High-Speed Distributed Data Handling for HENP[1]

W. Greiman, W. E. Johnston[2], C. McParland, D. Olson, B. Tierney, C. Tull
Ernest Orlando Lawrence Berkeley National Laboratory, University of California
Berkeley California, USA, 94720

Keywords: 07.05.Ys, distributed data, high-speed networks

Abstract

We describe a project whose goal is to demonstrate a scalable solution to the problem of high-bandwidth data handling for analysis of high-energy and nuclear physics data. The STAR experiment at RHIC is used as the basis for a realistic example in this project. The approach is based on a distributed-parallel storage system which collects data from the detector and serves data to a distributed cluster of symmetric multi-processor computers.

1.0 Introduction

Experience gained over the past several years in DOE and DARPA funded work in high-speed, wide-area distributed computing (see, e.g. [3]) has lead us to explore the potential of this approach for high energy and nuclear physics (HENP) data handling and analysis.

The general approach is to decompose the overall system into its component elements, and then to organize those elements into a widely distributed system. The advantage of this approach is several fold: first, it relieves the physical constraints of placing computing and storage equipment with the data sources; second, the decomposition allows maximum use of parallelism in the architecture, and; third, the use of widely distributed networks allows the independence of operation of the elements that is necessary to achieve non-interfering parallel operation of the system elements.

The first point - relieving the physical constraints - has been used to advantage in an experimental health care imaging system, where high-bandwidth data is collected in one location, processed and stored at a second, and used at other geographic locations (e.g. [7]). The second point argues that parallel-distributed systems provide a powerful and flexible computing and data handling environment. The third point - independence of components - has been used to advantage in several systems with coarse-grained parallelism, including the Distributed-Parallel Storage System ([3]).

We have decomposed the data handling and analysis for STAR [8]. Each of the major components - the detector, the event reconstruction, the event archive, the on-line storage for the second-level data (reconstructed events), and the physical analysis codes and platforms - have been distributed around a high-speed network.

The distributed experiment architecture is based on an OC-12 (622 Mbit/sec) ATM metropolitan-area network, which may be configured to about a 1000 km diameter, separating the simulated detector from the storage systems (the components of which are also scattered throughout this network). The processing elements include Sun E-4000 and Ultra-2 SMPs, DEC Alpha SMPs, and SGI SMPs, all of which are also scattered around the network.

The goal of the first phase work is to demonstrate that the networks that we expect to be in place at the turn of the century, together with the distributed systems approach described here, can collect and process all of the event data in real-time. This has the potential for new directions such as making decisions on how to best organize the data on tertiary storage and what sub-sets to keep in on-line storage, prior to committing to an off-line storage strategy; using such processing to feed information back to the experiment operators so that near real-time adjustments might be made to the experiment parameters based on the preliminary analysis, etc. The goal of a second phase will be to build a large enough system so that we clearly demonstrate that the approach is capable to the aggregate 1 GBy/s and 100 GFLOPS that will be needed for the production data analysis systems.

In the remainder of the paper, we will briefly discuss the software architecture, system architecture, middleware and middle systems, and prototype nuclear science data system and demonstration.

---

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-39992.

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.

2.0 Software Architecture

The "lower" and middleware architectural elements (Figure 1) for data structure management, machine independent data coding, and storage and communication that have evolved in the STAR Analysis Framework (STAF) are very similar to the general architecture that has evolved in other high-performance distributed-parallel systems (e.g. the terrain navigation application in the DARPA MAGIC gigabit testbed ([5]).

3.0 System Architecture

The idea of network-based instrumentation is no longer unusual (see, e.g., [2]) and is driven by the natural geographic distribution of the principal components: data sources, storage and analysis systems, and human analysts. Our premise is that production network capacity will expand in the next five years to routinely accommodate 20-40 MBy/sec data rates. (The National Transparent Optical Network Testbed [6] - see below - is a model for this.)

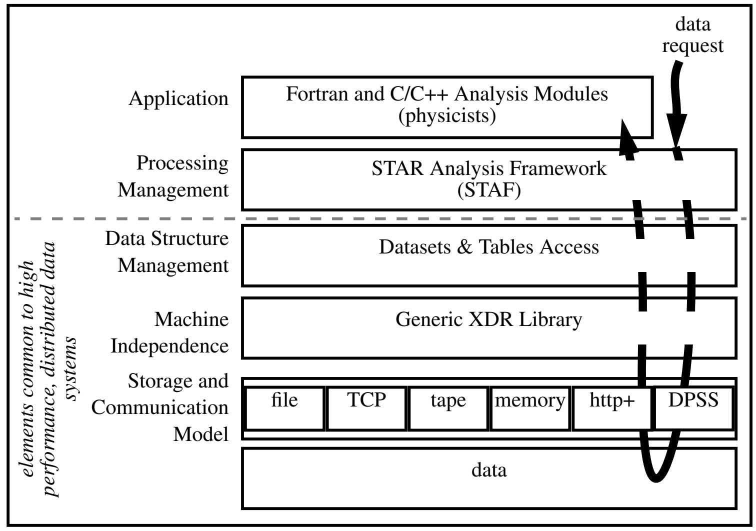

Figure 1 **Software Architecture for Analysis**

3.1 Event Reconstruction

The model for event reconstruction is as follows. Data is sent from the detector to a local workstation where it is formatted, buffered (to protect against network interruptions) and tagged and multiplexed onto an OC-12 ATM network. At the destination - which could be many sites - ATM switches demultiplex the data streams and send them to reconstruction systems (the gray lines in Figure 2). The intent is that there will be enough of these systems operating in parallel to do event reconstruction in real-time. At the same time that event data is archived, either by multicasting the data streams in the ATM switch, or having the reconstruction systems forward the data. The archive is likely to be in one place, since such systems exhibit good economies of operational scale. The processed data resulting from reconstruction (roughly an order of magnitude reduced in size) is sent to an on-line cache where is will made available to the analysts. The on-line cache is a Distributed-Parallel Storage System, described below.

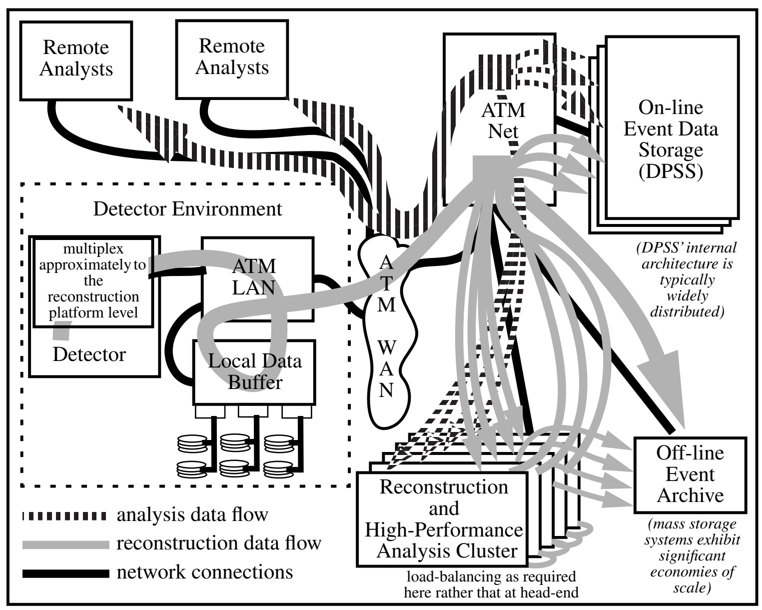

Figure 2 **System Architecture: Data flows**

3.2 Analysis

The model for distributed analysis is that the reconstructed event ("DST") data sets are available from an on-line cache that is large enough to hold all of the data of "current" interest. While the event reconstruction involves a single process reading, processing, and writing a single unit of data, the analysis involves many users simultaneously accessing the same data sets.

4.0 Middleware

The principal middleware (a general use, service-oriented element) involved in this architecture is the Distributed-Parallel Storage System (Figure 3). This system 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 or to supply data to high data-rate visualization applications (see [4]). The system is used in satellite image processing systems and for distributed, on-line, high data-rate health care imaging systems. Performance is obtained through parallel operation of independent, network-based elements, 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.) Many of the same design features that provide flexible resource management, as well as the aggregation of dispersed and independently owned resources provided by the network-based architecture, contribute to the scalable nature of the system.

The current system is based on disk storage of randomly accessed data blocks: The data is typically written to the servers in a round-robin fashion, but read randomly. As part of the STAR work we are looking at generalizing the DPSS architecture to include a tape backing store. Many parallel tape drives would provide a backing store for the disks on each server. Unlike the disks, data will be loaded and unloaded to tape in relatively large units. The disks and server memory both act as an LRU-based cache. The addition of the tape backing store is seen as significantly increasing the size of the DPSS as a *cache* - we do not envision all of the auxiliary functions required to provide archival storage.

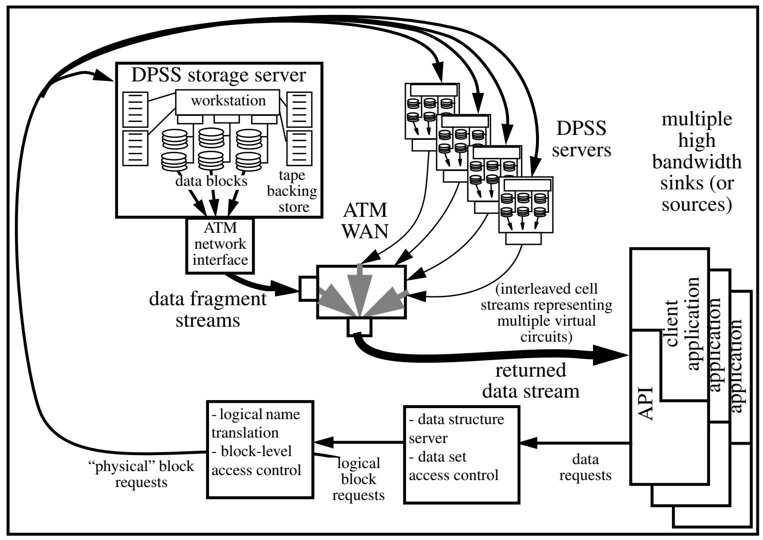

Figure 3 Distributed-Parallel Storage System Architecture

5.0 The Experiment Environment

The STAF and DPSS architectures provide "transparent" integration and demonstration of machine independent access to distributed, on-line data using STAR analysis code. The STAF architecture allows new data handling components inserted "under" the current analysis code interface. The distributed analysis of synthetic data streams involves simulating the detector with (pre-computed) data streams that are sent into a realistic network analysis environment at realistic rates. This is being done by using a high-speed workstation with an OC-12 ATM network interface operating as the detector element and the configurability of the NTON network to provide a 1000 km of OC-12 network. Full data rates will be sent into the prototype network-based storage and processing environment. The data is being analyzed on an experimental cluster of SMPs (see [1]). While full data rates are handled in the network, only every "n-th" event is being reconstructed (n depending on computational capacity in the prototype system). While we are confident of the scalability of the approach because of the independence of the network-based components, this will be demonstrated in the next phases of the experiment.

6.0 References

[1] W. Johnston "Clusters of Multi-Processor Systems in the Scientific Environment". See http://www-itg.lbl.gov/~johnston/COMPS/

[2] W. Johnston and S. Sachs, "Distributed, Collaboratory Experiment Environments (DCEE) Program1: Overview and Final Report, February, 1997. Available at http://www-itg.lbl.gov/DCEE

[3] B. Tierney, et al, "The Distributed-Parallel Storage System (DPSS)". See http://www-itg.lbl.gov/DPSS.

[4] S. Lau and Y. Leclerc, "TerraVision: a Terrain Visualization System". Technical Note 540, SRI International, Menlo Park, CA, Mar. 1994. Also see: http://www.ai.sri.com/~magic/terravision.html .

[5] "The MAGIC Gigabit Network" (http://www.magic.net/)

[6] "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.)

[7] M. Thompson, W. Johnston, G. Jin, J. Lee, B. Tierney, J. F. Terdiman, "Distributed Health Care Imaging Information Systems". SPIE International Symposium on Medical Imaging, 1997. Newport Beach, California. (Also available at http://www-itg.lbl.gov/Kaiser.IMG/homepage.html)

[8] C. Tull, W. Greiman, D. Olson, "MOAST - A CORBA-based Physics Analysis Framework". Proc. of the Conference on Computing in High Energy Physics, Rio de Janeiro, Brazil, 18-22 Sept., 1995.

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