HEPiX FNAL, October 23-25, 2002

Introduction

The Autumn 2002 meeting took place at Fermilab and followed on from the second Large Cluster Workshop (which will be the subject of a separate summary). Some 60 people attended (is this a record attendance for a US meeting, representing some 18 sites, more than half European. The three days will pretty full and thus so is this report. Highlights for me include (in the order they appear in this report) :-

- Several sites again report problems trying to run large NFS systems, the bigger the system, the more problems
- Most sites seem to be moving, or have moved, to Redhat 7.2. Is only CERN going to 7.3?
- Several sites happily using, and often expanding, their HPSS setups – BNL, IN2P3, LBNL and of course SLAC
- Similarly, Objectivity is still being used in BaBar sites with no statement about moving away, some "concern" over its future was admitted by one speaker from Lyon
- LBNL has decommissioned its last Cray, but adding more than 3000 processors to its IBM farm
- An LBNL group is evaluating various batch schedulers and should issue its findings soon
- LAL has implemented CERN's Print Wizard; but JLab finally wrote their own central Printing scheme despite an agreement to adopt the CERN model – more intellectually challenging I guess
- A monitoring package called NAGIOS has appeared at many sites
- IN2P3 have negotiated a national deal to use VMware; at SLAC, Fermilab and LAL, users who want both platforms are recommended to install Linux on the PC and connect to Windows Terminal Servers to get the Windows apps they need
- DESY's dCache (similar to CASTOR) is in production at both DESY and Fermilab
- For high performance NAS, LBNL have chosen Zambeel, DESY have chosen ExaStore
- DESY and SLAC have still not started to migrate their NT domains to Windows 2000 or XP although both state (again) that they are about to start
- A description of CERN's fight against SPAM (should be repeated at C5 and maybe elsewhere)
- It appears that no site represented offers a large central CVS service; all examples quoted were much less ambitious, intended for particular projects or teams
- Work on NGOP (Fermilab's equivalent to PEM) has slowed but not stopped; one reason given (but not confirmed) was that people had been re-deployed in an urgent effort to improve the beam luminosity which had been at RUN I levels and is only now improving
- Interesting session about a very high speed (up to 1.15 Gbps) lightspeed data transfer from TRIUMF to CERN, including back to back tests with a 12,000 Km loopback
- A few labs are starting to evaluate Blade systems
- After a lively discussion, it was agreed to dedicated typically one day during each HEPiX week to focus on topics of direct interest to LCG (and other Grid projects). Ian Bird and I will collaborate together with HEPiX to schedule this time
- There was an online survey of AFS at the various sites represented. I report this in some detail but there may be a fuler survey conducted later by Lisa Giacchetti and this may be a topic for the next meeting, which starts on May 19th in NIKHEF.

Site reports

These reports try to extract and summarise points from the various sites which have changed since the previous meeting in the Spring in Catania.

Yale

Still running a VMScluster to support printing and NT. Still using NFS and still having big problems with it and still does not know what to do about it. With 15-20 desktops and 20 more Linux servers coming in, she (Rochelle Lauer) believes she has serious management problems.

BNL

The RCF Linux farm (RHIC Computing Facility): should provide both general interactive and data analysis facilities as well as central data recording. Also houses the US ATLAS Tier 1 centre. Staffed by some 25 FTEs of whom 4 are dedicated to the RCF. 1200 TB capacity HPSS mass storage in 4 silos with connected 14 NFS data servers and a FC SAN; expect soon to expand the HPSS store. Also use AFS for personal files.

RCF is 840 1U and 2U systems, most with 1GB and up to 140GB of SCSI disc storage. The farm is logically split into 2 sub-farms for analysis (CAS) and for reconstruction (CRS). Running Redhat 7.2, images installed via Kickstart and customised via RPMs. The analysis farm is scheduled via LSF and permits logins on some nodes. No logins are permitted on the CRS farm and it uses a locally-written batch manager. Running a grid testbed for ATLAS.

RCF farm is surrounded by a firewall with only ssh through a gateway bastion. Monitoring via VACM and, more recently with xCAT from IBM (http://www.x-cat.org
) and CTS for reports. They built a web interface to their alarm scheme.

160 new IBM nodes coming in soon and should buy another 220 nodes when budget approved. Will upgrade to LSF 5 and looking at Condor.

IN2P3

Moving to Redhat 7.2 but have been asked for 6 different gcc compiler versions. Increased disc space, now up to 4 servers and 50TB disc space. Still doing small local developments on HPSS, to build APIs to RFIO from various languages. Also working on 64 bit support for RFIO. “BQS is in constant development” – support for parallel jobs, data bases for batch monitor control and integration to Globus.

Saclay

Biggest change was a new security policy imposed by the CEA: some ports were opened, some closed and filters are applied at the server level on all IP ports. No reasons known for the choices. However, their grid testbed is outside this protection. Will phase out NICE later this year and offer a choice of Windows 2000 or Redhat Linux. AFS cell consolidation in progress, second server, upgrading their OpenAFS version (on Linux) to 1.2.6 (or maybe .7) and adding new clients (MacOS X). A new French Grid is starting – e-Toile. Discussing upgrading Exchange to Exchange 2000.

LBNL/NERSC

Change in internal structure imposed by the DoE but effectively just a name change. Working at the Oakland site, finishing the machine floor and moving a wall 5 metres **inwards** to make way for more parking for a 40 story high-rise next door! Some concern on what physical effect this work will have, so installing seismic vibration measurement equipment to detect effects.

Added a Zambeel file server for home and spool space and 60 dual Athlons to the cluster. De-commissioned the Cray on October 1st. The IBM RS/6000 cluster being upgrade to AIX 5.1 and doubling the size by adding 52 frames, 3136 more processors. Working with IBM to get grid services running on this cluster (which is called Seaborg). Blue Planet was announced – NERSC, ANL, Oak Ridge and IBM working on an earth simulator.

Another group (ASG) are evaluating various batch schedulers including LSF 5.1, OpenPBS/Maui, SGE from SUN and MauiME: a report should be issued soon. They are also looking at various file servers.

The HPSS data store is being expanded with 20TB more cache and moving to 200GB tapes. They are starting to test Linux movers.

LAL

No major change since last meeting. Participated in IN2P3 national deals for VMware and Legato Networker. Moved to Redhat 7.2 and started a monitoring project based on Nagios; it seems to have a robust collection mechanism and permits development of a web view of the resources. Can use a database like MySQL to store its data. Will upgrade their IMAP server to Cyrus 2. On new desktops, they deploy Windows XP. They use the CERN Windows Print Wizard in production with only two minor issues (producing new drivers and using printers at both CERN and LAL – to support of multiple sites needs manpower to be assigned somewhere).

The IN2P3 Active Directory "forest" pilot (see report from the last meeting) has started and they hope to be at production level during the winter.

DESY

This reports covers the Hamburg and Zeuthen sites. Knut Woller said that the two sites are moving ever closer together (logically) although there are still distinct AFS cells and differences in computing architectures.

dCache is an ongoing collaborative development between DESY and FNAL to provide a distributed cache between clients and HSM. At DESY, it comprises 30TB disc pool on IDE RAID servers. Used by all major DESY and some FNAL groups and being looked at in other labs. Allows the use of cheap tape media by greatly reducing the number of mounts as well as the manpower needed to manage data management. Scales to thousands of clients and hundreds of servers; ROOT supports it as does GridFTP. It will be expanded to a 100TB pool early next year.

Many user groups request multi-TB high performance file system with random access. Tests have shown that AFS will not scale up to this level. Since users require direct access to files (not just staging), dCache does not help. Exanet have proposed their solution – ExaStore. It's a highly-scalable, high performance NAS (similar products exist from Zambeel and other vendors). Features of ExaStore can be found on the overheads. First tests, since April and running a beta system, have been quite successful although performance is not what it expected from the eventual production release. Target configuration should have 8 nodes and 12TB.

Reviewing their old user registry to take account of rise in the number of users, user groups and systems. It should be linked to the DESY human resources database. And it should have some degree of delegation. In design since January and evaluating commercial and public domain tools in parallel. The design is now thought to be complete and they have concluded that it needs a locally-written package. To be continued ….

The DESY Windows domain is still NT. Decided on an intermediary state based on this but with W2000 and XP on desktops, mostly the latter, especially on new PCs. These should still be installed via Netinstall, which now supports Windows 2000. Any new domain servers are W2000. A new project team is finalising the Active Domain design for the site and they hope to start migration early next year although NT must remain alive for the Controls Group.

Still in the "sad state" (quote by Knut Woller) of running 3 mail services - sendmail, Exchange, PMDF with load and capacity problems on all. Hope first to remove PMDF and eventually consolidate to one single service but no project yet.

RAL

Normal farm upgrade this year. RAL has been named to host a UK SuperComputer centre in a collaboration with Daresbury, Edinburgh EPCC and IBM. Consists of 40x32 pSeries CPUs based on the Power4 chip. But connected so as to look like 160x8 nodes. Over 1TB of memory. Target is to increase parallelism in the target applications.

Recent disc problems on their main cluster were blamed on a bad batch from the vendor who is now replacing the full batch. Felt the effect of needing to support users working in SLAC time as well as UK time. Have established a trial Redhat 7.2 service and hope to upgrade the whole cluster later this year. Will support Grid "generic accounts".

Will continue to run Objectivity as needed by BaBar users (RAL is a Tier A centre for BaBar). What was not said was that RAL also still run a VMS cluster, for a few recalcitrant users.

Exchange 2000 Pilot at CERN – Frederic Hemmer

No news for this audience.

Fighting SPAM at CERN – Emmanuel Ormancey

Currently, 25% of CERN's incoming mails are designated as SPAM. With the new tools, we expect to detect an additional 10% of mails, nearly reaching the average measured in a European survey of 36% of all mails being SPAM. US users report SPAM proportions up to double this.

Today's checks are at sendmail level and they check for mails from local lists of suspect or banned IP addresses or domains or for certain keywords in the subject or for header consistency. Evaluated some commercial products but found no real difference from what we do already and therefore not much better. Some don't delete mails but put them in quarantine and this requires a lot of manual overhead to filter. However one, SpamAssassin, incorporates several different tests and works in a client/server mode. Found good for SPAM detection but found some bugs and some particular problems on our Solaris mail servers (may be related to the version of Solaris we use). But judged that it would still be necessary to merge in our own tests (not so easy) and produced poor logging.

So we took the code of SpamAssassin, which is public domain originally, and added the SpamKiller code (in C#.NET) from Microsoft which is more programmable and more stable in operation. It is hoped that this will detect more mails than SpamAssassin alone.

Anti-SPAM tests found include regular expression checks on the headers but also on the body text, header consistency, blacklisted senders and some more complex tests. Each test results in a score and if a threshold

is crossed the mail is considered as SPAM and quarantined at the client level, delegating to the user what he or she wishes to do with the suspect mail, with the additional help that the "score" of the mail is given and the user can decide to accept or reject given levels. The client/server protocol is "HTTP-like", the server must run Windows since the code is in C#. Logs produced permit the system management to tune them over time.

The tool is integrated into the Exchange mail service today via SMTP and there are plans to integrate it also to the sendmail service. However, for users not using Exchange, it will not be possible to use the client quarantine feature and the user will only be able to delete the mail or send it to abuse@cern.ch.

LCG/EDG Grid Security – Dave Kelsey (RAL)

A status update since Catania and plans. In May they delivered to the EDG a full Security Requirements document and are now working on a security design for EDG middleware, due for delivery in January. They now believe they have identified all the needed components of security which must be catered for – see overhead.

Continuing to work to extend trust between Certificate Authorities (CAs) and individual site managers, extending this beyond the EDG to LCG and to other grid projects around the world. A minimum requirements document has been prepared to define what it needed in this respect and they are negotiating this with the GGF. There are already 13 trusted CAs with 4 more under active consideration.

The Virtual Organisation LDAP demonstrated in Catania has been extended into VOMS (Virtual Organisation Membership Service) to make user authorisation easier but there are more steps required, for example for ACL management.

LCG Phase 1 plans to deploy a production quality grid by the middle of 2003 and therefore requires a certain security level. The current plan is to create VO (Virtual Organisation) databases, perhaps maintained by experimental offices, and methods to allow sites to set up appropriate accounts on the fly or using dynamically-leased accounts. In the realm of CAs, both CERN and FNAL are proposing Kerberos-based CAs. But there is not yet agreement between EDG and LCG. And some sites are reluctant to accept long-lived private keys so some other scheme such as one-time passwords or smartcards may need to be considered. It is likely that a combination will be needed. In short, authorisation technology is today still immature and the experiments must participate to come to a workable solution in time for the first deployment without locking ourselves into the wrong choice too early. And there needs to be ongoing discussions between sites who until now have not seen the need for inter-site trust.

Computer Security Update – Bob Cowles (SLAC)

Usual update on the latest security incidents affecting Solaris, Linux, Cisco, Apache and Microsoft which might have affected HEP (indeed all) sites. And he came to his usual conclusions – keep up to date with patches, firewalls are not a substitute and poor system administration is still a major problem.

The FBI and the USENIX SANS organisation have put together a web site with the top 20 tips to protect your system – http://www.sans.org/top20/

Windows IP Security Filters – Joe Klemencic (FNAL)

IP Security filters are a new feature in Windows 2000 and beyond to authenticate and encrypt communications. They authenticate machine-based communications, not user-based, and then negotiate

encryption schemes to be used. This talk described in great detail the features and how they are used. But it was the last session of a rather full day so you will need to consult the overheads for details if you want them.

PDSF Host Database – Cary Whitney (NERSC)

A new database intended to track all information about nodes. NERSC use cfengine for system administration tasks and this feeds in the base information. Web interface written in ZOPE. But now being rewritten in PHP. There is also a command line interface so that scripts and cronjobs can update the database. There is a query/update tool. Information stored includes very low level items for operators to find switches, computers, consoles, etc and they plan to add photographs of physical modules.

CERN's New User Support Model – Maria Dimou

Maria described the new user support structure implemented recently with a refocused Helpdesk team and the Manager on Duty. She described how the team had studied past practice including reviewing many case histories as stored in Remedy tickets and how this had been used to define the new model.

She then summarised the results of a survey she had performed on how user support is organised at other HEP sites. The results are briefly summarised in her presentation.

Printing at JLab – Sandy Philpott

They have some 200-250 printers driven by 2 Windows and 1 UNIX print server with various incompatibilities between them and a number of long-term defects. Although they had looked at our CERN model (there was an agreed collaboration with JLab in 2001 and a JLAB visitor spent some months here with Ignacio and the print team to understand our model) they assigned a new team member to this task who basically designed and developed a new scheme.

Their new configuration is a Linux node running Samba to support SMB to gather Windows printing commands and LPRng for the UNIX connection. They built a new UNIX client with command line and graphical interfaces to send print jobs to the printer but they haven't solved the Windows client issues fully yet – still need to upload individual print drivers on demand.

Current status is that it is being used by a restricted number of users and should be made production soon. There is only one print server so no load balancing of fail-over yet.

Evolution of the CERN Printing System – Philippe Defert

Philippe reported on recent work by Ignacio and by Sam Lown (new student in PS Group) to

- Improve scalability
- Improve synchronisation of the printer configurations between all the servers
- Add better accounting and reporting

The key new feature is to create an SQL database of the printers and use this to feed, update and synchronise the printers. This gives only a single place in which printer configurations get defined or modified and then deployed to the different active print servers. Adding a new print server becomes very simple.

We will take the opportunity to upgrade the hardware, the Redhat version and the version of LPRng. The next steps will be to look at automatic fail-over of servers in case of problems, and load balancing.

CERN's Central CVS Server Status – Philippe Defert

Philippe presented on behalf of Manuel Guijarro the reasons why we created this new service, how we came to our chosen configuration (both as already presented to C5) and the status some 3 months after implementation (as presented to FOCUS).

When asked what other sites do, FNAL's system team use a stand-by server with an rsync of the 3GB file base every 15 minutes (does that scale? Is every 15 min enough?) but there are only a small number of users and experimental teams run their own CVS services. Even if rsync is replaced by CVSUP to mirror the file base, it still probably does not scale to our needs and any "fail-over" is far from automatic and leads to loss of service for some time and loss of any updates performed since the last rsync or CVSUP (although in general the user can subsequently re-commit any changes). No other site represented in the audience owned up to offering a central CVS service (IN2P3 offer a CVS service for DataGrid but only on a single server with a small repository).

NGOP Update – Marc Mengel (FNAL)

NGOP is Fermilab's system monitoring tool (cf PEM). Various improvements since the last meeting and more nodes are being monitored, including the mail, the web and the AFS servers and the Enstore cluster – now up to a total of 1100 hosts and some 25,000 components. However, performance monitoring is still on hold. There is a new GUI written in Java and a better web GUI.

But the main change is the suggestion to target NGOP to help achieve lights-out operation. This is causing a re-think as to whether and how NGOP might need to change to accommodate this. Also looking at a better admin GUI, using XML classes and deploying it at other labs (IN2P3 are looking at it and actively so during this week).

Windows 2000 Remote Installation Services – Michel Jouvin (LAL)

He presented work done to set up an RIS server to allow users to install their desktops and portables over the network using features provided by Windows 2000 or later. The talk was effectively a detailed technical description of the scheme and those interested should read the overheads.

CERN's Windows 2000 Migration Wrapup – Frederic Hemmer

A review of the migration as given to several audiences at CERN.

SLAC's Windows Migration – Bob Cowles (SLAC)

Presented on behalf of the SLAC Windows group who "could not themselves attend because they are in the middle of the migration." Plan is for Windows XP and Office XP on all desktops and other client systems (about 1600) with Exchange 2000 as the mail server by December 2003. A pilot migration of 5% of the user base, spread across all departments, will be migrated this November.

Meanwhile NT is being "stabilised" except for security fixes. Plan a simple Active Directory forest with multiple (about 4) domains; today there are about 12 NT domains. After discussions with Microsoft (and others, including CERN) they will perform in-place upgrades of the domain servers. The desktops themselves will be re-installed and not upgraded (as recommended in the CERN migration).

For inter-operability with the rest of the lab, they will install and support Samba and AFS. Fermilab report no problems with AFS integrated login but they do see problems with Samba under Windows. Like CERN, SLAC do not recommend or support dual-boot systems. If users really want both, they are recommended to install Linux and use Windows Terminal Server to get Windows. [This was echoed by both the Fermilab and LAL Windows managers.]

ATLAS Canada Lightpath – Corrie Kost (TRIUMF)

Created in Jan/Feb 2002. Initial goal was to transfer 1TB of ATLAS Monte Carlo data from TRIUMF to CERN across an end to end lightpath at speeds above 401 Mbps (the current land speed record). The path was opened on Sep 20th and 1TB transferred at 670 Mps on Sep 22nd, increasing to 750 Mps 2 days later. They used both bbftp and Tsunami, both in disk to disk mode. Peak rates of 1.15Gbps were seen in disk to memory transfers using Tsunami.

As a by-product of these tests, TRIUMF's computing infrastructure benefited from various upgrades, some sponsored by vendors wishing to participate. They also performed useful studies into 10Gb Ethernet technology and set a new benchmark for high performance disk to disk WAN transfer (see table in the overheads which also contain many details of the setup and testing and many "black magic" tricks discovered *en route*). During late setup they performed back to back testing, over a 12,000 Km loopback!

Lessons learned include

- Linux software RAID is faster than most conventional SCSI and IDE hardware RAID based systems
- For best performance use one controller for each drive, and the more disk spindles the better
- Unless programs are multi-threaded or kernel permits process locking, dual CPU will not give best performance
- Commands like tar and file compression take longer than the transfer and deletes take a *long* time.

In the longer term, the expertise, and some of the equipment, acquired, will be used to connect TRIUMF to the forthcoming Canadian WestGrid project.

Long list of acknowledgements including several members of CS Group, who provided help during Canadian west coast time.

Comparing Disk Benchmarks – Chris Brew (FNAL)

Still a work in progress. Required because they are continually acquiring more storage, how to compare the offers? Options seems to be Bonnie++, IOZone, Reader/Writer (produced locally within CDF), TIOBench. Tested on Linux, Solaris and IRIX so on which platforms can they all be built? The answer seems to be yes for all combinations. The overheads contain tables and charts of the main results of the comparisons.

His conclusions are

- Benchmarking tools show a great deal of variability even on simple tests, especially where memory caching may be a factor
- Bonnie++ and IOZone Read and Write speeds and roughly comparable in most cases
- TIOBench Write speeds are generally not comparable but Read speeds are
-

Are Blade Servers Ready for HEP – Rochelle Lauer (Yale)

Yale's physics department is only a small site so must make best use for the buck. Is Blade the answer for them? Do Blade systems really offer lower power and cooling needs, easier management, etc.? All the major vendors now propose them. She concentrated on the HP blades which support various Linux and Windows flavours, are reputed to be hot-pluggable and should be cost-comparable to an equivalent-power server. For a site with a small annual budget, a nice feature is that the configuration can be expanded in easy stages, taking advantage of technology improvements even within the same rack.

In reality the drawbacks include higher price for the processors compared to the equivalent power on desktops, why only 1.4GHz chips, why does it require management software (which only runs on a Windows box – is this only in the HP implementation?). Today, Blade systems seem to be designed and targeted for high-availability markets, how do they work in HEP environments? Are the necessary drivers available for Redhat Linux? More features of Blade systems that she discovered during her research, and questions she poses, are in the overheads.

Yale is likely to start with a small configuration and see how it goes. SLAC plan to evaluate Blades as part of their next acquisition cycle. FNAL had rejected them for the moment because of prices quoted to them but they had been quoted around $3K per blade while Rochelle quoted $1800 per blade from HP. LAL is looking also at Blades: their conclusions are that the first generation have too low powered CPUs but the next generation, due in the winter, should be more interesting.

Large Cluster SIG – Alan Silverman

I reported on two activities

- A very brief summary of the Large Cluster Workshop earlier in the week from summaries prepared by Ruth Pordes and John Gordon
- and a review of various site surveys conducted on operations methods of the different HEP computer centres, their videoconferencing facilities, how they select PC chips, etc. It was accepted that such surveys are useful and should continue on demand.

LCG and HEPiX – Ian Bird

Ian opened with a review of the project, its goals, timescales and the various interactions between areas within LCG and between LCG and outside bodies.

Deploying the grid middleware to remote centres is a multi-dimension problem. Many issues, some technical, some social. He exposed a few of these issues in more detail and what must be done to achieve the first planned deployment of LCG by the middle of next year. He then demonstrated where HEPiX could be a player in this, complete with a first, long, list of possible topics. His idea, as well as maintaining the link to the Large Cluster Workshops is to add a Grid Coordination/LCG interest group to work along with HEPiX with dedicated time at the meetings and sponsored or seeded talks. There was then an interactive discussion on how to integrate HEPiX into helping to solve some of these issues. I will report this in some detail because Ian and I need to take on board the opinions expressed in order to schedule further sessions; others may skip to the last paragraph in the section for the result of the discussions.

Dave Kelsey: tension between where to schedule sessions, HEPiX or this special day, restrict such special days to workshops and discussions and not hold formal sessions. Siroli: avoid parallel sessions. Ian: such sessions would become main-stream and assigned a dedicated day or half-day, not overlapping or parallel sessions. Ian and Alan – no more invitation-only days, part of HEPiX. Michel Jouvin – why create a separate SIG, use the Large Cluster SIG. Ian – ok but more frequently than every 18 months. Michel – needs careful scheduling to avoid overlap and not more than a week in total.

Alan – propose 1 day Large Cluster/LCG workshop, 1 day for Windows discussions if wanted (keep presentations to the main sessions in both cases), 3 days "normal" HEPiX. Timings could vary by agreement if one part of this needs less time at any given meeting. Lisa – meeting coordination is affected by multiple interests and who is able to attend particular meetings. But she agrees to something like our proposal. But how should "HEPiX" decide? In the meantime, she agrees that we should try the proposal.
Michel - schedule the special days to the end of the meeting, cover general issues first. John – if special sessions at the end, cannot report to the main sessions.

Bob Cowles – we will need to cooperate more between sites, how are we to decide on concrete questions affecting multiple sites. Ian – use HEPiX to discuss these issues but there must be smaller and more focused meetings between HEPiX meetings to resolve or work on the issues raised; many of these smaller groups already exist within grid projects, the HEPiX week should be used to bring these groups together and together with people not on the working groups.

So the model of future meetings seems to be agreed, first implemented at the NIKHEF HEPiX meeting, week of 19th May, one day LCG/Large Cluster workshop, one days for HEPNT Windows workshop or discussions, 3 days HEPiX.

What is Everyone Doing with AFS – Lisa Giacchetti (FNAL)

She will eventually prepare a site survey but in the meantime there was a brief discussion on what servers and what versions are being used at the various sites. FNAL still use Transarc AFS on the servers under Solaris. They use OpenAFS on Linux clients but Transarc on other clients. Few problems seen at this time. FNAL have tested Kerberos 5 for cross-cell authentication and got it working. Limited local support structure, who could do local support if they were to adopt OpenAFS. FNAL had heard that Transarc might extend their support beyond the announced end-date.

CERN has standardised on Solaris servers with OpenAFS since this summer. All architectures use OpenAFS on the clients. Some tests have been done on Kerberos 5 (as presented by Wolfgang Friebel to C5). After a bad experience with Linux servers some time ago, we are once again experimenting with OpenAFS servers under Linux for scratch space and so far there is no recurrence of earlier problems.

BNL has 2 cells with IBM servers and Transarc AFS; OpenAFS is used on Linux clients and a mix of Open and Transarc versions on Solaris clients. Looking at Kerberos 5 and what needs to happen to make this work. Also looking at Linux servers and OpenAFS but put off by rumours on the mailing lists. No Windows clients used.

SLAC: SUN servers, Transarc AFS, no plans for OpenAFS for servers but most Linux clients are OpenAFS. Problems until recently on OpenAFS on dual-node clients, now fixed. Transarc have indeed asked if SLAC wish to extend support beyond the end of the announced life but have not yet proposed a price.

RAL: production servers use Transarc AFS, clients use OpenAFS on both Linux and Solaris.

LAL: no servers but OpenAFS on all clients, including Windows where OpenAFS was the only possibility for Windows 2000 when the decision was taken last year. No problems.

DESY: major upgrade in progress – moving to a SUN SAN. During this they will migrate from Transarc to OpenAFS. On the client side, still Transarc on all commercial UNIX clients but OpenAFS on Linux and Windows clients. Some issues with Kerberos. DESY have some expertise in supporting AFS but only as

needed. They found that paying for IBM support for Transarc AFS was not cost-effective given the quality of the service provided.

INFN: national cell is moving to OpenAFS on Linux servers. Some local cells already use this combination. OpenAFS used on most clients, including Windows, except for the commercial UNIX clients. In Lecce, the local cell is based on Kerberos 5. Found that the OpenAFS mailing list was responsive enough for solving problems, much more responsive than relying on IBM.

Saclay: run OpenAFS servers on Linux and have successful experience in running a mixed Transarc/OpenAFS server cell.

BaBar Tier A Centre at CCIN2P3 – Jean-Yves Nief

He described the configuration installed and the resources used. He spoke of some of the problems encountered such as sometimes erratic network performance, Objectivity and HPSS bugs and problems and load peaks before the summer conferences. BaBar is interested in creating a Grid structure and the Storage Resource Broker (SRB) and MetaCatalog software have been tested at Lyon. Already, using EDG Middleware, it is possible to submit jobs from Lyon to RAL and to SLAC. More CPUs and more Objectivity servers will be added, HPSS load will be decreased by compressing the data on disc and a dynamic load balancing will be introduced on the Objectivity servers next year.

CVS Setup at CCIN2P3 and EDG CVS Tools– Yannick Patois

The CVS repository is on AFS disc. Originally login restricted to certain users and CVS actions were permitted via a special CVS pseudo-shell. This needed to be broadened to accommodate EDG users. They use CVSweb. The local service for IN2P3 users only started a few weeks ago while the EDG service has 120 accounts, 1 year's experience and 360 MB of information.

For EDG they created special build tools (autobuild) to take account of inter-dependencies and also tools to arrange and publish RPMs.

Measuring the Hardware Reliability of FNAL Farms – Steve Timms

The aim was to measure the hardware failure rate and the total cost of ownership. He first described the initial burn-in tests performed on new nodes; failures during this period are covered by the vendor's warranty. They created a so-called "Lemon law" — any node down for 5 straight days or 5 separate instances during the burn-in period (30 days) must be completely replaced. He gave some examples of frequent sources of problems; as in many places IDE discs form a major proportion of them.

In measuring costs, they defined a Fermi unit of power – can HEP not define a single HEP unit instead of each lab defining its own? They estimate that over the past 3 years, they have been able to buy up to 8 times the power per dollar as measured in these units. Analysing hardware failure rates, they see improvements with newer acquisitions as might be expected but large and unexplained differences between 3 more or less identical 50 node farms bought in 1999 and dedicated to CDF, D0 and Fixed Target experiments respectively.

Their total cost of ownership includes

- depreciation and residual costs
- maintenance
- electricity

- memory and other upgrades
- operating personnel costs (not development or user time)

Their total cost of the clusters bought in 1999 at a cost of $410K has been $415K by these calculations (assuming a residual value of $70K).

Final Session

Because of the timing of my flight, I missed the last afternoon but two of the sessions were from Philippe and Tim so I guess we can get the overheads from them or perhaps have them repeat the talks at C5 if they are of general interest. The last was on NFS performance which is of less interest to CERN.

Alan Silverman
27th October 2002