HEPiX SLAC Meeting Report

Alan Silverman

December 20, 1993

The Fall 93 North American HEPiX meeting was held at SLAC from October 27th to 29th. The attendance list had over 40 names but there was never more than 25-30 people in the meeting room at any one time. However, it was nice to see quite a few University people present, as well as a few actual users, a welcome development from previous European and US meetings which had been largely attended by system-providers and largely from the major labs. The overheads of all sessions are available in the Computer Science Library.

1 Introduction

Chuck Dickens, Director of the SLAC Computing Centre gave the opening keynote address and described how they had arrived at a new computing model for SLAC. He said the rate of change of technology was so rapid that they were afraid of specifying any given precise model for future computing. Nevertheless, their goals were client/server computing; a centrally-switched network with the eventual goal of ATM; desktops, primarily mass-market based (which he clarified under questioning as mainly PC and MAC but not necessarily NT). He said the lab was too small to build all their own systems and should concentrate on only one or two types of fully-supported desktops. There should be common applications and common access to servers and the network and these should also be mass-market based.

The general strategy should focus on applications not systems and they should select applications which run on multiple platforms. He stated that they should "use popular objects from profitable vendors" which would seem to exclude purchases from IBM and DEC! Solid support structures were important and outsourcing might be used for support. Another quote was that they should "buy adequate technology not necessarily the best".

2 Site Reports

2.1 SLAC

This was presented by Chuck Boeheim, Server Computing Group Manager. Central UNIX support is run by some 11 people providing UNIX administration for the central servers and a certain number of workstations as well as support for local administrators. They had a small prototype workstation farm of three RS/6000 systems connected to 6 10-tape Exabyte stackers. Their experiences so far, early in the evaluation, were that code conversion from VM was not easy, that the exercise was labour-intensive and that Exabyte was not yet production quality. They were working with IBM's Loadleveler but some vital features were still missing such as fair sharing among users and a policy-based resource allocation scheme. Work was continuing with IBM and with other US labs on this product.

They had established a first AFS pilot and were in the process of acquiring ADSM from IBM for UNIX file backup. The testing of UNIX systems connecting to tape silos was also underway and they would move the ADSM master to an RS/6000 in due course as part of the migration away from VM.

SLD and the newly-approved Beauty Factory (official confirmation came through during the second day of the meeting) would lead to greatly-increased data rates and they were upgrading the STK cartridges to 4490 drives giving 24TB online. The next stage could be helical scan devices from STK in 2-3 years and could offer 600-900 TB if the current 4 silos when re-equipped.

ZTAGE from IN2P3 was currently in use but NSL Unitree from IBM would be evaluated for staging and managing experimental data and FATMEN for tape staging and file management.

2.2 TRIUMF

It was stated that having an AT&T source licence gives unlimited user access to UNIX systems; to be verified in connection with the recent Solaris licensing issue. [In checking with Dietrich Wiegandt, this is in fact only true for the particular system mentioned in the contract with AT&T.] TRIUMF took a very pragmatical approach to UNIX support - select only well-tried packages and involve the end-users as much as possible. Probably very wise for a small site. Also, choose a single vendor (DEC in their case).

2.3 FNAL

See Pisa notes. Again use of Loadleveler and evaluation of NSL Unitree were mentioned. AFS at FNAL is discussed later.

2.4 Yale

The first University to report, the speaker was openly seeking answers to many questions (OSF, AFS, etc) and pleading for help and guidance from the major labs.

2.5 Notre Dame

A small US university site, partly VMS and partly UNIX, but a large AFS installation used by their students. Soon they will install an SP1 with 16 nodes. Running experiments at FNAL and BNL.

2.6 BNL

This was presented by Ed McFadden. A total lab population of about 450 UNIX systems, two-thirds SUN, plus 80 VMS and over 1000 PCs; most UNIX was around the Computing Centre (CCD). They had a small computing farm, mostly RS/6000; an Alpha loaned by DEC had been returned and a loaned SGI Challenge was likely to suffer the same ignominious fate. The problem in each case was lack of user interest, possibly caused by the fact that users were required to pay CCD for services.

They had bought a WORM device for tests and then discovered that the platters were too expensive ($325 for 6GB).

3 CVS

Paul Kunz gave a plea for people to use CVS for program source control instead of SCCS or vanilla RCS. It was now in wide use at SLAC and also at LBL, FNAL, CERN (Phil Defert), etc. The Internet news group for CVS showed a healthy interest from the commercial world also; at Thinking Machines, 100 software engineers worked on a project with a single 1GB source directory tree under CVS control.

4 GUI Programming

Paul Lebrun presented some work done at FNAL to develop a sort of style guide for MOTIF-like programming to promote simple-to-learn and consistent graphical programming. It was implemented as a set of widget libraries. In passing, Paul mentioned a new popular WYSIWYG X editor in the public domain called NEDIT.

5 AFS

5.1 CERN

I presented the AFS talk prepared by Rainer Tobbicke for the October IBM SHARE Europe meeting. When we came to the issue of long-lived tokens for batch jobs, Keith Rich of SSC mentioned a couple of points: Kerberos had the notion of a "promise" of authorisation which could be kept dormant until the job was actually scheduled. Plus he had written a wrapper round the token which could be used to prolong the life of the token while the job executed. He agreed we could contact him for more information. There was great interest in the AFS User Guide which Rainer is currently writing.

5.2 SSC

Their prime servers were on ULTRIX but they also had SUNs. There was a total of 7 servers, 41GB of disc space; some 150 AFS clients; 3 AFS/NFS translators to which some 200 workstations attached from time to time. No problems seen with AFS/NFS translation but, like CERN, they had not stressed this.

5.3 FNAL

Severe problems to report. There were few real clients as AFS was simply used to serve the CLUBS system via the NFS translator. When this had been run on an RS/6000 AFS server, it had crashed under load. Sometimes files had "looked" corrupt and if they were re-written in this state, could actually end up being corrupted. K. Rich (SSC) emphasised that the NFS translator should not be run on a server for reasons of load on a single system and noone else in the audience had seen a similar problem. FNAL had moved to using a beta test version of the SGI port; a production version was not expected until SGI would release IRIX 5.1.1, now due 1Q94. The impression was given that FNAL were alone in their bad experiences with AFS and had perhaps, with hindsight, taken some wrong choices in their setup.

5.4 SLAC

After initially (spring 92) deciding to wait for DFS (expected then to be 6 to 12 months away), SLAC had recently decided to install a small (6GB) pilot and now planned to expand this next year since DFS still looked 6 to 12 months away (this became the key phrase of the meeting). They used an RS/6000 for file servers, and SUNs for volume location servers. They had even installed the NFS translator on the IBM and had seen no problems yet but it was little used. After the FNAL experience, they would rethink this (HEPiX CAN be useful).

5.5 Other Points

There was a big demand from all AFS sites present for more and better AFS documentation for administrators. ADM was used by some sites for AFS administration tasks and a new tool, sysctl from the IBM T.J.Watson Centre was mentioned; it might soon become freely available (see my forthcoming LISA Conference trip report). Another worry was the price which might be charged for DFS licences.

6 ASIS

Philippe Defert presented the plans around ASIS. During this it was pointed out that it should NOT be described as offering "public domain" software but "freely available" software.

7 FREEHEP

A repeat of the Annecy CHEP 92 session. Later in the meeting methods were discussed to cross-link FREEHEP with some projected HEPiX initiatives in building up a database of useful UNIX tools for HEP sites.

8 Farming at FNAL

8.1 Data Mining at FNAL

Large computational needs led to workstation farms, some for CPU-intensive jobs, some more I/O oriented. The talk described some useful tools used on their farms (see overheads) and some interesting benchmarks were presented comparing farms with mainframes, supercomputers and MPP systems.

The CAP (Computing for Analysis Project) was an I/O-intensive farm utilising parallel I/O for experiments with between 25 and 200 TB of data each, using robotics and HSM. The aspect of data mining came from searching for specific data samples. The prototype was based on the SP1 and three were installed, to move later to SP2s. The data cache was a large Exabyte robot device with 8 drives initially. Files are transferred into the 8 I/O nodes using 20 MBps SCSI and fed to the 16 processing nodes. They expect an overall rate of 8 to 10 MBps for effective tape reading into the batch nodes.

In passing, the speaker noted that out of some 100 Exabytes drives in the Centre, they "only" had about one drive repair per day. The following speaker, from the same group, spoke of "many" problems with Exabytes, one per day. Life viewed from opposite ends!

8.2 Operator Console at FNAL

This was a tool to help with tape drive allocation and providing instructions to the tape mounting operators.

8.3 Robotic Tape Control

The other side of the coin, this was a utility to control tape mounting inside robots, Exabyte and other. It was not (yet) linked to the previous tool.

9 DNQS at BNL

BNL had taken DNQS from McGill and used it in CCD. But they "missed" HEPVM batch and would love someone to port the best features of this to UNIX. Meantime they were looking at DQS (also from McGill) and Loadleveler.

10 COSE

Judy Richards arrived hot-foot from San Jose further down Route 101 and presented her impression of the CDE initiative of COSE.

11 Vendor Talks

11.1 SGI

The first of three invited vendor speakers, Forrest Baskett is a senior VP of SGI (a founder?) and head of Research and Development. He presented the company and its products before moving on to describing why SGI believed in shared memory for parallel programming projects. He described a cluster of 16 20-node Challenges which had been used for a turbulence flow calculation with impressive results. He did however state that SGI were also looking into distributed memory solutions using High Performance FORTRAN for some future systems.

11.2 IBM

The most entertaining talk of the week was given by Merrit Jones of IBM's Metro Systems Division in Houston. He is a systems integrator who said he could hook any cluster configuration together as long

as he had the cables and the drivers. He would happily select any of the four most popular methods of clustering - bus, star, ring or cross-bar; he used a huge variety of links including VME, Ethernet, HIPPI, etc, etc.

He was involved in the project at the National Storage Lab (NSL) to develop a network-attached high throughput storage system. Many vendors are involved including IBM, Ampex, DISCOS, etc., plus user sites such as LLNL and four US National Science Foundation Supercomputing Centres. The inclusion of the Pittsburgh Centre implied the use of AFS which would thus have to be integrated with Unitree. It was from the NSL initiative that IBM had developed and were making available NSL Unitree.

The initial target was 50MBps sustained rates moving later to 500 MBps. He could handle many kinds of robot including IBM's 3480 cartridges, Exabytes, IGM (Summus) and "protein robots" - his term for humans.

A 128 node RS/6000 cluster is going into Argonne very soon linked by a HIPPI switch and using 4 file servers connected by Fibrechannel.

His talk was very well presented, he was eager to impart information. When asked about tapes, he explained that he was "not a tape expert but (he) had once sat next to a guy on a plane who had known one". He then launched into the relative benefits of helical scan (good for price and density but you sacrifice shelf life and reliability) against longitudinal (long lasting and low error rates).

11.3 DEC

The last (poor) talk was given by a guy from the High Performance Computing group in Maynard and was largely marketing about why Alphas were so good for HEP farming. Strange that it was almost the first time anyone had mentioned using Alphas in that role in the whole meeting.

12 Closing

The next meeting was provisionally scheduled to follow, immediately if possible, the CHEP94 conference in San Francisco next April and the attendees were then freed to enjoy the rest of the day with its 30 degree sunshine (a late-October record for San Francisco).