HEPiX FNAL 94 Meeting Report

Alan Silverman

February 16, 1995

The Fall 94 North American HEPiX meeting was held at Fermilab Laboratory on October 13th and 14th. Some 35 people attended, representing about 12 sites, including FNAL itself of course, SLAC, Argonne and BNL. CERN was represented by Frederick Hemmer and myself. The meeting was broadcast on MBONE and this was judged a success by the organisers since it was watched more or less continually by more than 10 people and at any one time there were approximately another 10 or more people watching. It was agreed during the meeting that such broadcasts should be attempted for all future meetings if possible.

Although I shall try to summarise the main points presented in this report, the reader is urged to consult the World Wide Web entry for the conference where most of the foils can be seen; the URL is - http://www-oss.fnal.gov:8000/hepix/1094/talks/

1 Site Reports

After an introduction by Joel Butler, head of the Computing unit at FNAL, the meeting proper opened with the traditional site reports with one difference; the speakers had been requested to concentrate on changes since their last report to HEPiX.

1.1 FNAL - M.Wicks

Matt Wicks opened the session. She reported that almost all data acquisition systems at FNAL were now UNIX and that FNAL also now supported an astrophysics survey group. The big growth areas were the IBM and SGI, both farms and non-farms, and X terminals where the population had risen from 76 to 126 in the past year, mostly NCD but with a rising number of Tektronix. The number of SUNs installed had remained around 70 (around 10% of the number of UNIX systems at the lab) and SUN had become an officially-supported platform; there were still less than 10 HP Series 700 nodes and their first 2 DEC Alpha/OSF systems had arrived, including one for Business Services. But VMS was alive and well at FNAL although not expanding. Although AFS at FNAL would be covered later, it was noted here that FNAL's emphasis on SGI as a supplier affected their view of AFS (where Tranaarc's port had came late) and also DFS (where there were no known plans). They had

recently discovered that AFS did not work correctly on MIPS 4600 chip systems but that Transarc were working on this.

1.2 BNL - E.McFadden

Brookhaven's computer population could be considered as a number of distributed fiefdoms where many groups provide their own services and others charter the Computer Centre for assistance. The Centre offered remote system management and hardware support, the latter in competition to vendors, as well as some specialised services. They were getting an SP2 on loan from IBM and they already had a small Cray and hoped to obtain a Convex Parallel Server later in 1994. There was an ATM testbed.

1.3 University of Notre Dame - J.Bishop

That week, during HEPiX, their SP1 was being upgraded to an SP2 with 16 nodes. They used AFS on the SP1(2), shared with a SUN cluster. Their workstation population was relatively constant but their disc base had nearly trebled, including a recent addition of their first 4 Seagate 9GB drives. Physicists at Notre Dame participate in experiments at BNL and FNAL.

1.4 CERN - A.Silverman

The speaker presented a report of developments at CERN using the talk prepared by A.Lovell for the European HEPiX meeting scheduled in Paris in 2 weeks time.

The range of systems supported had not increased since the last HEPiX but the number certainly had, especially HP Series 700 workstations and NCD and HP X terminals. The CERN AFS service had been greatly expanded but would be covered in a later session. The Print Spool service was very popular and could be used to send print jobs from CERN to remote site printers or from remote sites to printers at CERN. Orders had been placed for dedicated X terminal boot servers, both centrally and remotely in an area with a concentration of such devices.

Atlas were now successfully using their Work Group Servers, and CMS were about to start using theirs. Similar services had been brought into production for several smaller experiments and a first Public UNIX Login Server (PLUS) would be established on the IBM SP2 then being installed. Much work had been done on the HEPiX login scripts and these were coming into general use.

A description of CORE services would be left to later.

1.5 CEBAF - S.Philpott

CEBAF were trying to set up a video conferencing service on an HP server but a non-standard kernel was giving support problems. This site was almost entirely HP-based but they were now looking at a possible second platform and were evaluating several different architectures. Another open question concerned a possible fast file service: should it based on NFS or AFS or DFS? And should they go for a centralised or distributed model?

1.6 Argonne - J.Volmer

They take security very seriously and so make heavy use of Kerberos, including use over WANs. They were in the process of setting up a heterogeneous DCE cell, including the use of the Gradient client for PCs. They had an SP1 with 128 nodes which was in the process of being upgraded to an SP2. Attached to it were 220GB of RAID and an 8TB DD-2 robot. NSL Unitree was also used.

A most interesting project was the Argonne "Cave", a virtual reality "room" where the walls were the VR screens. Part of the power for the models projected in the Cave came in fact from the SP2 as well as an Onyx for the graphics power.

1.7 SLAC - R.Melen

SLAC had come late to UNIX but were building fast. A B factory was planned for the end of the decade and this would use UNIX platforms. A central RS/6000-based batch UNIX farm was being built up currently and the speaker showed the plans for 1995; it would include some nodes for interactive use. They had investigated LoadLeveler and were now looking at LSF. SHIFT software was in use. They used STK robots and hoped to move to the "Redwood" product in the future.

The challenges for 1995 were VM migration, file backup and restore, the production use of AFS, moving SLD to UNIX, building up UNIX support activities and building a prototype computing farm.

2 Deficiencies in AIX for Large Systems - R.Alexander

This was a report from a Birds-Of-a-Feather session at a recent SHARE (IBM Users) meeting. A group of AIX users from large installations had prepared a White Paper on deficiencies found in AIX when used on large systems, which they restricted to single nodes but with hundreds or more interactive users and with or without a significant batch load. They had submitted this to IBM and presented it to the AIX development team in Austin. There were three main areas covered - scaling, infrastructure and the interlinked areas of security, integrity and reliability.

In the time since preparing the paper, IBM claimed to have fixed several problems in AIX version 4 but they had not yet replied to the other points listed and the speaker requested the HEPiX group

and/or individuals with similar systems to add their voice to the concerns expressed in the White Paper, which can be consulted on the Web at URL http://www-oss.fnal.gov:8000/hepix/1094/talks/general/

3 HEPiX Structure

There was an open discussion on a suggestion, originally put forward by Matt Wicks some months before, on unifying the two Chapters of UNIX (North America and Europe) into a single world-wide group. Discussion on the news group following the original proposal had resulted in broad agreement that the suggestion was popular, that meetings should be arranged to be alternatively in Europe and North America with specific meetings to coincide with the Computing in High Energy Physics (CHEP) series; and that HEPiX should encourage the formation of small working groups to work on issues of interest to HEP sites generally.

The audience at this meeting expressed some apprehension on the possibilities for intercontinental travel (both the cost and the bureaucracy involved). One lab stated that they believed it would be justifiable if HEPiX continued to prove useful. Another suggestion was to try always to transmit HEPiX meetings over MBONE as was being done for the FNAL meeting.

Possible topics for working groups included AFS, product distribution tools and trouble ticket systems.

The consensus view of the meeting was positive and the issue was passed to the European meeting scheduled in Paris two weeks later.

4 Batch Systems

4.1 SHIFT/CORE at CERN - F.Hemmer

F.Hemmer presented a comprehensive review of the progress and status of the SHIFT/CORE systems at CERN over the past year. The decision to rundown CERNVM over the coming two years had led to an expanion of capacity in SHIFT and also to the acquisition of an IBM SP2, due shortly. Much emphasis had been put in increasing the reliability of SHIFT and CSF, two of the constituents of CORE, especially the disc and tape service reliability.

CORE's share of tape mounts compared to those on CERNVM had risen significantly. Various robotic devices were installed or planned. Ultranet is still heavily used but more FDDI was appearing. A new tape stager was being implemented with enhanced robustness and better handling of concurrency and the control of tape stage space. Future work on the tape stager would include the provision of access control and request prioritisation.

Future plans for SHIFT included ports to new software releases, inclusion of SHIFT code in the CERN Program Library and various enhancements to RFIO.

4.2 LoadLeveler at FNAL - K.Fidler

LoadLeveler is the batch queueing system used on FNAL's CLUBS system. Among the reasons for choosing it were the fact that it is vendor-supported (IBM), it has better interfaces than NQS, it permits the definition of a specific job mix per node (but see below), it is supported on both AIX and IRIX, it allows central operation and it is easily scalable.

Experiences has shown up some plus points and some negative ones. First the positive ones: implementation on CLUBS was successful, it has proved very stable with a nice GUI for operator control; it has been found easy to add or subtract nodes and it has a simple script language. There is also a good blend with the interactive service on FNALU.

On the other hand:

- there are currently no program exits for administrator needs (scheduled for version 1.2);
- it uses a FIFO queueing and has no concept of fair shares in its job selection;
- job chaining and job dependencies are not supported (but some features are coming in version 1.2);
- no UNIX group controls (partly in version 1.2);
- CPU time limits are not scaled by CPU speed.

Better administration tools are required, for example a tool to rundown and restart by job class and better monitoring tools. We need more flexibility to define the job mix by node and the new alternate central scheduler to handle failover to be provided in version 1.2 will be very welcome.

4.3 LSF at SLAC - E.Russell

As noted above, SLAC had earlier evaluated LoadLeveler with a view to adopting it as their UNIX batch scheme. Now they were similarly testing LSF, although they were still only at a very preliminary stage. The speaker showed some very interesting overheads comparing the two products. LSF (Load Sharing Facility) originated from a small Canadian company. Platform Computing, and SLAC were using a beta copy of version 2.0. Among its features were the handling of batch and interactive loads, job sharing based on options more than simply CPU load on the target nodes, and so on. Refer to the overheads for more details.

Platform has stated that version 2 will support AFS but doubts were expressed as to whether this was standard or required a special version. LSF uses FlexLM as a licence manager and appears to have a number of advantages over LoadLeveler, at least the versions of the latter now in use.

Although LSF is Platform's only product, they have signed agreements with Digital, Convex and SGI already. The pricing of the product is broadly similar to that of LoadLeveler.

4.4 DNQS at BNL - E.McFadden

The original DNQS had come from FSU, had been modified by McGill and adopted by BNL. Changes at BNL were driven by user demand except for the addition of a GUI. The changes included -

- Graceful termination to allow a job to close itself down during a shutdown
- job dependencies
- delayed scheduling of jobs

5 AFS

5.1 AFS at FNAL - S.Hanson

FNAL has had mixed experience with AFS. They complained that AFS releases for the different architectures often fell behind the operating system release schedules; different bugs appeared on different platforms and particular examples were quoted. Sometimes bugs were fixed by patches to the operating system as opposed to AFS itself. Many problems had been found with the NFS exporter.

Currently FNAL used RS/6000s as servers but they were in the process of moving to SUN servers because they appeared to offer a more stable service. FNAL had no definite plans for DFS production, the lack of knowledge of plans for an SGI version being the most important open question, but they hoped to start a pilot DFS service in the near future. There would be a formal review of AFS in FNAL in November.

The speaker ended on a positive note, despite the preceding criticisms. AFS now offered a stable service, with well-working file backups, moving shortly to using DLTs. Future areas of interested included an HSM package from the University of Michigan, AFS tools, performance monitoring and a shrink-wrapped AFS installation procedure.

5.2 AFS at CERN - A.Silverman

The speaker presented the status of AFS at CERN using the talk prepared by R.Tobbicke for the European HEPiX meeting scheduled in Paris in 2 weeks time.

CERN currently ran 6 AFS servers, all RS/6000s, with some 220GB of disc space. Most discs were in Digital Storage Arrays and backup was performed to local DLT units using the AFS backup facility. The version of AFS was 3.3 (base) and there was a site licence covering all major architectures. The main uses were for user home directories and for the CERN Program Library and the ASIS public domain software repository. There were some 750 registered users including many from the major LHC groups and for them some 20GB was mirrored using standard AIX volume mirroring. Also

available via AFS were some vendor software kits and patches, X terminal fonts and control software and certain PC interfaces and applications.

Administration was performed using scripts and the sysctl utility from IBM and these included disc space administration tasks which could thus be delegated to project level. A scheme had been devised to extend the lifetimes of AFS tokens for use with batch jobs. The speaker had established dial-in linkage to AFS from a computer at home but of course performance was considerably influenced by the available line speed.

Key issues for the future included the consolidation of the service, the development of more administration tools, implementation into production of the token extender scheme for batch jobs. As the service grew, it was hoped that we could handle more disc space per server and that some form of hierarchical storage management would become available; it was known that Transarc were looking at this latter but there seemed to be no short term solution.

5.3 AFS at BNL - T.Nguyen

The speaker related a number of problems he had found using AFS on a SUN, both under SunOS 4 and Solaris 2. Some were recognised as having been seen elsewhere but several members of the audience claimed not to have seen others in similar circumstances. It was suggested that BNL are concentrating a number of AFS services on a single system, including NFS exporting which is generally recommended to be offloaded from an AFS server. FNAL in particular had specifically taken the decision to move to SUN servers precisely because they appeared to offer a more stable environment.

The speaker pointed out that they had found experimentally that performance was heavily dependent on setting a large enough file cache on client nodes.

5.4 Supporting Distributed Computing with AFS - M.Wicks

FNAL's UNIX Product Support (UPS) package has been described in previous HEPiX meetings. They now felt the time was ripe to introduce AFS into this package, for example creating AFS read-only replicas of UPS products spread across the lab. It was realised that both the products themselves and the product release mechanisms would have to be re-evaluated in moving to AFS and there were a number of open questions on how best to merge UPS and AFS.

6 Central General Purpose Computing at Fermilab - S.Wolbers

There was a still a serious use of VMS at FNAL, especially in the D0 and CDF experiments, and a wide belief that the price/performance of current VMS systems was equal to that of UNIX-based systems. Despite this, the FNALU service, described in previous HEPiX meetings, would be expanded, adding a 4 processor SGI Challenge and a 4 processor SUN SPARC 20 (a new architecture to FNALU) to the existing 2 SGIs and 2 RS/6000s.

CLUBS, the batch service also described previously, will be upgraded, as will CDF's Challenge; and D0 will also get a Challenge. Plans for the central batch farms were less fixed but more memory and I/O capacity were likely in the short to medium term, more CPU power later perhaps.

One interesting phenomenon was that the Fermilab business community was moving off an IBM 4381 on to an Alpha OSF/1 and an RS/6000.

Among packages being evaluated currently were ADSM and AFS Backup and Unitree for file archiving.

7 A Common X Environment at FNAL - J.Kallenbach

FNAL now had a significant number of X terminals, mostly NCD but with a few from Tektronix, including some 50 at people's homes, and a common X11 boot service had recently been established on FNALU. It offered a range of configuration files and these would soon be extended to permit users to save their own configurations. More information on services at FNAL for X terminals can be found on WWW at URL http://cdibm.fnal.gov/x-support.html.

8 Nirvana - P.Lebrun

Paul Lebrun gave a most interesting demonstration of Nirvana, a project to produce high-quality graphical user interfaces to HEP packages and tools. The presentation showed Nirvana as applied to the analysis of histogram data. More work was planned on this and future extensions were possible. Among the topics being considered were the use of procedural or iconic programming languages, C or C++, Motif or TK/TCL.

9 Installing Public Domain Software on UNIX at BNL - E.McFadden/T.Nguyen

This session was a description and demonstration of a tool known as SQIRT - Software Query, Installation and Removal Tool. It provided access to some 500 packages on 6 platforms and offered the users one line commands to copy packages to local discs or create links to packages; to remove them; to show space used; etc. All publicity was via the Web including access to the man pages.

10 Building Packages for Multiple UNIX Architectures - M.Mengel

The speaker presented some procedures and scripts used to produce binaries of FNAL and public domain software packages from a common source tree for use on multiple UNIX architectures via

FNAL's UPS scheme. There were options for version control and the actual builds were performed via rsh to the target platform. The method was based on NFS for source and build file access with lots of Makefiles and made much use of templates. Some lessons were presented based on the experience in use such as the advantages of creating build makefiles automatically rather than relying on users to fill in the templates by hand.

11 TCL/TK used in Data Acquisition - R.Pordes

This session described how TCL/TK was being used as a command line and graphical user interface in the development of a data acquisition system for some fixed target experiments at FNAL and for the Sloan Digital Sky Survey. Various reasons were given for the choice of TCL/TK and examples of applications with their corresponding positive and negative aspects were described. There were also examples of some added-value extensions made at FNAL to permit command line editing and parsing and the aliasing of verbs.

12 POSIX Standards Update - M.Wicks

FNAL has representatives in the POSIX groups on systems administration and on supersomputing. The speaker noted that in general POSIX attendance was down, perhaps as a result of economic pressure. They were no longer producing language-independent interfaces nor requiring test methods which may be helping to produce more readable standards more speedily although arguably affecting the quality of the resulting standards.

He gave a list of where POSIX was today with respect to the interests of HEPiX members. The 1387.4 sub-group on printing, part of Systems Administration, was basing its recommendations on Palladium. Both it and 1387.2, the sub-group working on software management based on HP's swinstall package, have completed their first ballots and the second one for both groups should be complete by early 1995.

13 Computing For Analysis Project

Two speakers closed the meeting with presentations on this new data mining experimental project at FNAL, one from a user's viewpoint, the other as a system administrator. Data Mining was defined in terms of selecting partial events from a very large data sample. They were using an SP2, an HSM store (IBM's version of Unitree) and a data object model to select desired events quickly and efficiently for detailed analysis, using a simple query language. D0's projected data sample was some 30TB of data and their DST was 1TB. The first prototype was just starting work.