



Fermi National Accelerator Laboratory

FERMILAB-Conf-94/112

Analyzing Terabytes of Data at Fermilab

Stephen Wolbers

*Fermi National Accelerator Laboratory
P.O. Box 500, Batavia, Illinois 60510*

May 1994

Presented at the *Computing in High Energy Physics 94 Conference*, San Francisco, California, April 22-27, 1994

Disclaimer

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

ANALYZING TERABYTES OF DATA AT FERMILAB *

Stephen Wolbers
Fermi National Accelerator Laboratory
P.O. Box 500
Batavia IL 60510 USA

Abstract

Computing demands of High Energy Physics are increasing steadily due to the demands of larger datasets and increasingly sophisticated detector systems and analysis techniques. Fermilab has been meeting these demands by the use of many different computing techniques. Most of these techniques attempt to utilize the most cost-effective computing resources while providing effective solutions to the problems that are created by multi-Terabyte data samples and large collaborations. New strategies are being developed to allow improved access to the data.

INTRODUCTION

During the past 5-10 years at Fermilab the typical experiment has written increasing amounts of raw data to tape in each data run. In addition, the events have become more complicated due to increased energy and intensity of the incoming beams and the improvements that have been made to the detector systems. The increasing availability of computing power has also allowed experiments to become more sophisticated in their analysis programs. The final data sets used for physics analyses have also increased dramatically due to the larger data samples and due to improved triggers and reconstruction algorithms that allow larger and better final event samples to be kept for analysis.

All of these trends have forced Fermilab to focus on providing improved and cost-effective computing to handle the massive amounts of data that are being generated. Different computing solutions have been used for event reconstruction, splitting and filtering, and physics analysis. Though some notable successes have been achieved there are still improvements to be made to continue to keep up with the demands of the experiments.

INCREASING COMPUTING NEEDS

Experiments at Fermilab have been writing and analyzing large and increasing amounts of data. This is certainly not unique to Fermilab and reflects many trends in scientific computing. During the 1990-91 accelerator run the "typical"

*This work is supported by the U.S. Department of Energy under Contract No. DE-AC02-76CH03000.

experiment wrote approximately 2 TB of data to tape, with one experiment writing over 40 TB. During the 1992-93 run CDF wrote 2 TB of data and D0 wrote 8 TB of data. It is expected that CDF will write 6-8 TB and D0 18-24 TB of data during the 1994 run. Larger datasets are anticipated from future data runs. There is no reason to expect the trend of writing increasing amounts of data to tape will not continue.

CPU Needs

The CPU required to handle the event reconstruction (and other processing and analysis steps) are constantly increasing. The largest single requirement is event reconstruction. There are many reasons for the increase including; the increasing number of events, the increased complexity of each event, and the increased sophistication of the reconstruction algorithms. More effort is being used to maximize the physics potential of the data sets and less to saving CPU time due to the increasing availability of more CPU power. Capabilities exist to reconstruct events more than once in many cases. The availability of more CPU power has certainly had an influence on the amount of CPU time used for event reconstruction. The trend toward increased CPU usage for all parts of offline processing will continue.

Final Dataset Sizes

Each experiment has to reconstruct their dataset and split it into samples which can be quickly analyzed to produce physics results. The final samples are growing extremely rapidly. Experimental collaborations are becoming larger, making access to the ever-growing datasets a

more and more crucial problem for fast and effective data analysis. The bottleneck of slow serial access to large datasets has the effect of limiting physics analysis, especially as the datasets continue to grow.

UNIX FARMS

The UNIX Farms at Fermilab are used for the reconstruction of raw data [1]. The CPU-to-I/O ratio is sufficiently high that loosely-coupled computing solves this CPU-intensive task. The farms can be characterized by describing the hardware, software, and the running experience of the last few years. More details can be found in another contribution to this conference. [2]

Hardware

The UNIX Farms consist of extremely cost-effective UNIX workstations connected via ethernet and divided into worker and I/O nodes. The majority of nodes are worker nodes and consist of rack-mounted UNIX workstations (minus the keyboards and screens) containing the minimum memory needed to avoid swapping, and a small amount of local disk for the operating system. The I/O nodes are UNIX servers or workstations and are connected to local SCSI disk and 8mm tapedrives. The UNIX farms currently consist of 10,000 MIPS of computing with over 300 workstations and about 100 tapedrives.

Software

The parallel processing code cps (cooperative processes software) allows the events to be sent to processes running on worker nodes and is an effective way to

provide parallel computing. One of the most important tools used on the farms is the tape-mounting software ocs (operator communications software) which is used to handle the large tape-mounting activity on the farms (and elsewhere at Fermilab). In addition, a batch system that allows queuing of jobs has been developed and is in use. Finally, many utilities are in place to allow debugging, optimization, and viewing of jobs on the farms.

Experiences

The UNIX farms are extremely successful in providing large amounts of cost-effective computing to the experimental users. Both CDF and D0 are able to reconstruct data as quickly as they collect it. The farms have sufficient processing power for present needs and no upgrades are necessary in the near future.

SPLITTING/FILTERING

After reconstruction the data is divided into many physics subsets. The subsets each consist of a sample of events relevant for a set of physics analysis topics and/or is a sample of events useful for background measurements and studies. Though not required, it is oftentimes the case that each event is compressed into a much smaller format containing only quantities essential for physics analysis.

Techniques

Each experiment chooses to handle this step of analysis in the way that matches best the physics of the experiment and to match the computing systems that are available for the process. Ideally the task would be specified inde-

pendently of the hardware available but physical limitations (tape, disk, CPU) all dictate that many different systems are in use.

Examples - CDF and D0

CDF and D0 use somewhat different systems for performing the splitting and compressing of physics datasets after reconstruction. CDF splits datasets on the I/O portion of their UNIX farms. The SGI system consists of a Challenge XL with 4 processors, 80 GBytes of disk and 12 8mm tapedrives. The IBM system consists of an RS6000/590 and RS6000/580 with 140 GBytes of disk and 14 tapedrives. The events as they are reconstructed on the worker nodes of the farms are split into 25 physics streams (with some events being written to more than one stream). The output events can be stored in one of two formats - DST (full information) or PAD (physics analysis dataset). It is expected that the DST dataset will occupy approximately 6 TBytes and the PAD datasets about 875 GBytes from the 1994-95 run.

D0 reconstructs the data on their farms and produces two sets of output events. The first is the full-size (STA) dataset and the other is a compressed format dataset (DST). These datasets are further split on other computing systems. The STA sets are split into physics streams which fill approximately 17 TBytes. The DST sets are split into many more physics streams that sum to about 4 TB. Due to this large size a new reduced format (MDS) was invented to reduce further the size of data so that the whole sample could be compressed to about 250 GBytes. The STA sets are split on two dedicated SGI Crimsons fit-

ted with disk and tapedrives. The remaining splitting and filtering is done on DQFS, a large VMS cluster that is also used to serve data for analysis.

ANALYSIS

The physics analysis of the final datasets is accomplished on a wide variety of computing systems and in a number of different ways. Better access to datasets can and will improve the physics analysis of the data. There are many ways to characterize the many styles of physics analysis that are used. One way is to examine the three main sets of experiments (fixed-target, CDF and D0) to see what is being done.

Fixed-Target

The fixed-target experiments at Fermilab have access to a wide variety of computing systems both at Fermilab and at their collaborating institutions. In general it is possible to handle small datasets and final analysis steps (PAW) on local workstation clusters, most of which tend to be UNIX-based. The point at which the data becomes too large to handle locally varies but normally a sample which is smaller than 10 8mm tapes (or about 20 GB for single-density tapes) can be handled on a local system.

Fermilab has established two central UNIX systems which are meant to allow the fixed-target user community to access larger datasets for physics analysis. The two systems are CLUBS and FNALU. CLUBS consists of a set of SGI and IBM workstations connected with Ultrinet for fast data access, the Load Leveler batch scheduling system, UNITREE hierarchical storage management, and ac-

cess to external 8mm, 3480, and 9-track tapedrives and an STK silo. This system, with a CPU capacity of over 500 MIPS, allows users to analyze data that is staged from tape to disk via a staging system. Data is stored in the STK silo for rapid, reliable and multiple access. A dataset, once read in to the STK silo from 8mm tape, can be repeatedly read much more rapidly and reliably than from 8mm itself. A hierarchical system has proven effective in the past on the Amdahl mainframe in reducing manual tapemounts and providing reliable access for many physicists to a common dataset.

FNALU is a central UNIX system consisting of IBM and SGI computers divided into interactive and batch components. The AFS file system is being used on this system to provide home directory, product and some data access. The interactive systems are also used as front-ends to the CLUBS system. Batch jobs are prepared and submitted from FNALU to CLUBS. Local batch capacity on FNALU is available for users and applications that are not well-matched to the CLUBS system.

CDF

CDF uses two large central facilities for access to their datasets. The first is a large VMS Cluster (FNALD) consisting of about 500 MIPS of processing and 400 GB of disk space and a connection to an STK silo. In addition there are a large number of 8mm tapedrives directly connected to the Cluster to allow access to datasets. The STK silo contains CDF PAD datasets (and a small amount of DST datasets) which can be staged to disk and analyzed on the VMS cluster. The second system used is a UNIX sys-

tem consisting of an SGI 4D/480 (cdf-ga), about 100 GB of staging disk, 16 8mm tape drives, and a connection to the same STK silo. The connection to the silo allows CDF UNIX analysis access to the datasets via a staging mechanism.

In addition to the central systems CDF has large local VMS and UNIX clusters which are heavily used for data analysis. The analyzer and the analysis project determine where each analysis will actually be done. Access to the data on the local machines is available from local disk or tape or from data copied over the network.

D0

D0 has established a large file server (D0FS) to provide access to large datasets. D0FS is a VMS Cluster consisting of about 500 GB of disk and 30 tapedrives along with 2 exabyte tape robots. The disk is currently spread across 34 workstations. The cluster is connected via FDDI to the D0 analysis clusters consisting of VMS (predominantly) and UNIX workstations. Users have the ability to access data using D0FS to serve data over the network or by reading local copies of the data.

Analysis strategy

The techniques that have been developed for analysis are not sufficient to meet the current and growing needs. The limitations of the current systems tend to limit or make difficult many analysis projects. The larger data samples that exist and that are going to be created in the future make it more important that strategies for data and computing access be investigated. The use of robots and hi-

erarchical storage is essential in order that the huge manual tape-mounting load be reduced. One of the most difficult problems is the lack of scalability of most solutions now in use. Each increase in system size tends to create bottlenecks which are difficult if not impossible to overcome.

FUTURE DIRECTIONS

Analysis needs and computing in general benefit from having data as close as possible to computing power. Experience has shown that coupling computing and data as tightly as possible leads to large improvements in the ability to handle the data effectively. An attempt to implement such a strategy in order to analyze ever-growing amounts of physics data is the CAP (Computing for Analysis Project) system at Fermilab. In addition to putting data close to computing the project is scalable, leading to the ability to naturally handle increased data needs.

CAP

CAP is a project with a goal of providing HEP experiments with quick and reliable access to large amounts of data. The design goals are to store up to 100 TB of data in a tape robot (or robots), manage the files via a storage management scheme, read the data quickly onto a large (> 300 GB) disk pool and provide sufficient parallel I/O and computing to read through large datasets quickly. Various schemes of data management and storage are being prototyped in order to understand which techniques will be most effective in handling the needs of experiments in processing large amounts of data quickly.

The CAP hardware currently consists

of an IBM SP/1 multiprocessor system as the parallel compute and I/O system, the VESTA parallel file system from IBM Yorktown, UNITREE to manage the data, and a cartridge tape robot.

One of the major challenges of the system is to provide data organization and access that provides the speed and functionality necessary both for system performance and for analysis needs. Possible solutions include object-oriented data structures and other techniques for storing pieces of events to allow more efficient and logical access.

A successful implementation of this strategy promises a much improved system for handling the ever-increasing size of datasets. The CAP project should allow datasets of order 100's of GB to be handled with short turnaround times. This will create opportunities for data analysis which either do not exist today or are rendered rather difficult due to the long time and difficult data handling problems involved. The ability to handle the large datasets quickly will lead to better physics.

CONCLUSIONS

High-Energy Physics data sizes are growing ever-larger and this size increase is focussing effort on data-handling and data-processing of these growing sets. Fermilab has been blessed (or burdened) with a very large data-handling problem. A mix of ad-hoc solutions to the various aspects of the problem has been invented and is reasonably effective given current demands. This mix includes UNIX Farms for event reconstruction, a combination of UNIX and VMS Clusters for event stripping and filtering and a different and more diverse combination of UNIX and

VMS Clusters, robotics, staging software and applications for physics analysis.

One of the ideas for improvement is to integrate fast data access with fast CPU resources. The CAP project at Fermilab is a system that can provide increased capabilities in an integrated system for handling the increasing demands of experiments. This approach can help us in making data analysis simpler and more effective for the ever-growing data samples of HEP.

ACKNOWLEDGEMENTS

Many groups of the Computing Division at Fermilab have contributed to this work, including the Farms, Unix Systems Support and VMS System Support Groups, the CLUBS group, the CDF and D0 groups and many others who contribute to the successful integration and maintenance of all the many computing systems. Data Center Services has done a magnificent job handling the huge amount of data that comes into the Feynman Center.

REFERENCES

1. F. Rinaldo and S. Wolbers, "Loosely Coupled Parallel Processing at Fermilab," *Computers in Physics*, 7, 184, Mar/Apr, 1993.
2. M. Fischler, F. Rinaldo, S. Wolbers, "Production Farms at Fermilab", Contribution to CHEP94.