speed data handling model is based on the idea of a standard interface to a large, application-oriented, on-line cache. Each data source deposits its data in the cache, and each data consumer takes data from the cache, usually writing the processed data back to the cache. In almost every case there is also a tertiary storage system manager that migrates data to and from the cache. (See Figure 1.)

Depending on the size of the cache relative to the objects of interest, the storage system manager may only move partial objects to the cache; that is, the cache is a moving window for the object/dataset. The cache - application interface can (and for this application, does) implement disk read semantics: upon request available data is returned, requests for data in the dataset, but not yet migrated to cache, causes the application-level read to block.

Generally, the cache is large compared to the available disks of the computing environment, and very large compared to any single disk (e.g. hundreds of gigabytes).

This general model has been used in several data-intensive computing applications. For example, a real-time digital library system (see Figure 6 and [DIGLIB]) provides the architecture that supports the distributed imaging prototype mentioned above. This system collects data from a remote medical imaging system, and automatically processes, catalogues, and archives each data unit together with the derived data and metadata, with the result being a Web-based object representing each dataset. This automatic system operates 10 hours/day, 5-6 days/week with data rates of about 30 megabits/sec during the data collection phase (about 20 minutes/hour)

6 PROTOTYPE ARCHITECTURE FOR HENP DISTRIBUTED ANALYSIS

The prototype architecture for HENP data analysis is illustrated in Figure 7. The analysis phase is a second level of processing, and typical data volumes are 1,700,000 megabytes/day, with a processing requirement of 6 KSpecInt92/Mbyte.

The analysis framework generates queries that produce a list of objects of interest. This list of objects, then, has to be retrieved from tertiary storage, and loaded into the cache for processing. The loading process involves parallel transfers from the tertiary storage system to the cache. When an object (or partial object) has been loaded into the cache, the object manager is notified, and in turn it notifies the analysis code. Multiple instances of the analysis code (operating under the control of a work flow

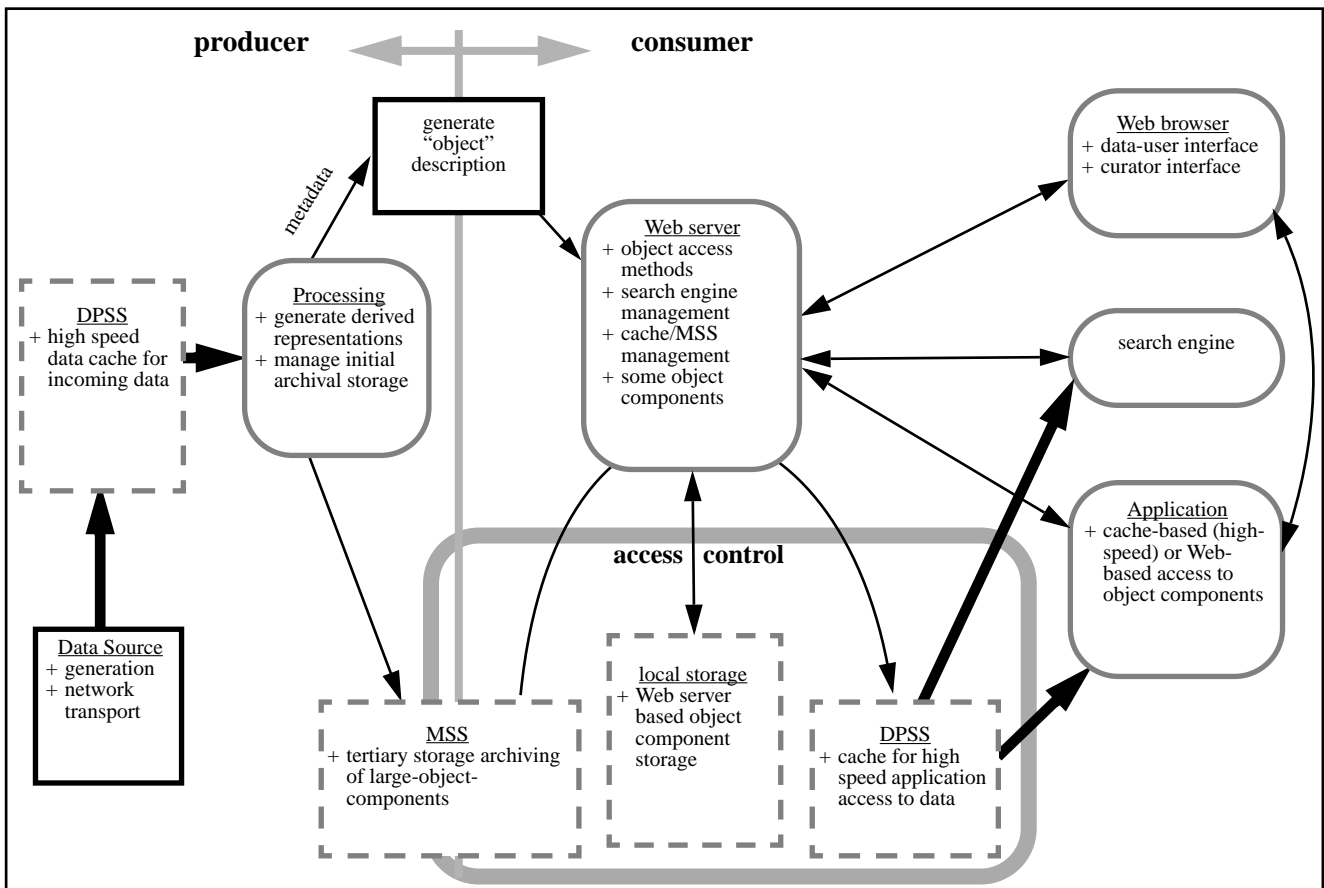


Figure 6 Distributed, Realtime, Large-Object Digital Library Architecture and Data Flow

manager) running simultaneously on many different systems then read data from the cache into memory, and processing commences. In a typical configuration [figure in final paper] the analysis systems may be widely distributed, and they all consume data from the cache, and return results to the cache.

7 STAR ANALYSIS FRAMEWORK

The STAR analysis framework (STAF - see [Tull97]) is being used to provide a realistic application environment in which to validate and refine the data handling architecture and implementation.

Generally speaking, STAF (represented as “Analysis Framework” in Figure 7) manages self-describing data structures on behalf of analysis modules. Data is requested through a standard interface that supports several communications models, including the DPSS cache. The data is

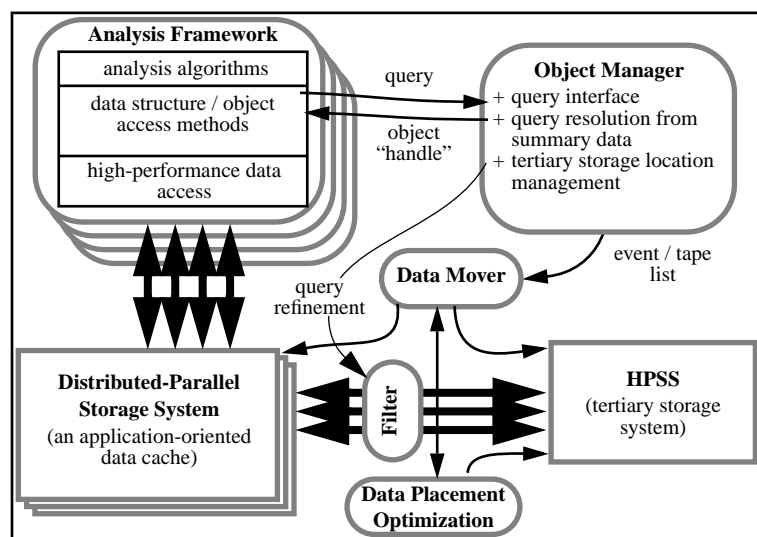


Figure 7 Architecture

converted to machine-specific format and placed into memory data structures, whence it is accessed by the analysis modules.

8 DATA MANAGEMENT

The two major data management problem are:

- ◆ The organization of the initial storage of data so that subsequent queries will have efficient access to the data on tertiary storage;
- ◆ Implementation of a query mechanism that will select subsets of the data based on values of sets of data parameters.

In the first case, data is partially analyzed and index structures built to characterize the parameters of interest. The data is then stored on tertiary storage based on the expected access patterns through the index structures. In the second case the query mechanism operates on summaries of the data to recover the data sets that have desired parameter characteristics. The queries will produce lists of events and their tertiary storage location. The existence of the index structures can be used to optimize the tertiary storage access. (E.g. minimize the number of tape reads.)

9 THE DISTRIBUTED CACHE ARCHITECTURE

The distributed-parallel storage system [DPSS] serves several roles in high-performance, data-intensive computing environments. This application-oriented cache provides a standard interface for high-speed data access, with the functionality of a single, very large, random access, block oriented I/O device (i.e. a “virtual disk”). It provides a high capacity (we anticipate a terabyte size for the STAR analysis environment) and serves to isolate the application from the tertiary storage system and the instrument. Many large datasets can be logically present in the cache by virtue of the block index maps being loaded even if the data is not yet available. In this way processing can begin as soon as the first data has been migrated from tertiary storage.

There are several features of the DPSS that make it an important and unique capability for distributed architectures. These features include application-specific interfaces to an extremely large (16 byte indices) space of logical blocks; the ability to dynamically configure DPSS systems by aggregating workstations and disks from all over the network (this is routinely done in the MAGIC testbed); the ability to build large, high-performance storage systems from the least expensive commodity components, and; the ability to increase performance by increasing the number of parallel operating DPSS servers.

As illustrated in Figure 8, the DPSS is a “logical block” server whose functional components are distributed across a wide-area network. The DPSS uses parallel operation of distributed servers to supply, e.g., image streams fast enough to enable various multi-user, “real-time”, virtual reality-like applications in an Internet / ATM environment. The DPSS is fundamentally a random-access logical block server: There is no inherent organization to the blocks, and in particular, they would never be organized sequentially on a server. The data organization is determined by the application as a function of data type and access patterns so that a large collection of disks and servers can operate in parallel enabling the DPSS to perform as a high-speed data source or data sink.

The high performance of DPSS - about 10 megabytes/sec/commodity disk server - is obtained through parallel operation of independent, network-based components. Flexible resource management - including dynamically adding and deleting storage elements, partitioning the available storage, etc., are provided by design, as are high availability and strongly bound security contexts. The scalable nature of the system is provided by many of the same design features that provide flexible resource management, which has the capability to aggregate dispersed and independently owned storage resources into a single cache.

When datasets are identified by, e.g., the STAF object manager, and are requested from tertiary storage, the logical to physical block maps become immediately available. The data mover that migrates data from MSS to DPSS operates asynchronously, and if an application “read” requests a block that has not yet been loaded, then the application is notified (e.g. when using the file I/O interface the read operation blocks). At this point the application can wait or request information on available

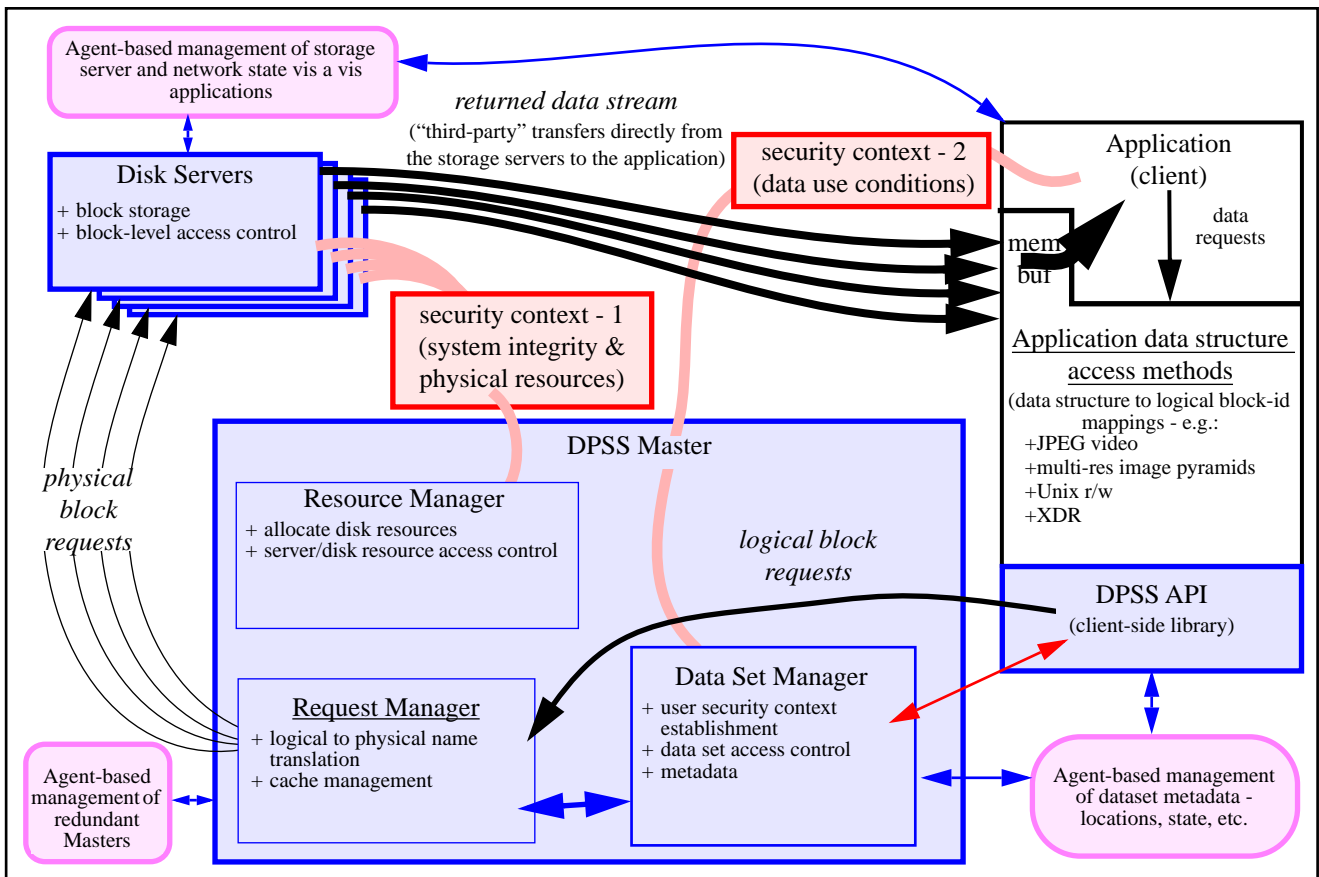


Figure 8 Distributed-Parallel Storage System Architecture

blocks in order to continue processing. (STAF reads megabyte sized data units, but processes these units independently, so processing scheduling can be data driven or organization - request order - driven.)

The STAF application uses file I/O semantics provided in the DPSS access interface, and reads to do not complete until data is available.

Typical DPSS implementations consist of several low cost workstations, each with several SCSI controllers, and several disks on each controller. A three server DPSS can thus provide transparent parallel access to 20-30 disks (150 disks in a high capacity configuration). 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" (stripe across servers and disks), otherwise the block locations are randomized when they are written. (Our experience has shown that random placement of blocks provides nearly optimal parallelism for a wide range of read patterns if the number of independent disks is large.)

The DPSS provides a scalable, dynamically configurable, high-performance, and highly distributed storage system that is usually used as a (relatively long-term) cache of data. It is typically used to collect data from on-line instruments and then supply that data to: analysis applications; to high data-rate visualization applications as in the case in the MAGIC wide-area gigabit testbed where the DPSS was originally developed (see [Lau94], [DPSS], and [MAGIC]); or to mass storage systems (after cataloguing, reorganizing, or pre-processing). The system is currently used in satellite image processing systems and for distributed, on-line, high data-rate health care imaging systems.

10 PERFORMANCE

As was mentioned, a typical DPSS server consists of a commodity workstation (e.g. a 200 MHz Pentium) with one high speed network interface (100 Mb/s Ethernet or 155 Mb/s ATM), three or more SCSI adaptors, and three or more disks on each SCSI string. Each such server can independently

deliver about 10 megabytes/sec of data to a remote application which sees the aggregated streams for all servers in a DPSS system.

Performance for the STAF application was measured using a two disk server, four disk, DPSS configuration. Data requests are made through the DPSS file semantics interface which collects blocks from the DPSS servers, buffers them, and provides serial access to the buffer through an API. The throughput rates are measured as data is delivered to the analysis modules, a path that includes translating the data to the appropriate machine format and structuring it in memory (both of which are very fast operations). With STAF running on a Sun E-4000 system with an OC-12 (622 Mbit/s) ATM interface, a data rate of 19 megabytes/s is achieved for reading data (as expected for two disk servers), and 16 megabytes/s for writing data.

Running 10 instances of the application simultaneously results in about the same aggregate throughput.

11 AN EXPERIMENT IN HIGH-SPEED, WIDE AREA DISTRIBUTED DATA HANDLING

It is our contention that in the time frame of the next generation of physics experiments (2000-2005 AD) that wide area networks will be easily capable of distributing the instrument output data stream anywhere in the US (and probably to Europe).

There are two advantages to this scenario. First, the first level processing (which is easily parallelized) can be done using resources at the collaborators sites (each experiment typically involves 5-10 major institutions). Second, large tertiary storage systems exhibit substantial economies of scale, and so using a large tertiary storage system at, say, a supercomputer center, should result in more economical storage, better access (because of much larger near-line systems - e.g. lots of tape robots) and better media management, especially in the long term, than can be obtained in local systems.

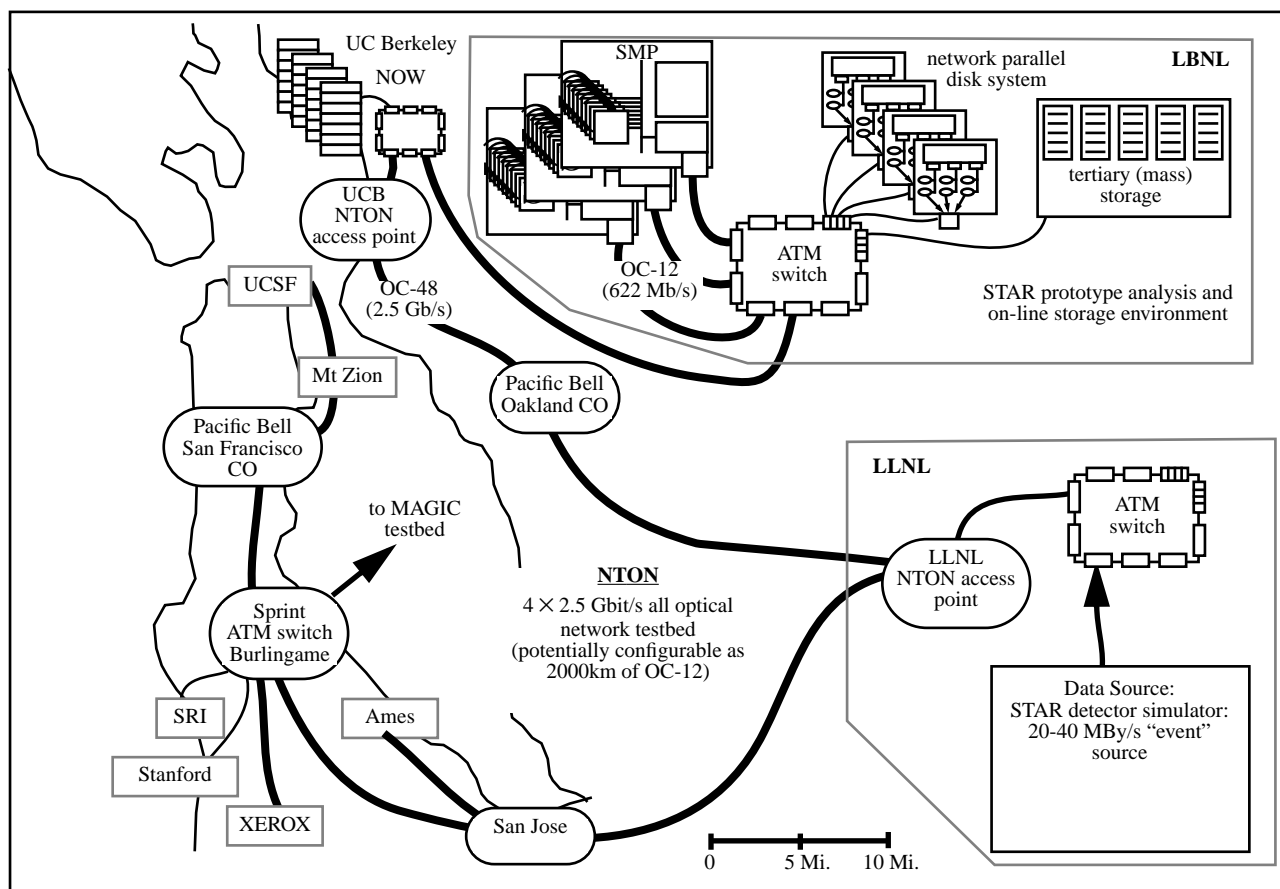


Figure 9

The Distributed Processing Experiment

To this end, we propose that the data handling architecture developed for the real-time digital cataloguing system noted above (Figure 6) can be used for this purpose

This scenario will be tested in the following experiments. A DPSS and a computing cluster are located at Lawrence Berkeley National Laboratory. The NTON network testbed [NTONC] that connects Berkeley and LLNL can be configured for a 2000 km, OC-12, path (by using the 16 OC-12 SONET paths that make up the 400 km underlying network). A high-speed workstation that has a collection of STAR events stored on its disks is located at LLNL and connected to NTON. This workstation will emit events at the same rate as the STAR detector, and this data will be cached on the DPSS at Berkeley. The computing cluster will process data out of the cache (doing “reconstruction”) and those results will be written back to the cache. A storage manager will migrate data to tertiary storage (or a “null” system that has the same throughput characteristics, as there is little point in actually storing this synthetic data).

12 CONCLUSIONS

The experiments described here are work-in-progress. The use of the DPSS as cache has demonstrated the required performance, but a complete demonstration of scalability requires running hundreds of analysis processes (which will be done in the near future). The wide area, high-data rate experiment configuration is nearly complete, and results are expected in the near future. We expect that this experiment will be successful, because several precursors have been carried out in the MAGIC testbed. However, experience has also shown that every significant increase in throughput and/or scale raises a new set of issues. We anticipate that this work will both provide a prototype for, and migrate to the Next Generation Internet environment in order to provide realistic environments for distributed science.

13 REFERENCES

- DIGLIB** “Real-Time Generation and Cataloguing of Large Data-Objects in Widely Distributed Environments”, W. Johnston, Jin G., C. Larsen, J. Lee, G. Hoo, M. Thompson, B. Tierney, J. Terdiman. To be published in International Journal of Digital Libraries - Special Issue on “Digital Libraries in Medicine”. Available at <http://www-itg.lbl.gov/WALDO> .
- DPSS** “The Distributed-Parallel Storage System (DPSS)”. See <http://www-itg.lbl.gov/DPSS>.
- Greiman97H** Greiman, W., W. E. Johnston, C. McParland, D. Olson, B. Tierney, C. Tull, “High-Speed Distributed Data Handling for HENP”. International Conference on Computing in High Energy Physics, Berlin, Germany, April, 1997. Also available at <http://www-itg.lbl.gov/STAR>.
- Johnston95V** W. Johnston, and D. Agarwal, “The Virtual Laboratory: Using Networks to Enable Widely Distributed Collaboratory Science” A NSF Workshop Virtual Laboratory whitepaper. (See <http://www-itg.lbl.gov/~johnston/Virtual.Labs.html>)
- Lau94** “TerraVision: a Terrain Visualization System”. S. Lau, Y. Leclerc, Technical Note 540, SRI International, Menlo Park, CA, Mar. 1994. Also see: <http://www.ai.sri.com/~magic/terravision.html> .
- MAGIC** “The MAGIC Gigabit Network” (<http://www.magic.net/>)
- NTONC** “National Transparent Optical Network Consortium”. See <http://www.ntonc.org> . (NTONC is a program of collaborative research, deployment and demonstration of an all-optical open testbed communications network.)
- STAR1** “Relativistic Nuclear Collisions Program”, H.G. Ritter. <http://www-library.lbl.gov/docs/LBNL/397/64/Overviews/RNC.html>
- STAR2** “High Speed Distributed Data Handling for HENP”, W. Greiman, W. E. Johnston, C. McParland, D. Olson, B. Tierney, C. Tull. http://www-rnc.lbl.gov/computing/ldrd_fy97/henpdata.htm

Tull97 “The STAR Analysis Framework Component Software in a Real-World Physics Experiment”.
C.Tull, W.Greiman, D.Olson, D.Prindle, H.Ward, International Conference on Computing in
High Energy Physics, Berlin, Germany, April, 1997.

Tierney “Performance Analysis in High-Speed Wide Area ATM Networks: Top-to-bottom end-to-end
Monitoring”, B. Tierney, W. Johnston, J. Lee, G. Hoo. IEEE Networking, May 1996.