HEPiX-HEPNT NIKHEF Report
19-23 May, 2003

Alan Silverman
23rd May 2003

Introduction

As usual it was a well-attended European HEPiX meeting with 64 registrations from sites in Europe and the US, including some sites in Germany, Italy and Switzerland which do not often attend. The meeting was scheduled over 3 days and followed by one and a half days of hands-on tutorials given by some of the CERN EDG WP4 specialists on the installation, configuration and monitoring modules being built for EDG WP4 as well as a couple of Fermilab people speaking on the monitoring tool ROCKS. As usual most of the overheads are available at the web site (see under the link Programme).

The organisation of the meeting was efficient and there were no visible problems. The weather however left much to be desired but at least there was no temptation to skip sessions!

Highlights

- Lots of sessions on tests and evaluation of different storage solutions, some reporting results, others ongoing and promising results at the next meeting
- CASPUR and DESY Zeuthen are moving to Kerberos 5 with AFS using the Heimdal version.
- Like CERN, SLAC computer centre is running out of power and is building a new sub-station
- Most of the few remaining HEP sites still running Transarc's AFS are moving or thinking of moving to the OpenAFS version
- Interestingly, Fermilab is planning to migrate their NT-based direct-attached IMAP application to a SUN/NAS solution.
- Regarding Windows migration away from 95/NT, Fermilab and SLAC appear to have made significant progress, DESY less so but they have begun
- TRIUMF appears to be the first HEP site fully endorsing blades for bulk processing, other sites remain unconvinced at this time of the relative price/performance although a few (Fermilab and INFN Pisa) are investigating them
- DESY and SLAC are considering using a content management scheme as part of their web site remodelling
- Spam Assassin is used in a number of sites, DESY Zeuthen in particular has performed significant tuning
- Objectivity and HPSS are alive and well in BaBar and HPSS elsewhere (e.g. BNL). The monitoring tool Ganglia appears more and more often.
- There was a tantalising presentation on AMD's forthcoming 64 bit chip (Opteron) but most of the presentation consisted of AMD marketing slides and only a few in-house tests have been done at this time But the early results do look promising with regard to running existing 32 bit codes with no penalty and even seeing some speed improvements. They will now see what happens when such codes are recompiled for 64 bits. *A suivre*

- There was a discussion on setting up some mechanism to discuss security issues among the HEPiX sites, possibly via a mail list of site security officers, possibly by occasional face-to-face meetings. Some proposals will be discussed by the HEPiX board and then published
- Next meeting will be in TRIUMF, Vancouver, during the week of October 20th. After discussions inside and outside the sessions, we are thinking of proposing Grid Security as the theme for the Large System and Grid special day(s), perhaps inviting some high-level DoE representative
- The Large System sessions at the end of the week were better attended than we had expected, all 30 seats for the hands-on terminals were taken and there were some people sitting at the back of the room in addition. A feedback sheet was handed out and I will circulate a summary in due course.

Site Reports

CASPUR

Moved to a new location and expanded to 45 FTE. Added 3 more frames to their IBM SP2 on which they still use Transarc's AFS and SUN Grid Engine. They have a new 8 node quad CPU HP Alpha SMP but they have problems with Transarc's AFS under the latest version of Trucluster. Currently installing 5 dual CPU Itanium-2 systems on which they are collaborating with Intel. Finally, they have just added an NEC vector system. Their storage cells (4 on site, 1 off-site) been migrated to OpenAFS without any problems. They plan to migrate to Kerberos 5 after successful tests of the MIT and HEIMDAL versions and they expect to collaborate on this with DESY. They are still supporting ASIS repositories for those architectures no longer supported by CERN.

SLAC

BaBar has now stored over 1PB in HPSS. SLAC has been awarded a cosmology institute. The central compute farm now has 850 SUNs, 512 VA Linux dual-CPU PCs and most recently 512 Rackable Inc. dual-CPU PCs at 1.4 GHz. Looking for this year at Pentium IVs or high performance Xeons. With 200TB of storage deployed, looking at adding 100TB this year, expect to pay about 0.75 cents per MB. They have started a storage evaluation project to support both Objectivity and BaBar ROOT databases. Thus they need to run ROOT demons and some NAS systems only support NFS demons so not all recognised suppliers qualify although some claim to run in direct-attached mode as is required. They will report on next HEPiX on the results. The test setup for this acquisition is 3 SUNs hosts and 50 Linux clients. Manageability is an important criterion.

HPSS is running smoothly, storing up to 3TB per day with an equivalent read rate; there are two instances, one dedicated to BaBar, the other shared by other experiments. Currently running on Redhat 7.2 and 7.3; expected to go to Redhat 8 but BaBar is looking at Redhat 9 with a target to migrate over the summer. Almost all Solaris is version 8 except that all new servers get version 9. Running latest Transarc version on their AFS servers, all of which are SUNs, now running Solaris 9. Running LSF 5.0 and looking at 5.1 but needed to upgrade the scheduler to a quad node because of load generated by this new release.

Almost no more power capacity so they need to build a new sub-station and turn off old systems in the meantime since the new power supply is at least 9 months away And they need to replace the seismic capacity of the computer room floor!

Fermilab
AFS: beginning a migration to OpenAFS via a third-party company (Sine Nomine) for SUN (all servers are SUNs) and SGI platforms; the cost is $43K for 24x7 support and allows for some minimal code changes. All AFS and core server backup done by the TIBS product from Teradactyl; it is an incremental backup scheme and reduces full AFS backup significantly.

Found a strange 9940A tape drive problem connected with appending files – see overheads for the full (non-trivial) explanation. Storage Tek are working on a firmware fix. With respect to the FNAL Kerberos infrastructure, the first applications are appearing which make use of this and they are looking to collaborate with other sites on this.

A Windows Policy Committee has been created to monitor their Windows domain. Over 5000 users and 2500 computers, including most desktops, are now migrated to Windows 2000. They are currently looking at VPN solutions, MAC support via Thursby software and at Windows Terminal Server.

Migrating their PMDF Mail gateways to iPlanet and using an Alteon layer 4 switch for load balancing between their mail servers and investigating anti-SPAM software. Interestingly, they are investigating their NT-based direct-attached IMAP application to a SUN/NAS solution.

On their compute clusters, they currently see IDE disc problems on their Western Digital discs. Later in the meeting was noted that at least three HEPiX sites suffered from this without being aware of each other's problems.

The new Fermilab computer service management appears to be investigating more buy-in of products and services than before and are interested in experiences of other labs.

GSI
First site report from GSI. They run a 200 node Linux farm, distinguished by the fact that it runs Debian Linux. I also remarked the continuing presence of VMS systems, including on some desktops.

TRIUMF
Like CASPUR, TRIUMF have inaugurated a new computing facility permitting, among other things, upgrading of their network. They took the opportunity to perform a major legacy clean-out. They are using CUPS[1] to share printers between UNIX systems and, via Samba, with Windows. And they share their two large-volume printers with copy services and e-mail scanners. The Westgrid adjudication for a 1000 node cluster and associated storage was won by IBM with its blade systems. Installation should take place in the summer.

PSI
Another first site report, this one from the PSI in Villigen. The main platforms are Windows (Windows 2000 since December last year) and Redhat Linux (including a 56 Athlon CPU farm) Heavy use of OpenAFS and use of many public domain management and monitoring tools.

[1] CUPS – Common UNIX Printing Service

DESY
Major progress on the dCache project, now used by a growing number of experiments at DESY, CMS at San Diego and Fermilab experiments. It is Grid-enabled via the Storage Resource Manager and in fact DESY has joined the Grid world by being a founder member of the German D-Grid initiative. DESY has moved to OpenAFS with Kerberos 5 integrated – see separate talk. They have created a Web Office to coordinate the DESY web site and they are planning to use a content management system so that should be interesting to monitor. They are looking to create a Linux laptop project. Other DESY activities are covered in other talks

CERN
Helge Meinhard gave the CERN site report.

BNL/RHIC and ATLAS
Using HPSS to manage 4 tape silos, recording raw data at 350 MBps and believe they have enough capacity to go faster. RHIC so far has accumulated over 600 TB of data. The 140TB of disc storage is managed by 24 SUN E450s and they can aggregate 600 MBps to and from the farm. The central Linux farm has nearly 1100 dual CPU PCs running Redhat Linux 7.2 and 7.3. The software is described as GRID-like (Ganglia, Condor, GLOBUS, etc). They use Grid monitoring tools such as Ganglia and they are concerned with Grid user issues (remote site authentication for example). They have a well-defined security policy – see my CHEP trip report on the IT/DMM web site.

RAL
Added new hardware this spring – 80 dual processor 2.66GHz hyper-threaded Xeon systems; more are expected this summer when the budget is finalized. They wonder how much hyper-threading will or is being used. [Steve Timms of FNAL reported that they turn it off to ease system complexity.] Variety of Redhat systems in use but trying to standardize on version 7.3. There has been no serious problems with their disc farm since all 600 MAXTOR discs were replaced and they have recently added 11 more servers to the installed 26. Looking at performance monitoring with Ganglia and a new batch scheduler (maui) has replaced parts of PBS but not all of it works. Adding home-grown PBS[2] accounting. Proposal to implement a new helpdesk system based on Request Tracker to replace a home-grown e-mail scheme.

LAL
They have replaced their main interactive server by 10 dual-CPU Xeon systems and an HP Alpha server. Further expansion is severely affected by government budget cuts. Looking at authenticated SMTP to restrict mail relaying and Spam Assassin plus MimeDefang for SPAM filtering. They hope to deploy this first in client mode and later using Sieve at the server delivery level. The IN2P3-wide Windows domain forest is now in production across 9 labs so far. Three more labs propose to join, including LAL. For this they need to go back to NT and then re-upgrade the Active Directory. Implementing a VPN pilot based on the CERN model as presented at the last (FNAL, Oct 2002) HEPiX. Still no formal support for Linux on the desktop, they recommend VMware instead. They are developing a cross-site backup based on Legato with the nearby IPN lab.

[2] PBS – Public Batch Service

INFN
INFN sites use mainly PCs running Linux or Windows but the number of MACs grows and grows. Commercial UNIX systems are only used for dedicated services. Network security is implemented only by access lists on routers connected to WAN, no firewalls. On the desktops, Citrix and VMware are recommended for cross-platform access; some INFN sites are beginning to outsource desktop support.

A recent meeting of administrators concluded that INFN has between 6000 and 7000 Windows systems, 50% Windows 2000 but still 25% Windows 95. The domain structures are generally simple with one third of the sites having only stand-alone systems. Many sites have no central storage management, others use AFS or Samba, others have central backup. Generally little wide area access, little Windows Terminal Service and today almost no VPN. On the positive side, anti-virus software and Windows Update use are common.

CEA Dapnia
Since the last meeting, the threatened security measures were implemented. All offsite access was closed and had to be renegotiated site by site and then via firewalls and only to certain hosts. They are in the middle of migrating to OpenAFS. For new PCs, Windows XP is the default but no migration from W2000 is planned and Windows NT support has been stopped. Moving to Redhat Linux 8 and MacOS X. Like LAL, they are looking at an Exchange solution with the eventual aim to phase out the SMTP servers. They participate in many Grid projects, present and planned, European and French.

Normal Sessions

CERN Computer Centre Farms
Vlado Bahyl gave a short but efficient and confident summary of the current status of the farms and some of the EDG WP4 tools being used on them.

High Availability Central Services at DESY
Thomas Finnern described how they use a UNIX cluster (the f5 product from BIG IP) as a switch using network layer 2/3 routing to route a request for a service. It can be configured in several modes and these are variously used to load balance the DESY font servers, network install servers and the dCache servers and they are considering extending it to the public logon servers and they believe it could be applied to many other services in due course.

CERN's Solaris Services
Manuel Guijarro presented a summary of the Solaris services offered by IT/PS Group, including the present state and plans for the SUNDEV service.

Fermilab Storage Tests
In the first of a series of storage talks, the speaker described some tests performed at Fermilab on disc configurations from Spinnaker and Zambeel as well as on the in-house disc farm. There were many graphs and those interested should consult the slides on the web site. Using the test suite developed, he hoped to continue the tests on other systems, hopefully in collaboration with DESY.

CASPUR Storage Lab
These tests were conducted in collaboration with several other Italian organisations, with CERN, and with the sponsorship of a number of leading vendors. The devices tested included:

- SAN Valley IP-SAN Gateway which allows to connect two SANs over IP in a mode equivalent to having the devices connected to a single SAN.
- Nishan's multiprotocol IP Storage Switch which offers a similar functionality but with more advanced features although it is much more expensive than the SAN Valley kit.

The first study was to see how a SAN system could be used for distributed staging purposes. Several remote staging schemes were considered with concentration on seeing if Fibrechannel-connected tape drives could be used at native speeds across a WAN using SAN-over-WAN middleware. This test failed last year but a successful setup was achieved this year using each of the devices described above although disc performance was not as good as that from tapes in respect of loss of performance across the WAN. [More detail on the overheads.]

Next, they studied various NAS protocols in detail. Various kernel and cache settings were investigated and both write and read tests were performed. The results were presented for RFIO, NFS and AFS protocol. RFIO was the most performant, approaching raw disc speeds followed closely by NFS.

Lastly, they investigated IBM's GPFS and Sistina's GFS as possible base protocols for a scalable NFS server. In 2002 the results had not been convincing, the protocols being unstable and inconsistent. Do the newer versions perform better? The answer for GPFS is "slightly" but not enough to look really interesting at this time. GFS on the other hand has improved much more and throughput is nearly at native disc speeds in multi-stream mode (the tests were performed with up to 4 nodes exporting data).

Storage Hardware Survey at LAL
LAL has an installed SAN, a consolidated file server based on TruCluster and a Compaq Storage Array but they needed to acquire up to 2TB more. They looked at several options.

- Server-based IDE RAID systems are cheap but do not scale and there is no (cheap) high availability solution
- SAN-based IDE RAID is also cheap but again does not really scale (cheaply), is not generally available with redundant servers and has no TruCluster support
- SAN-based SCSI RAID would be easy to merge with already-installed systems, offers high availability and more scalability but is much more expensive
- SAN-based FC RAID has similar advantages as the previous solution, but also the same drawbacks, especially cost.

Finally they selected the FibreChannel solution and now needed to select a vendor:

- The first option was to negotiate a configuration from HP (thus guaranteeing TruCluster support) within their (limited) budget. But HP's solution had hidden costs such as the need for a dedicated (HP) control PC and a dedicated (HP) rack.
- Storage Tek's offer seemed more attractive but did not guarantee TruCluster support and the price-performance target could only be achieved with larger configurations
- Hitachi's offer was similar to the previous one but did promise TruCluster support and seemed to offer the best option so this was the one selected.

Their big concern now is the interaction between different vendors in a SAN environment (existing and new disc configurations).

Security

There was a discussion on what role if any HEPiX-HEPNT could or should play in security. Options include mailing lists for incidents, more HEPiX sessions and increasing contacts between security people. The first has been tried and failed but it was an open list as opposed one closed to all but declared security officers. How to get these people on to such a list? Should HEPCCC or HTASK get involved? In fact HTASK had setup such a mailing list but it too is virtually unused. Expand this to university sites which may be affected by incidents at major labs? Should these security officers meet on a regular basis, at least to establish a sense of community. LCG has created such lists for their Tier 1 sites; could it, should it, be expanded? Another alternative to join the list may be to have people named or moderated by their official site representative for HEPiX. And we should invite people on this list to meet at least once in conjunction with a HEPiX meeting. Some security people will discuss these issues offline (already started) and come back with a proposal.

Kerberos 5 at DESY

Wolfgang Friebel described a continuation of the work he started while at CERN. The Heimdal Kerberos implementation has integrated version 4 and 5 support and allows a good conversion of the version 4 database although some aspects of the implementation could be improved. The MIT version is more widely used, especially in the US, but other aspects are poorer, especially integration with AFS. The Windows 2000 implementation is impossible for this use, for one thing because it is incompatible with AFS database servers. He has decided to concentrate on the Heimdal version and he showed the steps he took to make it work with AFS.

He has setup a working service at DESY Zeuthen with three servers and he has an openssh client, and IMAP mail clients (e.g. pine) working with it. There are a few disadvantages such as the loss of load balancing of the AFS kaservers, no lock-out for too many failed password trials (but an alternative lock-out does exist), no cross-realm trust so no easy sharing access between UNIX and Windows domains. The next steps planned are to produce a native Kerberos 5 arc server and client (for batch use) from Rainer Többicke's original and tighter integration with Windows 2003 server. For these he hopes to work with CASPUR later this year. His summary is that Kerberos 5 deployment is easy, service integration is more tedious and users did not notice the change. But in the current implementation, cross-realm authentication is not available at this time and thus integration with Windows more difficult although not impossible. Matt Crawford from FNAL reported that the Kerberos 5 arc conversion had been done in the US – see Google.

Automating FNAL Data Centre Support

The project is known as Auto Assign and Notify (AAN); it uses Remedy to move towards lights-out operation. Remedy is linked at one end to NGOP to discover when there is a problem and at the output side to a pager/voice system, a mail agent and something called Airwarrior (!) which activates a cell phone when needed. The decision or not to open a new ticket is taken by an NGOP module. The workflow inside Remedy was programmed by a consultant via a 14 day contract. Since implementation at the end of 2001, NGOP has created some 1000 tickets out of a total of 11,000 in that time. Deployment of AAN started late in 2002 but a human was then still needed to confirm the decision to page someone. They have declared some 474 conditions in the scheme and gradually certain conditions (26) are being verified automatically, removing the need for human checking to release those tickets to the notification modules. Others (302) are declared as "page someone next working day". Remaining issues include detecting when a problem triggers

multiple tickets and how to transfer tickets between support groups as tickets are better understood and debugged.

The Grid at RCF[3]/BNL

An update on his CHEP 2003 talk. The installation was described in his earlier site report. The RCF is used to process RHIC data and to offer an ATLAS Tier 1 centre. The Grid-like facilities comprise the use of Ganglia for job scheduling and monitoring and Condor. Ganglia supports federations of clusters and can be used as a job scheduler in a Grid environment although they have not yet used it in that mode yet. Thus far via its web interface they are testing it, particularly its scalability, for running RHIC and ATLAS jobs. They have found however that customising the code, although open source, is not at all easy. More tests will take place as it gets rolled out to more systems.

Condor, actually Condor-G interfaced to Globus, is being readied to offer batch access to the Linux farm in a Grid-like way. They would like to investigate how this setup scales, what are its security issues and can it successfully wok with their HPSS system.

EDG Security and Site Authentication

David Groep of NIKHEF and EDG WP4 described some aspects of grid authentication and authorisation. He described how the concept of a virtual organisation is used in the 1.x and 2.0 releases of EDG to grant authorised access to a Grid site, the so-called VOMS or Virtual Organisation Management Services. He described how grid user authorisation is a two phase process – first user authorisation via LCAS (Local Centre Authorisation Service) and then, for successful authorisations, the mapping of credentials for that user (LCMAPS).

Grid Security for LCG-1

Dave Kelsey reported on discussions in LCG on security issues and how they might affect Tier 1 sites, almost all of which are represented in HEPiX and indeed present in this meeting (the University of Tokyo is the only exception). In his introduction to the LCG, he listed some of the committees and their hierarchy in respect to security and how the LCG Security Group was created.

For LCG-1 in July, time constraints force the project to use what already exists, largely from EDG, supplemented by a number of tools and mechanisms from various US grid projects. In the longer term, there may be some development required. There are still a large number of open questions on security policies and procedures where some decisions are needed and the LCG Security Group are discussing these and will make proposals to the Grid Deployment Board (GDB). The bulk of the presentation covered these issues in some details, what are the open questions and what are some of the options for each of these.

EDG WP4 Implementation on Solaris

Manuel Guijarro described how the philosophy and APIs of EDG WP4 are being implemented for Solaris, work partially sponsored by SUN.

[3] RCF – RHIC Computing Farm

EDG Testbeds and LCG-1 Planning
Markus Schulz presented the history and status of the EDG testbeds and the current planning for the imminent LCG-1 release, an update of his recent C5 talk.

Grid Computing at JLab
The Grids in question are PPDG and the Lattice DataGrid. As reported at previous HEPiX meetings, JLab uses Jasmine to control their mass storage and they are extending Jasmine to work with Grid technology. Storage Resource Manager is being added in collaboration with Fermi and LBNL to link to PPDG to coordinate and schedule file movement. Each site has its own implementation but tests have shown that they work together. It is hoped that a wider collaboration, including EDG, will work on version 2, concentrating on web aspects. Jasmine SRM has been deployed for experimental physics groups and another SRM implementation for the lattice gauge group (LQCD). The hope is that Jasmine will evolve to become a framework to build grid services upon. But in all this work they have discovered that running a grid is "real work", there are many issues to cover (security, firewalls, certificates, etc).

DESY Windows Project Progress Report
The official project start to move away from NT was taken in March 2002 but has been overtaken by Windows XP becoming the preferred client platform and it appears that some Active Directory functions seem to be incompatible between this and W2000. These seem to be solved in the latest versions and now the decision is to move to XP SP1 on the clients and move the servers to W2003, which has many attractive new features such as a Group Policy Management Console, Volume Shadow Copy Services and so on.

The server hardware will be built on HP (Compaq) Proliant servers, distributed between the Hamburg and Zeuthen sites. All have an integrated "lights-out" board permitting remote control. The DFS file base is on a StorageWorks array controlled by a Proliant cluster. They soon expect to add an HP Blade server with more StorageWorks disc space, 6TB in Hamburg and 1TB in Zeuthen.

The plan now is, having installed the 3 domain controllers, to make the setup stable and well-understood, to test the applications, to build the production Active Directory and then try to have some 500 client users working successfully in the domain (target 2004) and then, if all goes well, to migrate the remaining users into the domain.

Backup is done using Tivoli file backup and there is McAfee virus scanning and AutoUpdate Architect. Initially mail and printing will use the old (NT) domain (implying a second login). Later they will look at Exchange 2003, e-policies and Samba version 3.

Windows Progress at SLAC
Moving from NT to a single Windows 2000 forest and domain with multiple Domain Controllers (DCs). They looked at various modes to migrate, finally chosing an in-place upgrade of the domain controllers to native mode Active Directory. Bypassed W2000 on the desktop, moving directly to XP. XP desktops and portables are built from a boot CD, one version of which can then install via the network, the second from the CD itself.

They use MSI and Group Policies for delivering software. The anti-virus tool is Inoculate. Hot fixes are packaged and delivered monthly (compulsory) although users can do it by hand at any time and security patches can be delivered at any time. They share central print servers with some UNIX systems. The server hardware is Dell PCs. They make use of Citrix ICA clients.

After investigating many monitoring tools from Microsoft and others (full details available on request), they bought NetIQ's AppManager and Administration Suite. Backup today is Veritas Backup Exec but it is not flexible or powerful enough and they are looking a SAN-based architecture and the various tools currently available for this. Citrix will be upgraded to the latest version. The Dell SAN installation has suffered two major outages in 2001 totalling 6 days and one in March 2003 of 28 hours so other options had to be considered. The outage problems are thought to be caused by a mismatch between signatures written to the discs by the NT hosts of the array and those used by the production W2000 servers. They decided to move to a multi-tier scheme for storage – one with a 4 hour recovery SLA and the second using low-cost storage with a much longer recovery time, up to days. In conjunction with this, quotas are being introduced. After reviewing different vendor solutions for storage, they purchased one from Hitachi and have recently completed the data migration. Details of the actual survey can be got from the speaker. They purchased Veritas's StorageCentral SRM Tools for service and end-user usage reporting although it is not yet available for the users.

They expect soon to migrate to Exchange 2000 when additional storage is acquired, although it is partly available now, to system administrators. Next projects include implementing common user authentication (single username and password but not necessarily single login) between Windows and UNIX (although they don't yet know how they will chose to do this); creating an Extra Private Network using firewalls to protect their business applications (Oracle and Peoplesoft) and perhaps some old experimental systems which cannot upgrade from NT; a new Backup architecture and a web content management scheme.

SPAM and Virus Filtering at DESY

Like everywhere else SPAM mail threatens to overwhelm real mails. Apart from user annoyance and wasted resources, a small but significant number of these contain viruses which must be blocked or quarantined. In this situation, the different DESY sites have chosen their own solutions. For virus-checking the Hamburg site has chosen the Mimesweeper tool using rhe F-Prot Scanner; Zeuthen used McAfee embedded in the open source tool amavisd. The Hamburg solution works well and is in production, the load distributed over 3 machines. At Zeuthen, the additional load generated by their scheme would overload the mail server and anyway all Windows PCs have virus scanning. They may try again when they can add more powerful mail servers.

For SPAMs, mails are not filtered but those arriving at the Hamburg site larger than 250KB are tagged using Spam Assassin and Mimesweeper provides additional checks; in Zeuthen all mails are tagged using Spam Assassin. In Hamburg, tagged mail has the string "[SPAM]" added to the subject line allowing easy user filtering (although wrongly tagged mail may get lost or misrouted). Zeuthen adds an extra header line complete with the SPAM score given by Spam Assassin. Hamburg mail servers also modify the mail contents of tagged mail (!) which can cause confusion for mails forwarded on.

Inside Spam Assassin, they perform checks against all languages and set a trigger score of 5 to declare a mail SPAM or not. The rate of false positives has improved recently although less so in Hamburg where there is less tuning of the checks used. Zeuthen claims a recent SPAM recognition rate of 95% and a false positive rate of less than 1 in 10,000 mails.

They are now looking at ways of filtering SPAM mails and have decided to leave this to the user agent, storing the SPAM mails in a special folder with published recipes for users on how to apply filters within the various mail agents used on site.

CERN's Print Service
Sam Lown presented the work he has done to upgrade the service, a talk he gave earlier this year at C5.

CERN's Central CVS Service
Manuel Guijarro gave an update on the current state of our central CVS service and some future plans, including support for LCG who insist on a service not depending on AFS.

Remote BIOS Upgrades of SLAC's PCs
There is a big difference in what is offered by different vendors for upgrading their BIOS's, for example to fix bugs. The talk reports not a scientific study but rather a report on what was found to work in various setups. One method is "sneaker net" but this is manual and takes a lot of time per node. Alternatively, you can set up a DOS partition on the system, make it writeable, boot from it and use a script to write the BIOS. He tried a remote method using the DOS BIOS floppy image and booting via PXE but he failed to flash the BIOS. However, adding networking to the DOS image – very hard but has been done by Sam Flory of Rackable Inc - then permits the flash operation to work and indeed this was used by SLAC to upgrade their 512 Rackable PCs in 2 hours but it failed for the VA Linux PCs.

IBM has a propriety scheme involving special software on a special card. DESY quoted a similar scheme in an earlier session using a special card (their so-called "lights-out" card). Dell has a similar scheme but only from Windows on PCs with certain motherboards. Other motherboards offer possibilities to use this feature from Linux and Intel is in fact planning some motherboard upgrades which will permit online updates easily.

Finally there is the openBIOS project – which loads a special kernel module and copies this into the BIOS but today it only supports a small number of PC systems, mostly older systems.

AMD Opteron Processor Early Access
Performed by INFN Pisa. Opteron is the 64 bit chip from AMD with built-in backwards compatibility with IA32 (cf. the Intel 64 bit Itanium chip family which has a 32 bit compatibility mode and which interprets 32 bit codes – I know Sverre will correct me). Thus with AMD, existing codes should work without recompilation and using no emulation. Of course optimising the code for 64 bit operation should improve its performance. Various vendors, e.g. SuSE already provide 64 bit Linux for these chips.

They developed a testbed in collaboration with AMD and Ferrari (AMD is a Ferrari sponsor) with the aim of seeing how they can move to this technology from today's 32 bit PCs. How much of the HEP environment is really re-usable? They have installed a 2 CPU system with 16GB RAM and they have tested Linux in 32 and 64 bits and Windows 32 bit. Ferrari has a similar setup in Marinello. Using SeSE 64 bit and standard CERN Redhat 7.3.1, both installed immediately without change. Using a Ferrari benchmark 32 bit code, they saw better 10-25% performance was seen over a 32 bit CPU, as promised by AMD. They are now porting their own software to 64 bit to see if there is a real performance improvement. On their initial experiences they have pre-ordered a blade server capable of accepting 160 CPU Opteron processors per rack.

Linux Server Vendor Qualification at Fermilab
The latest qualification is not yet complete so the final result is not yet public. They focused on 1U servers, 2.4GHz or better, with 1GB RAM. The qualified vendors will then be invited to submit offers and the top 5

will be put on the recommended vendors list. Future hardware purchases strategy dictates that purchases should include complete systems including racking, installation and warranty. The last qualifications were in 2001 for 2U servers and for desktops. Since the first desktop/server qualification in 1999 they have seen a price/performance factor improvement of 6 whereas Moore's Law would have predicted only 4.

The evaluation covers multiple nodes to get a more realistic scenario and small business cannot always provide what is needed in the time needed. References were required. To avoid being caught by firms going out of business, they will track the performance of all qualified vendors and may add more to their recommended list if needed. The eventual purchases will be long after qualification so account must be taken of component change as they have been hit by this before.

Vendors have to have Linux experience, to offer quality components and systems and to have good support. 45 vendors were contacted initially, 29 attended the information meeting and 21 passed acceptability criteria for configurations and benchmarks and supplied units for evaluation, of which 18 have passed qualification (see the full list on the overheads). These vendors are currently preparing formal bids and the 5 best price/performance bidders will qualify to be included on the recommended vendors list.

The evaluation covered technical issues such as power supply, ease of change of component swaps, code performance, etc. Heating effects were especially looked at and all systems had to perform successfully a 30 day acceptance test. They found AMD units ran very hot and had a higher failure rate than Intel systems. The rest of the talk covered the tests in great detail, see the overheads. He closed with a few tips for other labs wishing to perform such benchmarks (including limiting the number of vendors to be compared!) and he is now steeling himself to qualifying blade server suppliers.

Various members of the audience sympathised with the dangers of the 6 month delay between qualification and purchase, a problem not unknown at CERN so it is nice to see that we are not alone in suffering from this. There was also a (heated) discussion on whether blades are ready for production deployment or not; TRIUMF thinks yes, RAL and BNL are less sure, believing that 1U systems still today offer better price/performance.

Linux Servers Experience at JLab

Jasmine is JLab's mass storage system; this talk covers the data movers and cache servers. The data movers are Redhat 7.3 PCs, dual 2.2GHz CPUs. Disc and tape tests were performed to check advertised throughput figures as well as RAID performance. It was discovered that the chosen RAID (LSI Logic MegaRAID) is only supported if standard Redhat is used, without XFS – which in fact is preferred. And the RAID monitoring program caused SCSI resets! Now looking at different RAID controllers.

On tests on the cache servers, 2.0 GHz Xeon systems with 3ware IDE/ATA RAID controller, they had a number of problems with the Western Digital discs until they upgraded their firmware from that shipped by the vendor – a problem which several other sites noted they had been hit with and had solved – how to share this information better?

ROCKS at US/CMS

During this talk, to prove the power of ROCKS, the speaker started a re-installation of a node at Fermilab and we could watch its progress on a window on the screen! The intention of ROCKS is to remotely manage nodes with minimal human interaction. He described how the CMS cluster at Fermilab has grown in terms of various hardware units, including the history of Redhat versions and how the two interacted. By

the time LCG-1 is due, they expect to have some 200 nodes split into 6 sub-clusters with a variety of middleware and software packages.

All this led to a search for some tool which could deal with this complex setup in a more autonomous mode, for example ROCKS from NPACI (fully described in the first Large Cluster Workshop in Fermilab in May 2001). Other tools were investigated, notably SystemImager and OSCAR, but ROCKS was judged superior.

ROCKS can be used as a complete cluster management suite although they don't use it in that mode. The key to it is – if in doubt about a system, re-install it. Nodes are 100% auto-installed by it. ROCKS can be used to create multiple distributions, customising configurations via RPMs and Kickstart files. Using these tools you define various node configurations offering different services. But ROCKS assumes a private network.

Fermilab confuses things by needing AFS and Kerberos (forcing use of ssh) and by the Fermilab public network so Fermilab had to extend ROCKS (open source of course) to handle these. Also they wanted dynamic disc partitioning.

Wrapup

It was confirmed that the next HEPiX meeting will be in TRIUMF, Vancouver during the week of October 20th. Full details will be announced via the HEPiX mailing list shortly. After various discussions in this meeting about Grid Security and ways to create some forum for this, and discussions offline, I have proposed to use Grid Security as the theme for at least the Large System sessions at that meeting. With Ian Bird we will investigate what form this might take but some options include:

- Inviting a high-level DoE official (Irwin Gaines?) to participate
- Closed session, invitation only, for recognised site security officers, Tier 0, 1 and 2 sites, possibly others; discuss grid issues with the goal of solving some of them? Or setting policies?
- Open sessions or presentations, possibly within the normal HEPiX-HEPNT sessions on security in general, grid security in particular.

These of course are not exclusive and we will consider all of them and more over the coming weeks.

The meeting after that, spring 2004, is likely to be organised by RAL but, happily for some of us, not in RAL itself but more likely in Edinburgh in May.

Finally, as the European co-chair of HEPiX changed last year, similarly the US co-chair, Lisa Giachetti of Fermilab stepped down this year and is being replaced by Chuck Boeheim of SLAC.

Large System Sessions

These filled the last day and a half and consisted largely of technical presentations of the installation, configuration and monitoring components of EDG WP4. They were given by German Cancio Melia, Piotr Poznanski and Sylvain Chapeland respectively and included hands-on exercises. There was also a technical description of how ROCKS works by Joe Kaiser and Steve Timms of Fermilab. There were 15 PCs for the exercises with 2 people per PC plus seats at the back of the room and in all there were some 35 participants, a certain number of whom had come only for these sessions. We are running a feedback survey to see how popular such sessions are in this context but already it is encouraging to see the interest shown in this first run, in fact an updated re-run of an EDG WP4 workshop held in CERN last December.