HEPiX SACLAY Meeting Report

Alan Silverman

April 12, 1995

Two weeks after the FNAL HEPiX meeting, the European Chapter held their autumn meeting at SACLAY, Paris, hosted by DAPNIA. There were about 70 delegates representing over 30 sites in Europe plus 2 US sites and KEK Japan. The meeting was opened by Pierre Borgeaud, the Chef de Service of DAPNIA who described the work done at DAPNIA and its relation to other science labs in France.

Many of the overheads presented at the meeting are available on the World-Wide Web at URL – http://wwwcn.cern.ch/hepix/meetings/saclay94.html.

1 Site Reports

1.1 DAPNIA - P.Micout

Since July this year, there are no more VM or MVS facilities at Saclay and users desiring these must access CCIN2P3 in Lyon. Some DAPNIA staff use BASTA in Lyon or the Cray in Grenoble. There is an SP1 in Saclay which it is hoped will be upgraded to become an SP2 before the end of the year but it is not yet used by DAPNIA groups. Since the last European meeting (Pisa in October 1993), DAPNIA have added some 20 SUNs and a few other workstations including HPs, Alphas and RS/6000s. The usual range of central services are available (news, ftp, WWW, etc) but are closed to external access by default; it is hoped to be permitted to open some of them shortly. In the future they plan to make use of CCIN2P3's Anastasie service and are looking seriously at an Epoch-like backup tool and at AFS/DFS.

1.2 CERN - A.Lovell

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.

1.3 FNAL - M.Wicks

FNAL reported that almost all data acquisition systems at FNAL are now UNIX and that they also support an astrophysics survey group. The big growth areas are the IBM and SGI, both farms and non-farms, and X terminals where the population has risen from 76 to 126 in the past year, mostly NCD but with a rising number of Tektronix. The number of SUNs installed has remained around 70 (around 10% of the number of UNIX systems at the lab) and SUN has become an officially-supported platform; there are still less than 10 HP Series 700 nodes and their first 2 DEC Alpha/OSF systems have arrived, including one for Business Services. But VMS is alive and well at FNAL although not expanding. Although AFS at FNAL will be covered later, it was noted here that FNAL's emphasis on SGI as a supplier affects their view of AFS (where Tranaarc's port came late) and also DFS (where there are no known plans). They have recently discovered that AFS does not work correctly on MIPS 4600 chip systems but that Transarc are working on this.

1.4 INFN Pisa - S.Arezzini

They have built up their UNIX population to some 20 stations, a mix of Alpha/OSF, HP Series 700 and RS/6000, along with X terminals from IBM and HP. They are using the HEPiX login scripts and they are testing COSE/CDE on RS/6000. They have two AFS servers hosting software distribution tasks and some user home directories. They have a version of CERN's ASIS which they call pisASIS which, like CERN's, is AFS-based. They have an Mbone interface on an SGI Indy and they are looking at adding more AFS in the future as well as an implementation of SHIFT and possibly Loadleveler.

1.5 DRAL - J.Gordon

They are using a version of the HEPiX login scripts and also CERN's NQS, but due to perceived limitations in the latter, they are looking at other batch queueing products, including LSF. HEP UK stopped using VM in April 1994. DRAL's OSF cluster has been upgraded and a 4 drive 3494 SCSI-based robot has been installed in the Centre. Work is underway to upgrade SUPERJANET to 10 Mbits. MBONE traffic to CERN is slow and it was discovered that it passes (twice) across the Atlantic "en route". They have a mail service based on POP3 and use Zmail for PCs; in the future they are looking at the IMAP protocol.

1.6 Prague Physics Institutes - J.Hrivnac

In fact there were 4 physics-related institutes in Prague with a total of some 60 physicists and the speaker summarised the different computer configurations being used and their links to international collaborations such as DELPHI, ATLAS and RD41 at CERN and H1 at DESY. They already used NQS and NQS++ to submit jobs from Prague to CERN and vice versa. They had HPs and SUNs, using VUE and Xdm respectively and were looking at COSE/CDE. They hoped to adopt the HEPiX scripts and would be most interested in a similar HEPiX initiative to develop a standard X11 environment.

They cloned ASIS locally by regularly transferring tapes and would like to find a way to fund a client AFS licence within the CERN AFS cell if they could be sure the network between Prague and CERN was reliable enough and had enough bandwidth.

1.7 CASPUR - A.Maslennikov

CASPUR was equipped with Alpha stations, SUNs, RS/6000s and PCs running UNIX. Their mainframe would be downgraded shortly from 6 to 3 processors and an IBM SP2 installed. On a central, mainly Alpha cluster they ran a CSF farm. AFS was in use with 3 servers in CASPUR and clients in Rome and Naples and was used to access ASIS mirrored from CERN and locally-stored vendor software. One AFS problem which remained to be resolved was a crash of the AFS cache manager when running on OSF/1 version 1.x and they hoped that moving to version 2 of OSF/1 would solve that. The speaker expressed his appreciation of help offered by CERN staff in setting up their configurations and also of Tim Bell of IBM, described as European HEP AIX/6000 support.

They hoped to move to using LSF version 2 with AFS support for batch job processing and they would move some parallel applications to the SP2. They would like to build up their AFS installation and perhaps create a single AFS domain for all of INFN in Italy. Tests were underway of a PC AFS client.

1.8 Uni Dortmund - K.Wacker

They had a 7 node RS/6000 cluster as well as about 7 workstations in offices and Loadleveler was now used across all of these instead of NQS as was the case at the time of the previous Dortmund report. The speaker showed an interesting chart of positive and negative points about Loadleveler. Their X terminal population continued to grow as does their disc space, including 7 new 9GB discs then being installed. While the full CERN SHIFT software was not used, the disk pool manager from SHIFT was, despite the inability to pre-reserve space and the fact that the manual was not up to date with the software.

1.9 CCIN2P3 - W.Wocjik

Since the last report, work on Sioux, their interactive UNIX farm, had reached an advanced state and production should be open soon. They had seen no drop-off in VM use but BAHIA, the front-end to Basta and Anastasie (see previous reports) continued to be popular. They were now using AFS more and more with RS/6000 servers and both SCSI-2 and RAID storage. Access to scratch and stage space was via RFIO calls.

The speaker showed some performance figures on tape access between RS/6000s and StorageTek robots.

1.10 KEK - T.Sasaki

In March this year, they upgraded their network backbone to use a Gigaswitch to connect their 14 FDDI rings. Some of their mainframe equipment was due for replacement over the coming two years and specifications were being drawn up for the computing needs of the planned B factory. Fujitsu had won a tender to supply a new Supercomputer, due in January 1995. Workstations supported included those from HP, SUN and Digital.

1.11 NIKHEF - W.Van Leeuwen

They have used two SGI Indy workstations for several successful teleconferencing pilots. They have obtained offsite AFS client licences from CERN and installed them on some nodes in NIKHEF. Although installation was easy, understanding AFS access control was less so and they have discovered that some applications with byte-locking do not work properly under AFS.

The speaker made a plea for more understanding on the use of postscript to prepare overheads which could be made accessible on the World Wide Web. He made a number of suggestions and he was urged to document these in some form which could perhaps later be expanded by others from their experience. In this way we could build a model of how best to prepare overheads for display on the Web.

1.12 DESY - J.H.Peters

DESY has reorganised its central computing structure and the new structure was presented. On the hardware side, there was now direct access to the STK (3480/3490 cartridge) and Ampex (D2) robots; the Ampex robot was already full after only a few months running. ADSM was used for UNIX file backups although access to SGI discs was only via NFS. 220GB had been backed up so far, rising at some 4GB per week. Archiving was not yet supported.

The main UNIX platforms in the Centre were still SGI Challenge nodes (there are now 6 of them) with 2 each for H1 and ZEUS. One of the H1 nodes was for interactive use but most of the others

were dedicated to batch reconstruction work.

It had been decided to concentrate work on parallel computing in Zeuthen and 2 Quadrics and an SP1, to be upgraded to an SP2, had been installed there.

1.13 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 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.

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.

1.14 FNAL HEPiX - J.Nicholls

A report was given on the North American Chapter meeting of HEPiX, held in Fermilab 2 weeks earlier. Some highlights were noted such as the use of WWW and workstation-based demos for some talks and concurrent MBONE broadcast of the meeting. For more details, the reader is referred to the minutes of that meeting as described in an entry under the FNAL Meeting page of HEPiX in the World Wide Web.

2 POSIX Report - 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.

3 High Availability X11 at DESY - T.Finnern

T.Finnern reported on progress in supporting X11 at DESY. Major deficiences found included the use of the unreliable tftp protocol and NFS-mounted fonts. To obtain high availability, one needed to be independent of the various servers needed by X terminals. One needed to be able to guarantee good performance, to provide redundancy (no single point of failure) and to have automatic failover when problems occur.

Definitions of a suitable server had been drawn up and different approaches were discussed on how to provide a highly reliable and highly available service. Finally DESY have decided on IBM's HACMP with a switchable configuration and the equipment was expected in November. They were planning to use PC-Xware and had purchased 500 licences. They would investigate load balancing across multiple boot servers.

4 cfengine - M.Burgess

This speaker, from the University of Oslo, presented a scheme he had devised to configure a large number of workstations of different architectures in a standard manner. It was based on defining classes of systems where a class could be defined in various ways - by hostnames in a list, by operating system, by owning group, by the day of the week in which the task was being done or via some arbitrary variable set on that node.

For these classes he defined certain actions, with the possibility of inclusion and exclusion lists. Actions might depend on combinations of classes to which a node belongs. Actions could be to create net masks, to check file links or file permissions, to mount file systems, to clean up temporary files, etc. Hooks existed to call user scripts.

The software was written in C and ran on the most common UNIX platforms. It was in fairly widespread use in the University although he had always to overcome understandable reluctance from other system administrators to hand over administration of their nodes to an automatic procedure. Nevertheless, the GNU Foundation had expressed interest in putting this software on their freely-available product list.

5 Supporting Packages on Multiple Architectures - M.Wicks

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.

6 Discussions on HEPiX Structure

This discussion was prompted by a suggestion made by Matt Wicks on the HEPiX news group in the summer to unite the European and North American Chapters of HEPiX and to hold joint meetings henceforth as well as to establish small working groups on specific topics. At the recent US meeting

in FNAL, the idea had received general acceptance although with some reservations on the cost and red-tape connected with intercontinental travel.

The Europeans were even more sceptical on these aspects although they too were in broad agreement with the ideas, especially the working groups. Such groups should work mainly using electronic means of contact with perhaps working meetings 1 or 2 days before future HEPiX conferences; and they should be encouraged to report at conferences.

After some discussion, and a tentative schedule for future meetings was proposed alternating between Europe and North America every 6 months with a meeting to coincide with CHEP meetings, it was agreed to form a single world-wide HEPiX group. It was further agreed to form a steering committee and some initial working groups, most notably one on AFS. It was suggested to try to include at least one UNIX user in the steering committee in order to encourage a constant focus on user issues.

7 AFS

7.1 AFS at CERN - R.Tobbicke

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.

7.2 ASIS at RAL - J.Gordon

RAL had implemented a version of ASIS from CERN with help from Ph.Defert in order to allow them to share software, including the CERN Program Library. Their first attempt was to use regular tape copies but this rapidly failed as it was always out of date. Next they had tried mirroring the ASIS software via ftp and then NFS exporting the result to the rest of RAL with the use of the EPIP utility for some clients to install packages on local discs. The revised ASIS was adopted in January 1994 and EPIP was replaced by version of make_asis, as used at CERN.

Finally they had moved to AFS, taking advantage of the purchase offer of an AFS client licence in the CERN cell. Now RAL used the lfu package via AFS to update a local copy of the software nightly which could then be accessed from within RAL using NFS. This was judged to be successful and more reliable than mirroring and ftp. RAL were now looking at a possible extension to this scheme to access some experimental data.

7.3 Cross-realm Authentication - R.Tobbicke

Essentially, a Kerberos realm is equivalent to an AFS cell, an administratively independent domain. The requirement is to grant to a user in a foreign cell access to local files via a valid token without requiring him or her to have an account in the local cell. One solution is that the remote user logs into his or her cell and then "crosslogs" to the local cell and the two Kerberos realms perform the necessary authentication. The user is then granted access to local files without being required to specify another password.

7.4 AFS at Pisa - M.Davini

PISA's cell consisted of 1 RS/6000 server with 3GB of disc space and some 16 clients, both RS/6000 and HP. A second server was planned. It was used for home directories and pisASIS, a local implementation of CERN's ASIS. Another use of note was piCAP, a scheme for the distribution of software for MACs and access to some home directories via Appleshare. They were now working on a similar scheme for PC users as well as permitting MAC users to manipulate AFS ACLs directly.

7.5 AFS at FNAL - M.Wicks

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.

AFS was now considered a stable service at FNAL, 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. In summary, the speaker stated that there had been a significant improvement in their view of AFS since the HEPiX meeting in Spring at LBL but the SGI implementation, very important to them, still lagged somewhat behind those on other platforms.

7.6 AFS and Batch Jobs - R.Tobbicke

The issue to be resolved was how to supply an AFS token for a batch job which would remain valid irrespective of the length of time the job stayed in the input queue, and/or its execution time. When a user responded to a request for an AFS password, he or she received a Kerberos ticket-granting ticket proving that person's identity. Thereafter that person could be granted a valid ticket for a given service such as AFS but that ticket had a maximum lifetime for security reasons, typically 25 hours by default.

One method of applying this to batch jobs could be to tell the batch queuing system, e.g. NQS, the password in coded form which could then be used to obtain a valid ticket at the moment when the job began execution. However, there were risks of password breach (which had originally been reduced by AFS since passwords under AFS did not cross the network and were not stored anywhere) and also a stored password would become invalid when the user changed passwords.

An alternative, used in IBM's Loadleveler product, was to pass the service ticket to the batch system with the job and then provide the batch system with a safe method of ensuring that this ticket, which would now be independent of the user's password, was re-validated when the job was queued and while it executed. This could be handled by a specially-written token server which converted old or expired tokens into valid ones. Such a server, which must of course be run on a trusted host, had been written using sysctl and acted as a front-end to Loadleveler. Discussions were underway on extending this scheme to CERN's NQS.

7.7 AFS in Saclay - P.Micout

Saclay had installed client licences for the CERN cell. Apart from a problem when the clients were booted immediately after the installation, the systems have behaved well since. They found world-wide file access most useful but sometimes got confused by the different file protections used by different users. The use of local cache had proved its worth in the repeated use of certain files. Now Saclay were considering how to expand their use of AFS or wait for DFS. In the meantime, the lack of AFS for the latest versions of DEC's OSF/1 is a problem.

7.8 Fully Automated AFS Backup at Caspur - R.Gurin

The main requirement was to simulate a UNIX drive in their STK cartridge robot. A pseudo-driver was written for UNIX, communicating with the VM host of the STK via TCP/IP. An AFS application on the UNIX side then uses this pseudo-driver to mount tapes on the robot and then "writes" files to tape. Speeds up to 430 KBps had been seen using AFS backup.

7.9 Progress Report on AFS at IN2P3 - S.Ohlsson

IN2P3 now had 2 AFS servers, both RS/6000s, with some 70GB of RAID and SCSI-attached disc space and more on order. They had found both AFS itself and AFS administrative non-intuitive to learn: in particular tuning was necessary, ACLs were "tricky" and hard to understand, and AFS modified the behaviour of some common UNIX commands.

They had migrated through several versions of AFS releases and were not impressed by Transarc's procedures, especially since some problems remained open and crashes still occured during AFS maintenance. Despite Transarc's recommendation to run with 3 database servers, they had found problems with this and ran with only one. Some performance figures were shown, comparing AFS and NFS.

They had devised a scheme for AFS use with their DQS batch system for acquiring long-life tokens but would be looking at Tobbicke's latest suggestion (see above). File backup was via WDSF, which saved ACLs as well as file date but missed files which users had a lock on. They had created a more user-friendly interface to the AFS commands for users and they had developed their own tools to delegate some privileged operations.

In summary, they found AFS very powerful and useful, better than NFS in performance and easier to use in a production environment but it needed training to use both for users and for administrators and there was lots of scope for improvement, some of which HEPiX could take on.

7.10 First DFS Tests at CERN - R.Tobbicke

For his last AFS presentation of the day, Rainer spoke of his initial tests of DFS, the presumed follow-on to AFS. Most vendors, SGI being almost the only exception in our interest area, promised future DFS support and he had early implementations for AIX/6000, Solaris, HP-UX, and DEC/OSF; not all worked as they should since only the first two were production releases at that time. DFS (will) offer (almost) all AFS functionality and ACLs were closer to UNIX-style permissions that AFS.

Rainer pointed out that a site needed to order a complete DCE server package, not just the DFS part; this should include a DCE security server, cell directory server, etc and an administrator needed to understand the various linkages inside DCE.

Obviously the migration from AFS was of paramount importance. With the AFS/DFS protocol translator from Transarc, one could access a DFS filebase from an AFS client.

Installation of the software had been straightforward, easier than AFS, but the terminology was different.

The first pilot application, accessing the HP-UX documentation, seemed to work successfully so far and further pilots were planned but much work remained to be done before opening any kind of general user service.

7.11 DCE/DFS at KEK - T.Sasaki

KEK had installed the HP implementation of DCE/DFS and HP supplied some local support, including a short course for prospective users. It was a small test cell together with the SUN DCE client from Transarc. This version had a number of missing features, including no ACLs. Setup via a GUI had been easy and much valuable experience was being acquired. Some performance figures were presented. However, current performance was still poorer than with AFS and more work was needed on interoperability.

8 User Migration

8.1 User Migration at CERN - M.Marquina

The target was to move the CERNVM community off the mainframe by the end of 1996. Already at the end of 1994 most of the batch load would transfer to CORE with a halving of CERNVM capacity at that time. Thus the major concern was on how to move the interactive users. One of the most important aspects of offering these users services elsewhere was to provide them with user-friendly tools such a mailer, a mail list handler, keyword search of the phone and electronic mail databases, a nice and simple editor, access to the user account registration database (CCDB), etc.

A group was set up to manage this - the UMTF or UNIX Migration Task Force - and it had established small working groups to make recommendations in specific areas, not only on which tools to adopt but on support policies for these. There now existed a menu of recommended products in different fields, sometimes one in a given area, sometimes two. Gradually more and more areas were being tackled.

It was felt that more publicity was needed for users to assist in the selection of such tools and in informing them when choices are made. It was agreed that more direct help should be offered to users, for example simple user documentation, short training courses (several hours in some cases, up to several days for more complicated products).

Marquina summarised his feeling that it was a tough job to bring UNIX to HEP user desktops and that user support is the key.

8.2 User Migration at DESY - J.H.Peters

Downsizing of the DESY IBM was in progress, the single large mainframe having been replaced by two smaller CMOS systems which had about 40% of the capacity of the original system but which cost some 10% only of its price.

Most batch was going (had mostly already gone) to SGIs (the three largest experiments accounted for some 80%). Data storage was based on SGI/Ampex/FDDI; H1 had migrated to this setup, ZEUS would move in 1995. The remaining batch was constrained in order to encourage migration.

The support team had written a special program to transfor and transform complete NEWLIB libraries to UNIX. They had produced a list of recommended tools for users and had organised user tutorials. Documentation available includes flyers and command cards as well as more traditional notes and guides.

9 Standard UNIX Environment

9.1 SUE - Shrink-wrapped UNIX Environment - R.Tobbicke

The goals of this development, which originated from an idea by R.Tobbicke and T.Bell of IBM, were

- to help the user without system admin skills to nevertheless install a workstation
- to produce common workstation configurations in order to reduce errors and to decrease eventual support workload for system staff both to install a new workstation and to help maintain it afterwards
- to encourage the use of recommended configurations.

It was assumed that if we defined a reasonable and working default configuration then users would not need or want to change this, at least not in a major way. It was destined for systems on the CERN site and required network connectivity and AFS.

Rules were defined for which SUE features were mandatory (like some network parameters for example) and which were optional and could be selected as part of a group-defined environment (if ADSM backup was to be run for example, or if zephyr was to be used).

After the initial installation, after the vendor's operating system is installed, SUE update scripts should be run on a regular basis, default nightly, to keep the configuration up to date and consistent with the latest software. The frequency was controlled by cron jobs and could be altered by local administrators.

The current status was that AIX had been the original implementation and was working and an HP-UX port had also been done. Work was starting on a Solaris port but a revised definition was

being prepared based on the experience so far and the existing ports would probably have to be revised according to the result of this exercise.

9.2 Introduction of the HEPiX Scripts at CERN - A.Taddei

The goal of these scripts, developed initially by DESY and further worked on by a DESY/CERN joint collaboration, was to offer a common user environment across a broad range of UNIX architectures and a broad range of UNIX shells. However it was pointed out that this gave a very large number of possible combinations of architectures times shell times methods of login.

The project had been split into two phases, the first concerned the login scripts themselves and this was what was entering production in CERN now; the second phase would be to work on a standardised X11 environment. Although our principle target was new users, obviously current users were welcome, even encouraged, to adapt to using our recommended environment as well. The list of CERN groups and experiments already using the scripts was shown.

Scripts can be installed on a system to be used by everybody (with the option of an exclusion list for those people who could not move to using them) or by individual users at their choice.

Features of the scripts included enabling powerful advanced shell features where these existed, the possibility of customisation of the scripts at system, group and user level and making them AFS-sensitive. Taddei had developed a "compiler" which took as input the standard script and output a version tailored for a given architecture, for example to replace the name of a utility by its correct path name on that architecture. The compiler also offered some debug capabilities. He listed some problems and drawbacks that he had solved while implementing the scripts at CERN and he closed by posing some open questions.

9.3 New Developments in UNIX Computing at DESY - K.Kuenne

The speaker described the structure in UNIX Computing at DESY as "islands of solutions".

- Apollo clusters: gave many problems, not quite UNIX at some level, vendor support dropping; but still 27 nodes with over 50 users.
- HP cluster: now 15 nodes with over 180 users; most nodes run dataless, stability problems but now improving; static assignment of a user to a node depending on where his home directory is; no batch and future cluster support from HP unclear.
- SGI cluster: 7 Challenges, some dedicated to a given application such as data reconstruction; problems include lack of features in batch (e.g. no quota per group) and user processes frequently in conflict with each other because of the lack of a resource reservation feature; a few stability problems.

The UNIX group have set themselves some challenging targets in RTBF and response time for trivial commands. They will also work to improve the batch system, provide a more common environment

across the clusters and a reliable file backup scheme.

Possible solutions to meet these goals include a UNIX mainframe where users connect via X11 but they believe user and resource conflicts would still remain and they have doubts on its scalibility. An alternative is a distributed workgroup model where the drawbacks would include the question of load balancing and handling large amounts of data online. However, this second option is more scalable, offers better overall stability and allows them the possibility of isolating particular applications and therefore DESY has decided to go in this direction.

A set of central services is being defined (mail, file backup, news, etc) and AFS file services will be used. Nodes for the first user groups have already been ordered.

10 CORE Services and the SP2 at CERN

10.1 CORE Update - 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 years had led to an expanion of capacity in SHIFT and also to the acquisition of an IBM SP2. 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 was still heavily used but more FDDI was appearing. A new tape stager was then 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.

11 The SP2 at CERN - H.Renshall

The speaker started with a presentation of the Service Definition for this new system which consisted of 64 Power2 nodes interconnected by a high performance switch. Disc space would be initially 120GB but more would be added in 1995. Some of this disc space would be dedicated to a public staging spool.

The principle applications would be as a replacement for CERNVM for certain services and as a data server. Some tests would be done on parallel applications suitability, for example PIAF, LHC design codes and parallel GEANT. It was expected to start with one quarter of the nodes for interactive work, one quarter for serial batch and the rest split between tests of parallel codes and dedicated server

tasks but the nodes were dynamically reconfigurable.

User access paths were expected to be either Ethernet-based logins or X traffic. ISS would be used to balance the load (see next session). The SP2 was connected to CERN's network backbone by FDDI (for AFS traffic among others) and to CORE data services by Ultranet.

Initially the system software would be AIX 3.2.5 with extensions to fit into the CN Workgroup server model as well as various IBM tools for system monitoring. The batch queuing system would be Loadleveler. All the nodes would be declared as AFS clients.

It was expected that much time would be spent at the beginning on tuning, especially memory paramaters. Delivery of the hardware was starting during October and the first pilot users were expected to begin using the system in November or December. A public service was scheduled to be opened early in 1995, building gradually to take over the CERNVM load.

11.1 The SP2 as a PLUS Server - T.Cass

PLUS - a Public Login UNIX Server - was modelled on the Work Group scheme described by the speaker at the HEPiX meeting in LBL; this instance on the SP2 would involve dedicating 16 nodes of the SP2 for interactive use by those users not assigned to a particular Work Group Server for one of the large experiments. Each of these nodes would be an AFS client having full access to AFS home directories.

Also installed would be the recommended HEPiX login scripts and the set of migration tools proposed by the UMTF team, most being available via AFS from the ASIS server. There will be support for users coming in from ASCII terminals but inevitably access from X devices would be more comfortable.

The SP2 would constitute a single service with respect to user registration. The speaker made a plea for a standard FORTRAN interface, perhaps via a front-end command name, to hide architectural differences and to select the most optimum set of parameters for a given platform, but there seemed to be little support in the audience for the suggestion.

11.2 Interactive Session Support - T.Cass

ISS was a part of IBM's Loadleveler product and was at heart a method of switching IP addresses to connect users to the least loaded node of a cluster. One associated a name and an IP value with a cluster or group of nodes to be used as hosts for interactive sessions; when users requested to connect to this name/IP address, a routine on the ISS server performed the name resolution to a physical name/IP address depending on the chosen algorithm. This could be as simple as round-robin or it could involve asking the nodes in the cluster or group via an rsh command for some measure of free resources and selecting the most lightly loaded. The measure of free resources was also very flexible, at the local implementer's choice.

The software had been installed and was in use for the SP2 and work was in progress to begin this is a general service but there were some limitations. It was also noted that ISS is usable for telnet, XDM and Chooser connection requests.

12 System Monitoring

12.1 GeNUAdmin, a System Management Tool - W.Friebel

This tool, written in PERL, was developed by a company in Germany and distributed as share-ware (free but users were expected to send a contribution to the authors). It was described fully at the LISA 94 conference. It offers management of UNIX systems from a central point using standard UNIX tools, with no kernel mods but with no graphical interface (yet). Also the documentation was best described as rudimentary.

Node configurations were stored in a database and these could be prepared by the tool itself or by hand. The tools contained consistency checks on this data and included the delegation of certain manager functions to different userids. It maintained files on the target clients up-to-date and in a standard form, files such as rhosts, services, and so on. It also checked the validity of soft links.

Installation at DESY Zeuthen was easy, including a simple port to their Convex. Zeuthen had requested several improvements which were now appearing in the latest releases of the tool. However, it was not yet in production use as this would imply major changes in the /etc directory of target clients and the dangers of this change would need to be assessed. Nevertheless, it was then expected to be activated before the end of 1994.

12.2 Roundtable Discussion

This discussion was led by Dave Underhill, Operations Manager of the CERN Computer Centre. His team has to deal with some 155 consoles in the Centre itself plus 30 in the operators control room; most of the non-prime time there is a single operator; and there is a wide range of monitoring tools for the different services. Finally, as if these were not enough, the operators have to respond to user interrupts for help.

Their plan was to move to a commercial tool - SPECTRUM - for network monitoring with the addition of MAESTROVISION for some system management functions. This would gradually replace a CERN-written alarm server. This last tool - SURE - was also in pre-production use at RAL and under test at DESY, who have also tested MAESTROVISION, which they report as being heavy on resources.

It was pointed out that alarm checking needed to check on the presence of certain processes or daemons, not simply if a node was alive or not. Should such checks extend to systems outside a Centre? Should the alarm system warn the operators or be linked to an automatic calling mechanism?

SLAC used another commercial network monitoring tool (TTS6000) and reported that this was also heavy on resources. They are also considering a simpler home-written tool based on the finger command.

Another popular tool in this area is Digital's Console Manager which is in use at DESY and CERN. Its main advantage is to reduce the need for systems to have their own local console. SLAC are evaluating an equivalent tool obtained from FNAL and which is in the public domain.

13 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.

In the discussion following the presentation, some ideas were exchanged about possible ways to merge the FNAL scheme with CERN'S ASIS scheme and the speaker will continue these discussions with P.Defert.

14 Problem Reporting Scheme - A.Lovell

The UNIX Workstation Support team at CERN have been investigating schemes for problem reporting and tracking and the speaker presented some results of their studies with respect to the aims of such a scheme and the requirements.

Problems can arrive via mail or from users in person or by telephone. After being assessed, they must be dealt with or passed to an appropriate expert; those which are not resolved immediately must be tracked to be sure an answer is provided in due course. It was agreed that problems submitted in written form, via e-mail, were easiest to deal with but providing input forms, TK/TCL or WWW, was essential both for front-desk support personnel and for users who might then be encouraged to submit more problems electronically.

The CERN investigations had concentrated on two tools, one home-written for use by the CERN network team and the other the public-domain tool gnats. Their respective strong and weak points were described; a final choice was due to be made shortly.

In the discussion, it was noted that form-based input must be kept simple and that the database of problems should searchable (but one person suggested that the submitter field should be hidden for search). One facility missing today in both tools looked at was the ability for a central dispatcher to allocate a problem to a specific support person. For a scheme which might be used to report problems from offsite (for reporting CERN Program Library problems for example), ASCII input must be possible.

From other labs, FNAL used a home-written tool based on ORACLE and they were looking at the REMEDY commercial package; SLAC made some use of REMEDY although one physics group used gnats; RAL had a home-written tool for internal use but were also looking at both REMEDY and a package from Legent.

15 Tapes

15.1 CERN's Tape Stager - J.P.Baud

The new version of the Stager supported staging files from tape to disc or vice versa, thereby overcoming a shortcoming of the previous version. Among its other features were

- one stager for even the larger experiments
- the stager was able to manage several pools
- it was fully distributed
- it permitted staging from outside CORE
- it ran as a demon with root privilege.

The speaker explained the internal execution flow and the various user commands. He also described the new concept of reserving space to avoid the problem in older versions of disc overflows. Finally he showed some impressive performance figures.

15.2 Possible Future Tape Strategies at CERN - C.Curran

Over the years there had been considerable improvements in CPU and disc performance but this had not been matched by equivalent improvements in magnetic tape performance. Today, the speaker had calculated, CERN operated some 122 tape units with a total of 9 varieties in its Computer Centre, differentiated by type and mode of use. CERN, like other HEP labs, was eagerly awaiting IBM's promised NTP but many open questions remained around this device.

CERN's tape library was some 750K volumes, of which over 400K were "archived", with a total of about 160TB of data, grwoing at 10TB per year.

There was a serious push for the use of robotics in order to reduce the number of manual tape mounts and various robotic devices which might be able to handle the quantities of data projected for the LHC experiments (1GBps) were presented. For LHC experiments, he estimated it would require 5 robots with next-generation tape devices or 200 Exabyte robots or 100 DLT robots or 20 3490E tape drives.

15.3 Status Report on XSTAGE at IN2P3 - S.Ohlsson/CCIN2P3

The speaker showed their xstage model with its control and data flows. It used 4 tape servers, 2 tape drives on each connected via System/370 channel interfaces and 2 disc servers for a total 200GB pool. The maximum throughput achieved was 2MB/sec, even with 2 channels active and using RFIO.

Future developments included work on administrative commands such as drain, clean, etc, as well as efforts to improve mount performance.