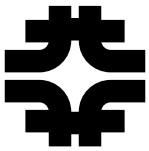

Fermi National Accelerator Laboratory

FERMILAB-Conf-95/328

Integrating Data Acquisition and Offline Processing Systems for Small Experiments at Fermilab

J. Streets, B. Corbin and C. Taylor

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

October 1995

Proceedings of the *Computing at High Energy Physics 1995 (CHEP 95)*,
Rio de Janeiro, Brazil, September 18-22, 1995

Disclaimer

INTEGRATING DATA ACQUISITION AND OFFLINE PROCESSING SYSTEMS FOR SMALL EXPERIMENTS AT FERMILAB

J.STREETS

Fermi National Accelerator Laboratory,
Batavia, Illinois 60510

B.CORBIN

University of California,
Los Angeles, California 90024-1547

C.TAYLOR

Case Western Reserve University,
Cleveland, Ohio

Two small experiments at Fermilab are using the large UNIX central computing facility at Fermilab (FNALU) to analyse data. The data aquisition systems are based on "off the shelf" software packages utilizing VAX/VMS computers and CAMAC readout. As the disk available on FNALU approaches the sizes of the raw data sets taken by the experiments (50 Gbytes) we have used the Andrew File System (AFS) to serve the data to experimenters for analysis.

1 Introduction

APEX and MINIMAX are two experiments taking data during the Fermilab Collider Run. Because of the small size of the experiments, their low data collection rates (50-100 Hz event rate and 400-2000 byte event size) and the very restricted manpower available, they have taken data with the legacy VAXONLINE system[1]. During the last twelve months of data taking, the experimenters have been proactive in extending the "counting room" to the central Fermilab UNIX analysis computers through the use of Andrew File System (AFS)[2]. Services on the central Fermilab UNIX cluster - FNALU[3]- are in the process of being developed to support future experiments offline processing and analysis requirements. The use of these facilities by APEX and MINIMAX at this time is acting as a prototype and beta testing project to help in determining the right configuration and mix of services to be provided.

We report on the infrastructure and services used - integration of the centrally provided AFS disks in the online environment, batch job scheduling and feedback to the counting room and data transfer and analysis rates obtained. We include observations of the positive and negative aspects of migrating legacy experiments and experimenters to this new operating paradigm.

2 Running the experiments

CDF and D0 are the major users at Fermilab during collider operation although several smaller experiments can be using beam from the accelerators at the same time. These experiments typically have up to 20 collaborators and usually run in a parasitic mode.

Two of these experiments are APEX and MINIMAX. They take data from CAMAC into a VAX/VMS QBUS computer at rates of up to 100 kbytes/second. The raw data is written to disk or 8mm tape, and analysed with several types of computers at Fermilab and at collaborating institutions. Accelerator information is read from remote computers via ethernet every few minutes and written to the main data stream. Acceptance calculations are also required, and both groups make use of GEANT[4]. The format for the Monte Carlo data is kept similar to the format of real data so that the same program can analyse either type.

By keeping similar online and offline software, we have been able to reduce the management of code and data for otherwise unrelated experiments.

2.1 APEX

APEX is searching for the decay of antiprotons at the Fermilab antiproton accumulator (AA). An early test[5] ran in 1992, and a run between April and July of this year with upgraded apparatus is hoped to improve the limit on the electron decay modes by two orders of magnitude. A first search for muon decay modes will also be made with this data. The experiment uses a large decay tank welded into the AA, hodoscopes for triggering, calorimetry for energy measurement, and scintillating fibre planes for track reconstruction. The event size is a few 100 kbytes.

Data was taken when no antiprotons were being stacked, this occurs only when a shot is being set up for the collider (every 24 hours), or when there is a problem in the accelerator complex. To take advantage of unscheduled down times, the front end computer monitored the accelerator and paged several members of the experiment when conditions were good for data taking.

The experiment ran for 3 months, and took 50 million triggers. Most of these were written to 8mm grade tape, although some were taken to disk for fast turn around analysis.

2.2 MINIMAX

MINIMAX is situated at the C0 intersection of the Tevatron Collider, and is designed to detect event fragments in the high rapidity region. This region is a possible place to search for Disoriented Chiral Condensates (DCCs) which could explain events with anomolous charge to neutral particle ratios seen in cosmic ray experiments[6].

MINIMAX first took data in 1993, and the full detector was assembled early this year. It comprises of 24 wire chambers, lead/scintillator calorimeters and trigger hodoscopes. Event sizes are around 3 kbytes. The experiment takes data during special low luminosity runs when the electrostatic beam separators at C0 can be switched off. MINIMAX measures the ratio of charge to neutral tracks by counting

```
fnapx1 6% ls /afs
afs1.scri.fsu.edu       ctp.se.ibm.com          nersc.gov
alw.nih.gov             desy.de                 palo_alto.hpl.hp.com
andrew.cmu.edu          dsg.stanford.edu        psc.edu
anl.gov                 ece.cmu.edu             pub.nsa.hp.com
athena.mit.edu          es.net                  rel-eng.athena.mit.edu
bu.edu                  fnal                    rhic
caspur.it               fnal.gov                ri.osf.org
cern.ch                 grand.central.org       rpi.edu
citi.umich.edu          hep.net                 rrz.uni-koeln.de
club.cc.cmu.edu         hepafs1.hep.net         slac.stanford.edu
cmu.edu                 iastate.edu             theory.cornell.edu
cs.brown.edu            ibm.uk                  transarc.com
cs.cmu.edu              ir.stanford.edu         umich.edu
cs.wisc.edu             msc.cornell.edu
ctd.ornl.gov            ncsa.uiuc.edu
```

Figure 1: The Andrew File System as seen from a client

the numbers of tracks before and after a neutral to charge converter, and looking at correlations in the calorimetry. Systematic studies with different thicknesses and types of converter are required to search for DCCs.

To date, 3.5 million events have been taken so far to tape and disk with various conditions. All the raw data are kept on AFS, as well as results from tracking algorithms and Monte Carlo generation.

3 What is AFS ?

The Andrew File System (AFS) was originally developed at the Information Technology Centre at Carnegie-Mellon University. It is now marketed and maintained by Transarc Corporation. AFS appears as a UNIX file system from AFS client nodes. The file system is separated into cells, grouped by internet domain. Each cell is maintained locally, and comprises of one or more AFS servers which define the directory tree for that cell. Figure 1 shows the file system as seen by one of the experiments file systems.

File protection is not handled by the UNIX subdivisions of user, group and other, but by Access Control Lists (ACLs), with special AFS commands. An ACL was created for each experiment, and the list of accounts associated with the ACLs was maintained by the experiment. When users log in to the AFS client the operating system inquires the Kerberos[7] Authentication Server in that cell for a token. This token is valid for up to 24 hours, and is used to authenticate the user for privileged transactions, such as writing to AFS.

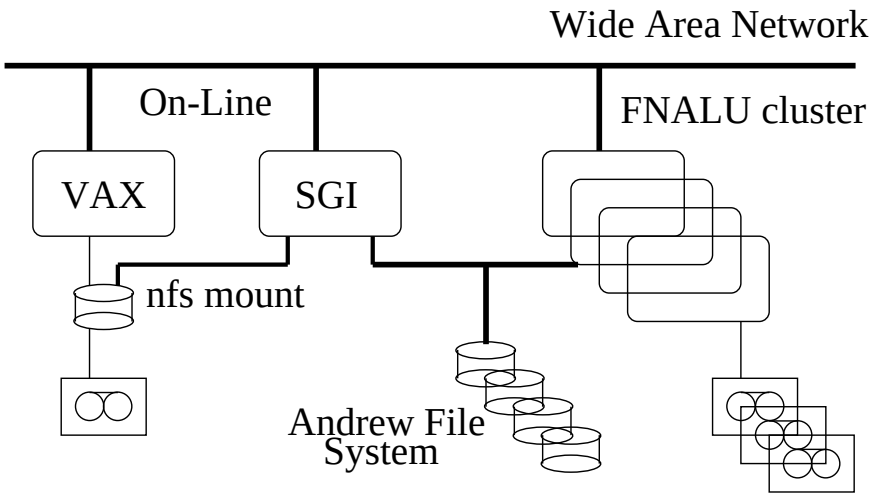

Figure 2: AFS at the experiment

4 Using FNALU

The FNALU system is a LAN of UNIX nodes which are being used to replace the existing FNALV VAX/VMS cluster. FNALU has computers running the AIX, SunOS, IRIX and OSF1 operating systems. Its primary use is for user support at Fermilab, such as electronic mail and documentation, however the large capacity for data processing makes it attractive for running analysis as well.

The batch facility is based on a package called LoadLeveler from IBM, however this is not fully supported on all nodes in FNALU. We have found that running jobs in the background, with lowered priorities (with the UNIX command *nice*) is sufficient.

To get the data from tape onto AFS disk, we looked at various methods. A tape mounted on the AFS client in the counting room could be dumped to disk at only 6 kbytes/sec. By using tape drives connected directly to the AFS servers we were able to attain data rates of 100-200 kbytes/sec. For runs which were already on disk, we used normal FTP which was faster than UNIX *cp* for copying to AFS.

5 Data flow at the experiment

Figure 2 shows the counting room view of the computing systems. There are two computers at the experiment, the front end VAX/VMS machine which collects data, and a Silicon Graphics R4000 Indigo used for both online and offline analysis. During the run, events are shipped to the online analysis computers over TCP/IP for monitoring and display. Once the data has been written to disk or tape it is copied to AFS, where it can be processed from either the experiments Indigo or any node on FNALU.

The data disk on the VAX is made available to the Indigo with MultiNet NFS, however, attempts to mount AFS to VAX/VMS failed.

Experiments also used the AFS disk for distributing code management between different nodes. Common code can be kept in one place without the need of tools to update code across the network.

6 Limitations

We found a few problems with AFS.

- Although 9 Gbytes disks are now used at Fermilab, there is a maximum size of 2 Gbytes per directory, adding an extra directory level to event data.
- There is no support for VMS, making the transition from VMS to UNIX more difficult.
- AFS clients have slow write access, forcing the use of other file copy programs, such as FTP.
- Due to the Kerberos authentication, logging in to the online SGI took several seconds longer once AFS was installed.
- Collaborators from institutions outside Fermilab would log in to FNALU, rather than use AFS to access data from their home computers.

7 Advantages

Our major advantage of using AFS was the reduction in load of disk management by the experiment. The large (50 Gbyte) disk supply is managed and backed up centrally at the Feynman Computing Center by computing professionals. Availability of the data is more reliable for several reasons: AFS (unlike NFS) can have multiple servers, the FNALU computers have better support and they are run in cleaner environments.

8 Conclusions

Traditionally experiments have used event selection, data summary tapes (DSTs) and ntuples to reduce processing time at the analysis stage. These methods generally reduce the time taken to transfer data into the CPU at the expense of throwing away information. Often this information is needed at the latest stages of analysis, to study systematic errors arising from software cuts, or hardware triggers. Using the large disk farms, it is possible to keep a large fraction of the raw data readily available. This is useful for experiments like MINIMAX where each run represents different trigger or apparatus conditions which require individual systematic error analyses. In experiments where the data were taken with identical running conditions, reasons for keeping all the data on disk reduce to problems experienced with tape handling. Not only is tape slower than disk, but the tape mounts are time consuming, and tape access is more complex. (On FNALU, the guideline is currently to service 50% of mount requests within 30 minutes.) Tape handling problems can be

reduced with tape robots and silos, however these are usually beyond the expense budgets of small experiments.

AFS clearly has advantages for APEX and MINIMAX, although it is new and we need to gain more experience with such file systems. In the future, more use may be made of AFS clients outside Fermilab. It should be noted that experiments will clearly expect to be able to use more disk in future experiments as it leads to greater flexibility in processing their data and computing centers will need to investigate new ways to manage this disk.

Acknowledgments

This work was supported by the Department of Energy and the National Science Foundation. We wish to thank our colleagues in Operating Systems Support for setting up and supporting the experiment AFS directories on FNALU. We also thank other members of APEX and MINIMAX for bearing with the new computing systems.

References

1. V. White et al., *IEEE Transactions on Nuclear Science* **34** 4, (1987).
2. AFS and Transarc are registered trademarks of Transarc Corporation.
   VAX, VMS, QBUS and OSF1 are registered trademarks of Digital Corporation.
   UNIX is a registered trademark of AT&T.
   SunOS and NFS are registered trademark of Sun Microsystems, Incorporated.
   AIX, LoadLeveler and VM are registered trademark of International Business Machines Corporation.
   SGI, IRIX and Indigo are registered trademarks of Silicon Graphics, Incorporated.
   MultiNet is a registered trademark of TGV Incorporated.
3. Please see **http://www-oss.fnal.gov:8000** for more information about FNALU.
4. CERN Program Library, CERN, CH-1211 Geneva 23, Switzerland.
5. S.Geer et al. *Phys. Rev. Lett.* **72**, 1596 (1994).
6. J.D.Bjorken, K.L.Kowalski, C.C.Taylor, "Proceedings of the 7th Les Rencontres de Physique de la Vallee D'Aoste", La Thuile. 507 (1993).
7. S.P. Miller, B.C. Neuman, J.I. Schiller, J.H. Saltzer, "Kerberos Authentication and Authorization System", Project Athena technical Plan, **E.2.1**, MIT, (1987).