HEPiX Fall 2005 Trip Report
SLAC
10th to 14th October 2005

Introduction

The Fall (autumn for us Europeans) 2005 HEPiX meeting was held at SLAC. Confirming the trend from last year, the US meetings now attract as large an audience as European meetings, over 80 this time from some 27 institutes in Europe and North America. The meeting layout this time started with a day of sessions on Collaboration Tools, 3 days of "normal" HEPiX sessions and finally a half day on talks related to hardware issues.

The meeting was impeccably organised by Chuck Boeheim and colleagues. Accommodation was in the on-site SLAC hostel. As usual, virtually all the overheads can be consulted from the web site. I have notes on all the sessions with links to all the overheads (with a single exception). There are also links in the notes where appropriate to relevant sites.

No-one is expected to read all of what follows but I recommend you consult the agenda and skip to the session(s) which interest you.

Executive Summary

The following points are highlighted. They are in no particular order and more detail can be found in the text and in the overheads.

- The first day was an interesting review of some common collaborative tools, highlighting possible trends in video-conferencing in particular. The afternoon session was largely drawn from information gathered by LCG RTAG 12 including a summary of that report by the RTAG12 chairman who closed his talk with some personal suggestions on how to move forward on the report's recommendations. There was also a short discussion on how to select the most appropriate EDMS tool for the ILC study.
- Collaboration was also evident throughout the week where many system and sysadmin tools are appearing in labs across the HEP community – Scientific Linux of course, Quattor, dCache, Xrootd and so on. And at this meeting came the first indication that CERN's monitoring tool, Lemon, may be the latest package to be of more general interest.
- As at the last meeting, most sites reporting CPU farm increases are buying Opteron systems, BNL and CERN being exceptions. In this area, there was a most interesting set of benchmarks on the different processors performed at GridKA and reported on the Friday – see the overheads.
- The CERN Certificate talk from IS Group raised a number of questions: why is CERN/IT building a second root CA (after the LCG one); which master CA has authorised this new one; why should outside users trust it; should it be issuing certificates to systems outside CERN as was suggested in the talk? Perhaps we need a discussion of this inside the house with some urgency.
- Batch workshop follow-up – after Tim Bell presented some changes made as a result of input from the last HEPiX meeting, Francesco Prelz (main EGEE batch system developer) presented his open questions and there was a productive exchange and a plan to move forward on a topic which has been unresolved for some time. We, including Francesco, have agreed to follow this up at the next meeting. Francesco said he was glad to have come and felt it had been valuable and that HEPiX has provided a welcome forum to make this happen. The HEPiX board have agreed to make this happen in CASPUR and perhaps in future meetings as needed.

- Linux issues: Fermilab proposed that we consider making HEP code LSB-compliant. There was a discussion about SL/SLC differences and compatibility. Some concern was expressed that CERN, unlike other labs, had formalised its tailoring of Scientific Linux (SL) into a formal SLC release. It apparently causes confusion among some users, at least those outside CERN, associated with LHC experiments. We either need to reconsider this branding or advertise better what it is, why we do it and the fact that it is, or should be, binary compatible with the SL from Fermilab.
- An interesting set of sessions on Thursday morning on disc and storage systems including DPM, SRB, Xrootd, dCache and Panasas
- CERN talks: all IS groups included a live demo; not always possible (Hege thought of doing one for LICmon but offsite access to the pages is forbidden) but it should be considered where possible.
- There was a deliberate effort made by the organisers to schedule a number of talks targeted more to the physics happening, or planned, at SLAC rather than directly at our usual range of only computing subjects. These (from the Director, Richard Mount and an astrophysics theorist) were quite interesting and, especially the last one, entertaining.
- Next meetings will be in CASPUR, April 3rd to 7th 2005 and next Fall/Autumn in Jefferson Lab.

Collaboration Tools

The week started with a day of talks on this theme, including some from non-typical HEPiX speakers who detoured past SLAC on their way to an ESnet conference on the subject later in the week in Berkeley. Thanks to Thomas Baron for helping me set this up.

Wiki at GSI – C.Huhn

Wiki at GSI was described at the last meeting when it was mainly used for documentation purposes but since then its use for collaborative tasks has grown significantly. There are now some 270 users, growing at some 50 registrations per month. The first collaboration tool added was the Holidaylist plugin for marking absences in a simple manner. A more advanced tool is for creating and managing meeting minutes where the search feature is very popular; there is a web form and a template for creating new minutes and there is an Approval plugin which implements a workflow for agreeing and finalising the minutes. Other wiki-based tools include a logbook using the Comment plugin and a light-weight trouble ticket scheme using the ActionTracker plugin. More tools are planned.

Wiki at Karslruhe – A. Grindler

In FZK, a major use of wiki is in change and configuration management. After describing situations where these could be useful, the speaker listed the objectives and tasks of configuration and change management and how this leads to a database, relationships between the entities in this database and actions on the entities. They describe the data from a customer view and they have developed wiki templates to model this data. Using wiki they have built a portal for various administrative tools such user management, change calendar, change logging, etc.

Form Factory – Wolfgang Friebel (DESY)

This is a tool being built at DESY Zeuthen to assist in the building of web forms by providing templates to make the task simpler for non-experts. Wolfgang chose PERL as the toolbox, selecting CGI.pm which is the de facto standard web-building module. The layout involves cascading style sheets but care must be taken to support old browsers (such as Netscape 4 which remains remarkably popular) so sample files are provided. Form Factory is actually a framework of about 500 lines of code which implements the workflow of building a web form, by reading and parsing a configuration file for preparing, validating,

installing and testing a web form, calling as needed the various PERL modules which number in total some 30,000 lines of code.

InDiCo – Thomas Baron (CERN)

Thomas started with the history of the tool from its base of CDS Agenda which now manages over 20,000 events. From this there was created an international, EU-funded project to build InDiCo, specifically targeted for conferences. InDiCo currently manages over 50 conferences and there are plans to migrate CDS Agenda into it. The main development tool is Python, it runs on an Apache web server and uses the Zope Object Database for storing the conference material and tools such as XML and XSL to build the timetables. It is fully open source and available under GPL. He described in detail the elements of a conference and the various interactions and the processes to be applied to these. InDiCo can be used for anything from simple meetings of very small teams up to multi-stream, multi-day conferences such as CHEP04. InDiCo offers a search tool, it can archive the material and it can be used to produce proceedings. He ended with a long list of planned and hoped-for improvements and further developments.

VoIP with Skype – Alf Wachsmann (SLAC)

Skype is a company based in Luxembourg, now owned by eBay, offering peer to peer telephony over IP. It uses proprietary signalling but the APIs are open and documented. Alf described the features and how they are used, some of which, such as a call to a non-IP telephone line, must be paid for. Calls within the Skype network are free and the rates for other use are cheap and published on their web page. There is heavy emphasis on data security and interchanged data is 256-bit encrypted. As a peer to peer protocol, while one attraction is that there is no central server, this is also a major drawback because if a Skype user machine seems to be reliable, it is declared a supernode and that system becomes a Skype router, potentialally risking a serious affect on performance of the system and the attached network. This is one reason why some sites, CERN is one example, discourage its use.

Access Grid in the UK – John Gordon (RAL)

Access Grid (AG) is widely used in the UK, more in the e-Science programme rather than in the HEP community so far. Coming originally from Argonne, it is also heavily used in the US. Compared with the more common (in HEP at least) VRVS, it is more aimed at high performance video conferences as it needs a serious infrastructure in terms of video equipment and it usually involves an operator. Like VRVS there is the concept of a virtual conference room and in fact it interfaces to VRVS. The UK AG community has set up an Access Grid support centre. John listed the various services offered such as shared powerpoint presentations, conference recording, quality assurance tests, training and documentation.

Collaborative Tools in NICE at CERN – Alex Lossent

Alex listed some of the features of CERN's NICE service for Windows and their application for collaborative work. DFS offers a unified Windows file repository for home directories and workspaces and sharing is available via a self-service administration interface including control of security and access rights. Outlook is used as the base for shared calendars as well as shared mails (e.g. for a manager's secretary). This helps for task assignment and for scheduling meetings. All the features described are available both inside and outside CERN using web-based tools, or WebDAV access to DFS, or Windows Terminal Services (WTS) and Alex demonstrated a few of them. Via a web browser or via WTS, they are also available on non-Windows platforms. Other tools include workflow management which has been enabled via the "send for review" feature of the latest releases of Office, Instant Messenger is offered via the IM Microsoft tool although a replacement tool is expected from the supplier. Current work is going on for the creation and management of e-groups for web-hosted collaborations for defined communities.

VRVS to EVO – Philippe Galvez (Caltech)

From its start in 1995, there are now some 18,500 registered VRVS users from 120 countries with 1100 meetings per month worldwide involving some 4500 participants per month. It currently is or has been used by at least 72 HENP experiments but also by many non-HENP teams (most recently including ITER). As mentioned earlier, the initials stand for Virtual Room Videoconferencing System. In its 10 year history it has bee re-architected several times and is now supports MBONE, H.323 and MPEG. There are 82 reflectors worldwide. Support is shared by Caltech, CERN and a team in Slovakia. Developments today are focused on making it even more robust and scalable in a real-time environment. The first step was to understand its failings, what can and does go wrong? Many of these (failures of end user equipment and international networking for example) are not within control of the tool. It appears that multicasting has reached its limits so we should look to an overlayed network by deploying Java-based intelligent software agents (called Pandas) around the network to work around recognised problems. These interact with Koala, an intelligent VRVS client agent in a dynamic manner, reacting and working around problems as they occur. Philippe compared this with Skype, described earlier, but with the major difference that the VRVS agents are under user control.

LCG RTAG Report on Collaborative Tools – Steven Goldfarb (Uni Michigan)

Steven started by presenting the mandate of the study group which began work in January 2004. Participation was drawn widely from the LHC experimental collaborations, from CERN/IT and including well-known experts in the field. There were weekly meetings and with people charged with providing conferencing services and with users of those services. Progress reports were presented in June and November and the final report in April 2005. The summary included phrases such as "large and growing gap", "increasing need", "growing popularity" and "lack of dedicated resources". Unfortunately they also concluded that solutions would be complex, that no turn-key solutions appear to be available and research and development would be needed. He then covered the principle recommendations in turn. See overheads for details. The cost of their recommendations was estimated at 1.5M CHF in equipment investment and 500K CHF per year in dedicated human resources but this should be compared with the several MCHF per year of lost resources in travel, dysfunctional meetings, etc. among the experimental teams. Since the report was presented, there has been some follow-up, some of it described already in this today's sessions, and others are planned but more needs to be done – when funding can be identified. He personally believes that the experiments should identify a liaison person, establish a collaboration programme within the experiments and what resources are available and then make a proposal to CERN/IT. On the other hand, he expects CERN/IT to formally reply to the report, what they agree with and what not and they have to agree to take the leading role described in the report.

Remote Conferencing Working Group Report – Brooks Collins (SLAC)

The RCWG charter is to provide insight and requirements from the DoE community to ECS (ESnet Collaboration Services) for usage, problems and service planning. They also test and feedback on new services and they conduct information exchange in the topic. There are 22 members, some users rather than IT professionals, mostly US-based but including the main players in Europe. Brooks described some ECS services in detail and features coming in the future. He presented the current RCWG activities and some which are planned, a number of which are fairly US-centric. He feels that more publicity of their activities is desirable and they should attract more DoE sites and universities. He ended with some observations on the fast-moving technology involved.

Collaboration Tools at IN2P3 and Beyond – Christian Helft

Christian described how the 20 HENP sites in France are connected for video-conferencing. This includes a home-written management system called VACS. It currently supports some 40 conferences per month, slowly increasing but approaching saturation at peak hours. Unfortunately, scheduling is

mandatory and has to be done by hand and the hardware is becoming obsolete and expensive to maintain. They are therefore moving to more modern equipment. CERN is participating and work is going on to update VACS for this new equipment, adding functionality made possible by features coming with this more modern equipment such as the ability to call participants, giving the floor to a speaker and an H.239 remote documentation presentation feature. It will be linked to InDiCo as a document repository and it will interface to ECS and to VRVS. At this point, he included an overhead asking why HEP is not using more web conferencing despite the availability of both ECS and credible tools (e.g. from Microsoft) for this. He then moved on to making a plea for a consistent video-conferencing service which linked the various tools such as those already discussed, for example a unified scheduling scheme. Also we should have a single instant messaging scheme but not a commercial one. He praised FNAL where care is taken in properly installing video facilities in their meeting rooms. He ended by suggesting the creation of a conferencing tools deployment team which could give guidance to the collaboration tools service which was one of the recommendations of the LCG RTAG on collaboration tools, of which he was a member.

Web Hosting at CERN – Alex Lossent

The goal of this service is to reduce the number of locally, and often badly, managed web services across the site by offering a staffed and supported central service. There are 25 servers, they host 7000 web sites and service 2M requests per day. Within the support team 1.5 FTE are dedicated to its support and development. Web sites can be stored on Windows DFS or AFS. He described the search mechanism, support for web applications via tools such as ASP, PHP, Java, etc and web authoring tools (Frontpage and Dreamweaver, and Visual Studio.NET for ASP.NET applications). In the future, they will move towards e-groups for web-hosted collaboration tools; add support for authentication with certificates and improve the search facilities.

EDMS

Although this topic was not covered (no-one volunteered a talk) Tom Markovitx[1] from the ILC study team based at SLAC asked the audience's opinion on how to choose a document management scheme and if we could recommend one which would scale to ILC size. Several sites reported different installed systems for similar tasks, usually more than one per site but targeted at different audiences (engineers as opposed to physicists for example). If there was any conclusion it was that one should not try to force one tool to satisfy all needs in this area, especially if the tool was initially designed only to satisfy a subset.

Site Reports

Introduction – Jonathan Dorfan (SLAC Director)

He presented the new structure of the lab, indicating its future new focus on non-HEP sciences, in particular astrophysics (the new Kavli Institute for Particle Astrophysics and Cosmology, KIPAC) and the Linac Coherent Light Source (LCLS) project. He described the first 3 space telescope projects planned for KIPAC, the first of which launches in 2007; there is also a plan for a land telescope dark energy search project for which "first light" is due in 2012. LCLS will be the world's first X-ray laser. It is being built in the old SPEAR hall and is due to begin operation in FY2009. It will have 10^{10} times the instantaneous brightness of a conventional synchrotron and he stated that no-one has ever done science at this resolution and simulation is an interesting challenge.

[1] Tom is a member of a study team looking into this subject, on which the CERN representative is John Ferguson. Speaking to Tom later, he is aware of both tools in this area in CERN (EDMS from TS Department and IT's CDS) as well as similar tools in other labs such as DESY and FNAL.

Turning to HEP, he described the plans for BaBar which is expected to collect 5PB of data by the time it completes at the end of 2008, a factor of 4 more than has been gathered up-to-date. SLAC is also involved with ILC simulation studies.

To support all this SLAC intends to build a computing structure to support these activities in a consistent manner across the various disciplines described, harnessing computer science as and when required.

SLAC – currently adding 360 dual-CPU, dual core Opteron systems, replacing their Netra T1 and VA-Linux clusters; LSF was upgraded that morning to 6.1 and moving the LSF master to Solaris x86 on a quad Opteron system. HPSS will go to V5.1 later in October, moving the metadata to a DB2 database format and moving partly away from DCE. While gearing up for RHEL4, they are also developing support for 64-bit mode for Opteron large memory applications. Solaris x64 will also be used for more dedicated services. To reduce in-house support costs, they have chosen to acquire an application package of pre-compiled binaries of Open Source software from The Written Word, chosen over various competitors partly because of their support offering, including regular security updates.

RAL – a new procurement aims at adding 1400 SPECinto2000 and is targeted for the first half of 2006. CASTOR 1 is under test but it is taking some time to remove many CERN-specific features. Procurement for a new robot is also underway, 3PB, 10 drives, to be ordered later this month and delivered in time for LCG Service Challenge 4 (SC4) for which also various links are in the process of being upgraded. They still run AFS on AIX but have plans to move to Linux hosts. The home file system has been moved from Solaris to a Linux host service and stability has significantly improved.

CERN – as traditional in CERN site reports, Helge started by describing the latest department re-organisation, this time of the physics groups. He then listed the recent hardware additions in farm PCs (+200 dual Nocona systems), disk servers (+116 nodes, +900TB), mid-range servers (+112) and a 32 node Infiniband-based cluster for the Theory Group. He described the recent dedicated ATLAS DAQ test where 730 systems were given to them and then re-installed very easily for general use – another successful stress test for Quattor. The serial console infrastructure is largely completed and is proving very useful. He mentioned the LHC@HOME service for LHC beam orbit simulations where around 100 old PC systems have been dedicated to running BOINC for this. Most other changes and new developments will be covered in particular talks later in the week.

RHIC/BNL/USATLAS – recently added 6 new staff members, arriving at a total of 31. RHIC run 6 will start in December. There are 20 now shelves of Panasas storage with 100TB but there are some serious problems (frequent node crashes, no user quotas, no LDAP) – see later talk. Recently bought a new STK 8500 robot with 20 LT03 drives and 6500 slots. Added 864 Dell dual Xeons for a total of 2000 systems; 677 TB of local disc storage. Both LSF and Condor in use but most of the farm uses Condor although LSF is heavily used on certain nodes. Ganglia is used for monitoring the farm and Nagios is being evaluated (it is already used for many infrastructure nodes). Performed tests on dual-core Opteron systems in preparation for the next purchase and found good power consumption results (20% less than Xeons); will test dual-core Xeons soon. They use OSG for production Grid work for both RHIC and US ATLAS and are deploying LCG. Becoming heavy users of dCache with separate instances for PHENIX and US ATLAS (see later talk).

INFN – the Computing and Network steering committee has been refreshed; the last act of the outgoing one was to sign a nation-wide licence for all INFN sites for Sophos anti-virus and junk mail control software. Local AFS cells are moving or have moved to Linux hosts and OpenAFS and Kerberos 4 has

now gone. There is increasing use of LSF in the batch farms. The Tier 1 centre at CNAF has recently been upgraded with 150 new Opteron systems.

Oxford University – an interesting view of running a small Tier 2 site with limited resources, human and material. Oxford is also a member of the UK Southgrid, all of whose sites have upgraded to LCG version 2.6.0 and some of whom also participate in the Biomed data challenge.

TRIUMF – acquired mini-Google - US$3K with 1 year support, supports up to 100K documents[2]. They are doing a lot of work on networking and they expect to have a 10GbE permanent Lightpath link to CERN by the end of 2005. They are participating as a Tier 1 site in the ATLAS Service Challenge with a small cluster. He has found that adding lots of memory greatly improves stability of central services. Starting to use Plone for content management for the TRIUMF web site. File backup is still an issue and they are moving to Amanda, a freely-available backup tool from the University of mMaryland.

CCIN2P3 – now 747 worker nodes and a range of file systems (AFS, NFS, HPSS, XROOTD and dCache). AFS is being used to install LCG middleware; this permits coexistence of LCG and gLite during migration to the latter. Thy have a small (42 nodes) parallel production farm for some non-HEP teams. Still IBM AFS V4 but plans for V5 and OpenAFS.

GSI – They have successfully moved to Exchange 2003 mail server and the batch farm has been upgraded with more Opteron servers which needed Debian to be upgraded for 64bit support. GSI participated in ALICE service challenges although their low bandwidth meant that this blocked all other accesses so they will need to upgrade their network connection (currently 30Mbps). They claim that adding grey-listing for automatically selected domains and Spamassassin 3.1.0 means that "spam is no longer a problem at GSI". After some tests and studies (see overheads for some of the reasons, mostly related to cooling and the need or not for fans on the boards) they decided to add 140 dual CPU, dual-core Opterons (560 processors) in the near future. Their positive experience with serial ATA file systems continues (lots of details in the overheads) and they have invested heavily in more. They are moving from VPN to Citrix Terminal Server and are happy with the improved security and they will close the VPN service very shortly.

GridKA – the worker node cluster has been expanded to 780 systems, 1560 CPUs, the addition being mostly of Opterons but heat problems meant they needed to shut down some older Xeon systems while the total cooling power is being augmented. Also they have expanded their disc farm and the dCache storage pool. They have started a large tender operation which, because of its size (>200K Euros), has to be EU-wide and takes about 6 months and they will do the same for more disc storage. Meanwhile, more water-cooled cabinets have been ordered and more rooms are being equipped for this as they require to expand their centre. They are currently enhancing various components ready for the service phase of LCG SC3

DAPNIA – DAPNIA has migrated from Exchange 5.5 to Exchange 2003 and other parts of Saclay and then CEA will migrate later. They have recently expanded their Alpha Tru64-based disc store. They are trying to build a Tier 2 site (the GRIF project – see later talk).

LAL – as new Director, Guy Wormser, has officially "blessed" their Grid activities and formalised a grid group under Cal Loomis who has been given a permanent position. They have added 20 dual CPU Opterons, one-third for GRIF, as well as 6TB of disc and 9TB on order (half this disc for GRIF). Using Quattor for LCG UI installation on SL 3.0.5. Quattor is in fact used for configuring all systems and

[2] It will be interesting to watch how this progresses. We will schedule a special session in 12-18 months on search engines when CERN has some experience with our planned mini-project on this.

although there is a learning curve, they are very pleased with it; Michel is discussing with the CERN support team on replacing the CDB component by more modern methods (SVN + http). They have deployed

- SPIP (originally a French product, now multi-lingual) for web site content management.
- TRAC, a wiki-based product for project management, trouble ticket tracking, etc

And they are investigating web-based agenda possibilities to replace their current scheme.

Manchester University – possibly the first appearance at HEPiX of a Tier 3 site. They use SL3 but the speaker wonders what are the differences with SLC3?[3] They support a number of HEP experiments, mostly on AMD systems but their ATLAS team is entirely based on Mac.

JLab – Upgrading soon to 10Gb MAN with connectivity to ESNet. Implementing secure wireless using WPA with support for all the usual operating systems. They have installed a secure mail server and upgraded their SMTP hardware. They, like SLAC, base themselves on Redhat's Enterprise Linux (RHEL) and they use the Redhat Network Satellite for support and patches. On the Windows side, virus protection is provided by Symantec. They have installed a 25TB Panasas system and are working with them on some issues. They have installed 2 STK B280 systems (30TB) and are evaluating a B680 which is similar but uses SATA discs. After recent problems with generators and UPS they have decided to build a new computer room.

Scotgrid – another LCG Tier 2 presenting at HEPiX for the first time (although they attended the Edinburgh meeting). There are 3 sites, (one of which one is actually in England) but the distribution of boxes, CPU power and disc capacity is spread inconsistently across the sites. Batch is based on PBS and they are gearing up to participate in LCG SC4 next year, for example with all sites moving towards (or already using) the LCG Disc Pool Manager (see later talk).

DESY – first experiments have started on their free electron laser facility and some ILC studies are based now in DESY. They have offered to be a Tier 2 node for LCG for both ATLAS and CMS for whom they have deployed LCG 2.6.0 on SL 3.0.x using Quattor and they will dedicate some 200 CPUs to this by the end of the year. A new machine room will be opened in spring 2006 in Hamburg. They participate in the ILDG, International Lattice DataGrid, for which DESY coordinates deployment of an LCG-based datagrid. Windows and Linux are the main platforms but they are preparing to move their (few) SUNs to Solaris 10. Apart from some dedicated older systems, migration of the Windows domain to XP client and 2003 server is complete. DESY is now SuSE-free, having moved mostly to SL-based Linux but there are also some (unsupported) Debian systems onsite. After prototyping for a year, the first production apeNEXT massively parallel supercomputer is being deployed, targeted at Lattice QCD.

NERSC – LSF has been replaced by SGE (SUN Grid Engine) 6.0. Looking at SL4; bringing in one-time passwords for administration functions to avoid use of root password. Lustre 1.2.4 is being used but not without some problems - "Lustre is still a little green". The Jacquard 360 node dual Opteron system with 30TB of discs has passed its acceptance tests; it is based on Infiniband and the PBSpro batch scheme. In the past 8 months, NERSC has had a number of planned and unplanned power outages. One of the planned ones was to power an "un-named" new system needing .5MW about to be installed.

HEPiX General Sessions

[3] Since he had brought up the question on Tuesday morning already, Andras had by then got the answer from CERN's Linux team and had mailed him the list of differences – see Linux BOF later.

OpenLDAP Configuration and Tuning – R.Petkus (RHIC/USATLAS)

Originally they were heavy users of NIS but it is insecure (clear text authentication), not scalable, more or less UNIX-only, flat name space, and there are doubts over continued support. Since they were switching Linux farms, why not use the opportunity to switch to LDAP which removes many of these drawbacks. OpenLDAP was chosen because it is open source, platform-independent and has a rich feature set. He presented some of the features and then some tuning tricks he had used such as indexing, setting the cache size to the amount of entries to be stored in memory, a sparing use of logging, increasing thread count and so on. LDAP load balancing helps if you need fault tolerance and/or high performance and he showed some possible configurations. OpenLDAP is not without its own drawbacks and he listed a few – logging is expensive in resources, some particular problems on Solaris and a few more but in general they are very happy, especially with the Linux version. See overheads for more details.

Software Licensing at CERN – Hege Hansbakk

Hege reported on the LICMON scheme devised in CERN to manage commercial licences. She described different licence models. LICMON not only manages the licences, it monitors their use so that we can better judge how many of each are really required. LICMON interfaces to many commercial licensing packages on UNIX, Linux and Windows, and MacOS support is being added. She showed how the usage data is produced, where it is stored and how it is displayed but because access to the database is restricted to onsite use, she could only show screenshots.

CERN Certification Authority – Ruben Gaspar

Ruben started by describing the secure architecture for creating a certificate. They support software and smartcard certificates. Use of certificates is growing for short term actions (e.g. to sign and encrypt mails) and long term uses are being envisioned (eg. embed a smartcard in a CERN access card) and he illustrated several examples. He said it takes into account the needs of grid authentication and work is in progress with TS Dept such that the next generation of CERN access cards may include a smartcard feature. As usual for all IS groups talks, he ended with a fine demo of the presentation material. During the questions, several people in the audience expressed some reservations on some aspects – why a second CERN root authority for example; what are the risks in issuing host certificates for hosts outside CERN?

Grid Single Sign-On – John Gordon (RAL)

Today there are different user bases for different communities on the RAL site and they would like to create a corporate database where some users have grid identities already. They would like to make the various access permissions available through a single password. If on-site, one would use a federal Kerberos id; if off-site, one would use a certificate. John explained how to get a certificate, renew it and how it is used to access the required information. The scheme is currently being linked to VOMS.

Scientific Computing at SLAC – Richard Mount

Scientific computing must be aligned with the mission of the lab and responsive to its needs but not subservient to it. The drivers, BaBar, KIPAC, etc were described earlier by the Director. Many of them are heavily data-intensive and/or require deep analysis. Funding for computing for these new projects is rather meagre. The Computing Dept (SCCS) runs some 4000 processors, almost all Linux PCs (Solaris will be gone by year-end as far as physics processing is concerned) plus a few specialised farms for specific applications and for disc and tape storage. Currently-funded research areas include huge memory systems for data analysis and scalable data intensive systems, grids and security, network research and internet monitoring and prediction. Driven by the new projects, competence is growing (and has to grow) in parallel computing and visualisation.

The Petacache project, where Randy Melen is the project leader, aims to revolutionise the handling of complex scientific databases by using huge memory systems. The motivation comes from HEP's experience in dealing with feature databases using sparse (or random) access. The key is to avoid as far as possible disc access with its inherent latency when performing random data access. The prototype has SUN V20z systems with Opteron chips and 16GB of memory each where up to 2TB is possible; the next generation should reach 10TB using new memory technologies known to be coming in 3 to 10 years. Via XROOTD, BaBar are able to benefit already from this system for random data access. Richard ended with two possible views of scientific computing at SLAC in 2012, one deeply pessimistic but the other has SLAC participating actively in all the planned projects for the lab plus LHC, ILC and perhaps others.

Security Update – Bob Cowles

The traditional "Bob Cowles Show" started with a list of the features and output of a new password cracking tool Bob had discovered (Cain and Abel). Then, with the audience holding its breath, he displayed the list of passwords he had collected the previous day[4]. Phishing and pharming are still on the increase as is spyware and identity theft in general. He recommends that one should never type in a username/password if the URL is not https. He described various new technologies such as Bluetooth, smartcards, one time passwords; then he described some of their weaknesses. He is sceptical if they will make a difference, other than to the price of bypassing them. SLAC recently had a visit from a DoE Site Assistance Unit who checked that SLAC indeed have the necessary documentation required by recent government standards. This included a penetration test which succeeded in gaining access to a Windows 2000 SP3 server still running and from there they accessed most other Windows servers on site. [SLAC claim to have been unaware that Windows 2000 SP3 was no longer supported after June 30.]

Fermilab Plone Update – Marc Mengel

Plone is a customisable content management framework supporting different kinds of content, not just wikis. It has a rich suite of features and tools supporting wikis, workflows, forms management, trouble tickets, searches and so on. It supports WebDAV and there is a myriad of plug-ins available. Marc's team have used this for example to create a new web site for HEPiX. He gave demos of using it for creating a log book and for approving committee requests. He said that most of the tools he had created involved only a few hundred lines of code built round a template. The speed of the demo was rather jerky but he claimed it was not usually that slow and they have plans for a new configuration which should improve this. Another performance gain should come from upgrading to newer versions but there is an issue with export/import between particular versions. Another issue is that it uses https and most browsers do not cache SSL content by default and most users do not know or think to change this. Despite these drawbacks, Marc and his team happy with the product and intend to work further on it, trying to make it more globally acceptable inside the lab although many potential users have been scared off by its original poor performance.

WSUS/SUS for Patch Deployment in DESY – R.Baltrusch

There are 2500 clients in the DESY domain to be managed. They started using SUS but moved to the successor, WSUS, when it was released. Reinhard described (in some detail, see overheads) the configuration and operation of SUS and how WSUS not only replaced it but improved on some weak points, in particular controlling the success of an upgrade on a client. While WSUS has some nice new features such as the definition of target groups, there is interference with DESY's chosen Netinstall tool for Office updates and there is not enough flexibility for a client to delay the installation of a Service Pack which could be lengthy process. The other main remaining problems are how to deal with

[4] I'm glad to say I don't have to change mine – thanks to the NICE and AIS teams for protecting their authentication processes and allowing me work securely from afar.

unattended or home PCs and with roaming laptops. There is an API so in principle it could be used to install non-Microsoft products but the API is not documented.

Microsoft Clustering – Sean Roberts and Jean Pierre (SLAC)

This was a review of the various offerings from Microsoft – network or component load balancing clusters, server or compute clusters. These were described as far as configuration, state-full or stateless, possible uses, etc. Sharepoint was chosen as a specific example and how this fits into a network load balancing cluster was described.

NICE Admin with no Admin Privilege – Ruben Gaspar

Why is this desirable? One reason is that not having admin privileges in normal desktop work increases system security. The aim is to emulate "su" on UNIX and something like this runs fine on Windows Terminal Server. Options include –

- Fast User switching on Windows XP would be one option but it is not possible in a domain client.
- Another option is a blank password which is in fact safer than a weak password (he claims) but this may break site security (as it does at CERN).
- The RunAs application – but then there is a problem running MSI.
- Use a local admin account but then you can't access network resources.

The proposed method is, on demand, to run a script which adds your current account to the Admin group, forks a process to perform the action and then removes the account from the Admin group. Ruben described the various steps in some detail, the processes which run on the user's desktop and the web service which authorises, or not, the user to use this tool (only authorised users may do so on a given system – normally only the owner). And as usual for NICE talks, there was a short demo. The scheme is currently in test in IS Group and there are still a few issues to resolve before release but it should be a good solution to the stated problem, until Windows Vista (Longhorn) arrives.

The Kavli Institute – Tom Abel

Another in the series proposed by the organisers at this meeting on the applications for which we provide support rather than the support activities themselves. He began with a very entertaining animation of a fly-through of recorded points in the universe. The task ahead (or at least one of them) at this new institute is to explain galaxy formation from its very beginnings. He presented graphically some of the techniques that he, as a theorist in the field, uses to project forward galaxy growth from the first moments of the universe. He ended with describing some of the other activities planned or already started in the institute. There is expected to be heavy use of parallel and supercomputing, in passing confirming the need for the huge memory systems mentioned by Richard Mount the previous day. Another difference from HEP computing is their need not only for double precision floating point (instead of almost all integer in HEP computing), they need quad precision because of the sizes involved. Visually this presentation was stunning.

Condor Flocking at BNL – A.Withers

They have 5 Condor pools for 5 experiments with the CPU resources split more or less evenly except for one small pool. Clearly this configuration is not the most efficient when viewed from a lab perspective. . Hence the proposal to use of flocking so that idle cycles can be shared among pools using algorithms such that "flocked" jobs do not take resources which could be used by "native" jobs in each pool; the activation of a "native" job would suspend or even kill a running flocked job. Techically it is fairly eqasy to permit this thanks to running with a standard O/S, common NFS and Panasas disc mounts and a

single LDAP database. Since large-scale testing has not started yet, there is still some concern about load on Condor processes. More serious are the concerns of some of the "owners" of some resources against sharing. There were several questions from the audience about the overall efficiency of suspending, or worse evicting, a running job which may be almost complete. It was suggested that promoting a campaign for fairshares which may be more overall productive.

Batch Workshop Follow-up – Tim Bell

This was a follow-up of issues raised at the 2 day Batch Workshop held at the previous meeting. Much work has been done on batch plugins to either fix them or offer more facilities for local admins to fix them. There is now a basic so-called "HEPiX batch web page" which is maintained and advertises the current state of the LCG sites[5]. A major issue in FZK was the interaction between the LCG job schedulers and local schedulers and there are many improvements in GLUE 1.2 to try to address these, especially in the calculation of ERT, estimated response times. The new ERT implementation has been done for LSF and PBS and volunteers are now sought for the other batch systems. Work on GLUE version 2 will start soon and suggestions are invited, to be sent to Laurence Field at CERN.

EGEE Batch System Work – Francesco Prelz

This was the "reality check" which the Tim had foreseen in his talk. Francesco is one of the lead developers of the EGEE batch scheduler. How to abstract the different batch systems and how to delegate job control from the grid batch controller to them? He has started from the view adopted by Condor, that there has to be a shared convention understood by the various job controllers and by the grid batch controllers to describe job attributes. He showed some possible attributes for compute elements, split into blocks such as host features, local policy, host state, etc. Most of the work goes into exceptions – jobs lost in bottomless pits, failed and undetected transactions in the transfer of control of jobs, tracing lost or forgotten jobs. Either match making must have an extra step at the CE level or we need to let the local sysadmins handle the mapping of job requirements. Both have serious drawbacks. [The feeling of the audience in a straw poll was for the latter option.] This is work in progress and choices must be made fairly soon. All input is most welcome. In the discussion, it was suggested that an expert in each individual batch systems should be put in contact with Francesco's team so that the latter could understand which attributes can be matched to which batch schemes and which cannot. Francesco welcomed this and after the meeting promised to continue this face-to-face dialogue as he had found it most useful.

Scientific Linux Update – Troy Dawson (FNAL)

Since the last meeting, usage of SL has hit almost 10,000 systems although SL3 growth seems to be levelling off. SL 3.0.5 and SL 4.1 are released as well as bug fix repository for each. Troy said that he appreciated input from other sites, in particular CERN. He would still like to organise an SL workshop, perhaps in conjunction with a future Fall HEPiX.. SLF (Fermilab SL) 3.0.5 for i386 and x86_64 and SLF 4.1 for i386 are released. FNAL users are being encouraged to move away from "unsupported distributions", which include RH 7.x, and these have been removed from the distribution servers. The SL team are working on 4.2 which Redhat released last week.

Linux Status at CERN – Andras Horvath

The phase out of Redhat 7.3 has almost completed apart from some which will stop at Christmas and a few isolated systems who cannot upgrade. While the recommended system is SCL 3.0.5, there is a beta test of SLC4 and certification is in progress but, as explained at the last meeting, this is only a safety net in case SL5 is too late for adoption for LCG final preparation and startup (decision late summer 2006). However, the SL4's Linux 2.6 kernel has attractions for desktop users and for storage hosts. SCL3 is

[5] Unfortunately, there is 1 production and 2 (!) test HEPiX home pages and Tim was persuaded to advertise his web page on the least popular of these! This will be fixed when the "official" new home page has passed a mini-stress test.

available for 64 bit support, both for Xeon/Opteron and for Itanium (ported by CERN) but CERN does not run a 64 bit server yet although all recently-installed servers are 64 bit-capable and the next generation of desktops to be ordered to CERN will also be 64 bit capable. Work is going on to understand the problems associated to a 64 bit service on multiple chip sets.

Linux Discussion

Why is there an SLC, what is different from SL? The answer from Jarek Polok via Andras Horvath is that most of the CERN tailoring are add-ons for AFS support (Kerberos 4 support is needed for example) and anyway both releases are binary compatible and if exceptions to this rule are found they will be looked at. Several members of the audience claimed that the binary compatibility message has not been properly passed and more advertising is needed. Michel Jouvin voiced fear stated by some users that SL and SLC will diverge. The Manchester Uni representative stated that some of his users say they will not use anything but SLC but he needs to support SL for FNAL experimentalists at his site. We will discuss these concerns with the CERN SLC team.

Marc Mengel suggested that we should join in some form the Linux Standard Base (LSB). It would enhance interoperability of our code, make Grid interworking and HEP-wide software publishing easier. Being LSB-compliant may assist in the migration of user applications from SL3 to SL4. On the con side, LSB is still evolving. There was some agreement that adding some of our applications to the LSB compliance suite would be desirable. One possible negative aspect is that being LSB compliant could be seen as replacing the current trend which is heavily towards a single platform based on the Redhat base, whether RHEL (SLAC and JLab) or SL/SLC (almost all of the other HEP sites, at least those represented at HEPiX). Perhaps we should pass the message to LCG management that they should consider the potential benefits, to themselves and to their users, of being LSB compliant. Meanwhile we should think about the advantages or not to join.

FNAL had originally said they would support SL 3.x for 3 years. Since it will soon be 3 years since first release, they are now wondering if they should extend it to 5 years. There was some support for an extension if FNAL feel they can resource it. Troy will seek permission to update the SL web site to promise support for 4 years.

Troy wondered if we cannot find someone to coordinate the SL applications repository but there were no volunteers in the audience. On the other hand, the audience agreed that it makes no sense at this time to do an Itanium port of SL4 but it could be reconsidered at a later date. It was further suggested that due to the large collection of 3.0.5 updates, FNAL should consider a consolidation into a 3.0.6 release and Troy will reflect on this but 4.2 will have priority (this was agreed). Sabah Salih of Manchester Uni offered to organise a series of monthly video meetings to discuss SL issues.

Light-Weight Disk Pool Manager – Gilbert Grosdidier

DPM's goal is to provide a replacement for a Classic SE for LCG Tier 2 sites, offering better manageability and integrated security à la Globus. Gilbert described the components and features. It has been installed in at least 18 sites. Installation can be done in one of 3 ways, one of which is manual and not easy but once installed it appears quite stable and it is being used on some sites for LCG SC3. Some missing functions have been reported (at both sysadmin and user level) and these are being worked through. He then presented an exhaustive list of release dates and short and medium term plans. He described how to manage a given DPM user file. Testing is a major activity, both for Oracle and MySQL backends, in order to improve its robustness and performance; it is estimated that testing took around 50% of the total development effort. It is hoped that a gLite production service can provide a plausible solution for small Tier 2 sites, migrating Classic SEs towards DPM SEs.

Panasas at the RCF in BNL – Robert Petkus

The main reasons for looking at Panasas are to offer a single, facility-wide namespace for data with POSIX-like, transparent access to the data. They have installed 100TB on 20 shelves over 10 Realms (like AFS cells). Current data transfer per shelf reaches 250MB/sec opening only 2 ports (4 are available per shelf). He then listed the main features and where these improve on those of NFS and Veritas. He also described other solutions they had looked at such as GPFS, Ibrix, Xrootd, dCache and Lustre. He finished by describing the configuration of the different blade servers in the Panasas box noting that each Storage Blade (up to 10 per box) has 2 SATA discs of 250GB or 400GB each. Remaining issues include the current lack of LDAP support, quotas only per volume, no failover if the Director Blade on a shelf dies[6] , some instability if a disc is almost full, possible memory leaks (a frustrating cause of a number of crashes) and a maximum of 350K files per directory[7]; they are working with Panasas on many of these. In summary, scalability and performance are a great improvement over NFS, it is reliable and easy to administer, and the supplier has competent and responsive support staff. Both BNL and JLab would recommend it but only after starting with a test setup.

BaBar Data Distribution using SRB – Jean-Yves Nief (CCIN2P3)

BaBar has produced so far more than 800TB of data and processing is spread across 6 Tier A and many Tier B centres. To ensure reliability, robustness and scalability, Source Resource Broker (SRB) from San Diego was chosen for data distribution. SRB is widely used in many sciences and provides a uniform interface to heterogeneous storage systems on multiple sites. Jean-Yves showed the steps to move data from HPSS at SLAC to HPSS at Lyon. SRB is running on SUN servers at both sites with an Oracle database on the SLAC side for the MCAT metadata catalogue (Lyon is only a client for now but an MCAT will be added soon). They have achieved transfers of up to 3TB per day, tape to tape. They are very happy with SRB, it has saved a lot of time in developing their applications and it is fully automated (automatic problem recovery in some instances for example). So far they have imported some 132TB in 232K files. The next step is to add the RAL Tier A as an SRB site.

SRB at CCIN2P3 – Jean-Yves Nief (CCIN2P3)

Outside the BaBar area, SRB is attractive for many projects, including as a collaborative tool to share files so CCIN2P3 have installed 2 SRB servers (currently) with clients across many operating systems and interfacing via the SRB API, Java APIs or a web interface. Although smaller scale than BaBar, the clients of this SRB instance span various sciences such as astrophysics and medical science. In total, SRB stores 140TB for the current projects with more coming.

Database Services for Physics – Radovan Chytracek (CERN)

Most of the support given by his team (CERN/IT/ADC) is for LHC experiments but they also support Compass and HARP. They based support on an Oracle 9 Sun cluster but since some performance problems have been seen they are examining load balancing and failover solutions such an Oracle 10g RAC/Linux cluster in coordination with other CERN/IT groups. They are setting up 4 small 2 node RAC clusters for the LHC experiments and other clusters for testing. From being a reactive support team, they are moving a more pro-active mode where they interact with application developers during the development phase, helping to optimise the application before production starts. They separate applications into those which can share a service and those resource-consuming ones which require their own service. There are 3 services, one for testing, one for integration and validation and a production service. They help developers in performance tuning, problem or bottleneck tracing and monitoring, They built monitoring sensors for Lemon (see later talk) in order to integrate their service into the central operations of the Computer Centre.

---

[6] JLAB, another HEP site currently using Panasas, intervened at this point and reported having updated the control software recently and failover seems to work now.

[7] Some people would say that setting a limit like this is a feature.

dCache at BNL – Z.Liu

dCache has been in production for US ATLAS since Nov 2004. The speaker enumerated the well-known advantages of dCache. At BNL they use a hybrid model for the read pool servers (332 nodes out of a total of 336 dCache servers) sharing resources with worker nodes so that each worker node serves both as a storage node and a compute node. They use the Oak Ridge Batch System as an optimised backend tape pre-stage system. There is a total of 100TB of production data compared to 123TB in the ATLAS archive in HPSS. dCache is used for ATLAS Grid production jobs and its use is being tested for LCG SC3. Peak transfer rates of 3.8TB per day were seen in July, mostly stores to tape, as part of SC3 testing. Current rates are around 1.8TB per day. Offsite clients access it via GRIDftp or SRM. While the hybrid model works well for read pool servers, write pool servers require dedicated systems with XFS rather than EXT3 file systems and reliable discs. They are generally happy with their setup although there was a learning curve for the sys admins. She finished with experiences from US ATLAS's experience with SC2, where they met the goal of 70-80MB/sec disc-to-disc CERN to BNL transfer, and SC3 so far where already they have achieved peaks of 150MB/sec sustained for a week. In SC3, 4 US Tier 2 sites are participating and achieving aggregated transfer rates of 30-40MB/sec.

LEMON Monitoring Update – Miroslav Siket (CERN)

There is a wide range of monitoring tools in the different HEP sites which means little chance of sharing the monitoring data or even building gateways or common alarms or inter-site problem tracking. Could Lemon fill this gap? There was a feeling expressed at the FZK HEPiX meeting that Lemon was still too CERN-centric so work has been done to make it more transportable. At CERN it is used in the Computer Centre (on over 2300 nodes) plus at 2 smaller installations. Other sites, mostly in Europe, are now evaluating it, see list on the overheads. Various platforms are supported but mainly Linux; various installation methods are possible including Quattor and Yaim. There is a wish list which is being worked through including improvements to the GUI, better web support, MySQL support (today only Oracle) and better security (a request from LCG sites). Today, security in terms of authentication and data encryption is rare in such tools but the next version of Lemon will include SSL support for authentication and the possibility of encryption and access restrictions. Other current work is to improve the CERN Computer Centre Alarm scheme, based today on an old service (SURE). The new scheme must be able to cope with an installation of over 10,000 nodes where one could imagine instances of 100s of simultaneous alarms. Thus LASER (the LHC Alarm Server project) is one possible option and IT are discussing this with the accelerator builders and a gateway has been developed for evaluation. Miroslav closed with the current enhancement programme and the new simplified installation procedure to install Lemon. After the session, several sites, including SLAC, expressed interest in evaluating Lemon.

OSG at SLAC – Matteo Melani

OSG grew out of Grid3 which itself grew from older grid projects. Grid3 had 30 sites with around 3600 nodes but it closed as of September 1st. Members of OSG include all the major US HEP sites, some universities, some older grid communities and representatives from all HEP experiments represented in the US (not ALICE or LHCb). Partners include LCG, EGEE and Teragrid. OSG has adopted a pragmatic approach where the experiments will drive the requirements and expansion of scale and complexity will be adiabatic. It is heterogeneous (all Linux but different flavours) but sites will have autonomy – no site will be forced to have grid software on compute nodes. He presented a slide of the OSG components, each taken from various other Grid projects. The production OSG consists today of 38 sites, 14 VOs and around 5000 CPUs. There is a smaller test grid with 24 sites and around 2400 CPUs. There is an Operations Centre and an incident response framework, the latter coordinated with EGEE. At SLAC there are 4 SUNs on the production OSG with 100 job slots and among the registered VOs are ATLAS and CMS and SLAC is considering becoming an LCG Tier 2 site. The SLAC Unix account scheme did

not match the OSG model so a new class of Unix accounts had to be created with minimum privileges (no email, no login). Their scheme of using unique Unix accounts for registered Grid users does not scale but they were forced into this by the lab's security policy. There was a race condition between the job gatekeeper and the LSF scheduler and the LSF job manager had to be partly rewritten. US CMS jobs ran without problem but US ATLAS jobs require access to remote databases and SLAC batch nodes do not have Internet access so they either need a local copy of the database or they need to create a tunnel to BNL or CERN.

XROOTD – Present and Future – Andy Hanushevsky

The application design point is 1000s of clients accessing the same data supporting small block, sparse random access to a large data set. Their performance tests showed that the Xrootd server was as good as or better than any other data server they tried (although he did not list which) and that performance was limited only by the hardware. He showed graphs of factors affecting performance. They have created an Xrootd service which they believe could scale to 256,000 nodes; SLAC runs a 1000 node test server cluster and BNL runs a 350 node production server cluster. The clusters use a minimal spanning tree algorithm to self-cluster which is remarkably fast. Today there is a simple agnostic mass storage support but there is a proposal from BNL and LBNL for SRM support. Turning to the future, "the next big thing", he described the Petacache project presented the previous day – ideally 30TB of memory at commodity prices.

XROOTD at CCIN2P3 – Jean-Yves Nief

Xrootd has run at CCIN2P3 since 2003, first for BaBar only but more recently for other experiments. There are 2 master servers and 29 slave Xrootd/Objectivity servers with 70TB of disc cache. CCIN2P3 are very happy with Xrootd because of its scalability, transparent access to data, flexibility of configuration and ease of operation. Nevertheless, he had compared it with dCache but he has been forbidden to reveal the results (at least at this time) – he said he had been accused of making an artificial test and thus biasing the results! For the LHC era, hundreds of disc servers will be needed to serve thousands of clients. We know Xrootd can do this. Are there others? Are there better ones?

Planning for LCG Emergencies – Dave Kelsey

Starting from the target of 99% uptime during LHC data taking, what are the security aspects of the T0-T1 network connections so that service is maintained during or after an incident, what are the effects on LHC data taking? A policy and procedures document has been produced, based largely on a similar document from the OSG. Incidents must be classified according to their impact and what should happen in each case. It was recently realised that such plans are needed also for other incidents, not just security. Many sites must have such plans, can we learn from them? Issues range from how to keep data flowing to how and when to inform the press. Dave then presented some overheads from Denise Heagerty on what actions emergency procedures could include. He ended by inviting all the HEPiX sites represented to submit their ideas, and even more so any plans they have, on this topic.

Fermigrid – Keith Chadwick

This can be considered a meta-grid of existing resources which today may be dedicated to a particular stakeholder. It should be used to share resources via VOs. Work is in progress to translate this requirement and the formal mandate of the facility into a working reality. Keith listed some of the facilities, clusters, storage and networking which will form the core of Fermigrid. Some of the required software components are also already in place, often adopted or adapted from existing grids (VOMS, Grid User Mapping, Myproxy. A major (psychological)) obstacle is the recalcitrance of some groups to share "their" resources with all or certain other groups but the resistance is being gradually overcome (see the interesting table in the overheads of who is or is not allowed to access whose resources). There is a working interface to OSG and jobs flow in both directions. From their initial experiences they

realise that more documentations is needed, as is more user education but there are also areas where the support staff need to learn more, for example about supporting multiple VOs with large populations and a number of grid security lessons.

Evolving Local Resources Towards the Grid – Fraser Spiers

This was a summary by a Tier 2 site, and the Scotgrid coordinator, on the difficulties of acquiring or making locally-purchased equipment available to the wider grid community. Like the Fermilab speaker before him, he touched on how difficult it can be to persuade resource "owners" that they had more to gain in sharing. He also had to persuade local users who are members of a VO that it makes more sense to submit via the VO to the "grid" than to submit to the local batch system.

The GRIF Project – Michel Jouvin

GRIF is a Tier 2 project for the Paris region. It should have 15K SI2K units by the end of 2007, with 350TB of disc, the sites inter-linked by a 10Gb backbone and linked to the Lyon Tier 1 by a 1Gb link. There are 4 local IN2P3 sites and one from CEA and up to 20 people are involved but mainly part time. The intention is to appear to LCG as a single Tier 2 site. There is a plan for a modest budget but it is not yet agreed. Today, the LAL LCG/EGEE Grid site has been reborn as the initial GRIF site and the other 4 sites have either identified or ordered resources to join. Quattor is being used for site customisation to ensure inter-site consistency and DPM is being evaluated. They are participating in LCG SC3; so far they have sustained 35MB/s over 4 days in the throughput phase and they will participate in the service phase from November. Next year they hope to have 20% of the final configuration and they will need to make decisions on CE/SE split, which batch scheduler to use (looking at SGE and LSF), which monitoring tool (will look at Lemon as well as others), etc.

Day 5 – Hardware and Infrastructure

A half day of sessions on the second theme topic of the week.

Testing High Performance Tape Drives at CERN – Hugo Cacote

From 2007, CERN will need to store some 15PB of LHC experiment data per year and we need to choose the tape hardware rather carefully. Currently, CASTOR has some 42PB of data in its store. Hugo described the tape models currently installed and under test noting that next year we should move to drives offering more than 100MB/s transfer speeds. The tests include exercising all the SCSI commands, tests on the mechanics via repeated (125,000 times) mount/dismount cycles, various performance tests and the validity of ANSI labels. The drives have been integrated into a production CASTOR system to check them in production mode. This effectively performs checks on the operations of the drives, error reporting, usage statistics, etc.

Computer Procurement at CERN – Helge Meinhard

Helge started by noting the broad outlines of the CERN purchasing rules and the constraints these impose on us. Partly because our tenders are open to relatively small suppliers, we insist on sample systems for evaluation along with the bid for all but the smallest procurements. We also include a 3 year onsite warranty for hardware purchases, executed either by the supplier or by an approved third-party. He explained why and how we insist on a 5% bank guarantee and explained that the contract includes penalties for late deliveries. We perform acceptance tests and during this period, and during the 3 year warranty, there are clauses for hardware replacement and in the limit for the cancellation of the contract. Helge described the purchase procedure step by step. One conflict in the purchase process is the desire to restrict the number of onsite suppliers balanced against the requirement to follow market trends with

respect to technology and pricing trends for successive purchases. He gave some typical configurations of recent purchases for CPU and disc servers. Experience shows us that, after eliminating the smallest firms at the market tender stage, box integrators normally win against so-called Tier 1 suppliers, probably because they take more care to supply exactly what we request while the larger firms offer us the nearest system in their range to our specifications. Stress tests are a very important part of the procedure, using procedures "discovered" at previous HEPiX meetings from FNAL and SLAC. Finally he showed a chart of power measurements which had been made by Andras Horvath. In the questions, John Gordon said that their external advisors have encouraged them to concentrate on Tier 1 suppliers, to buy one major purchase per year, scheduling deliveries over the year and basing the price on list price minus a percentage. In this way they hope to get a better bulk price.

CPU Benchmarking at GridKA – Manfred Alef

They use SPECINT benchmarks as being universally accepted. However, even on the SPECINT web pages there is often a range of numbers for a given chip, the difference surely coming from the chosen environment, O/S, compiler and level and choice of optimisation. Therefore it is important that for testing you select the most appropriate environment for the target applications, in this case Scientific Linux with gcc 3.4.3 and a particular choice of optimisation switches. The next problem is that SPECINT results are usually for a processor but HEP sites nowadays typically buy multi-CPU systems and now often dual-core. For this reason, GridKA runs several copies (2 or 4 respectively) of the tests in parallel, He then presented the chips under test and detailed sets of results; anyone interested in hardware performance is seriously encouraged to check out the overheads. From his tests he suggests that 64 bit mode gives a noted performance boost but hyper-threading seems to actually decrease overall performance (when running 4 tests in parallel), probably because of memory contention; multi-core systems should be configured with sufficient memory. Turning to power consumption and cooling, he noted that many sites reports issues and/or problems in this area. He estimates that the cost of 3 years of power for the systems is equivalent to about half the purchase cost, at least in Germany. Again he showed some charts per CPU type.

Accounting in LCG – John Gordon

John described the accounting (or metering) scheme for LCG called APEL, how the data is collected and how and where it is consolidated and published. For LCG they report on various flavours of PBS, LSF, SGE and Condor. APEL can also accept aggregated data from a site, Lyon is an example where its batch system, BQS has its own accounting scheme. The results are reported to the LCG and GridPP management to show how the sites are performing against commitments. APEL builds data at a site, it is middleware independent and it offers many views of the data. Today it only measures CPU use, is this enough? Other open questions include how to describe the accounting data, how important is it, are there better or more flexible way to collect it, especially across multiple grids. There is an issue about publishing information which allows to identify individuals since data privacy prohibits this in some countries. Work is in progress among the main HEP grid players to address these issues.

Tree SLAC Incident – Chuck Boeheim

SLAC's main power feed was taken out one Wednesday morning in May by a tree falling on 2 power lines. Initially it took a significant time (many hours) to appreciate the extent of the damage and how long it would take to restore. He described the steps taken to complete the monthly payroll run which had been interrupted mid-way through and to open a temporary web page to carry the latest news. Planning for the eventual recovery was done via external mail from home and by GSM. This has led to better documentation on how to get in contact with key personnel. They are also considering how and where to establish alternative emergency offsite web and possibly e-mail services. E-mail would be particularly tricky in terms of reconciling an emergency service with the production service on recovery. [Power was eventually restored some 50 hours later and recovery took all weekend to complete.]

CERN's Computer Centre – Tony Cass

Tony described the where, why and who of Buildings 513 and 613. In the former there are 2 rooms of 1500 and 1200 m^2 with 1.5kW/m^2 upstairs and 500W/m^2 downstairs. The services go beyond physics and some of them are deemed essential to maintain in operation. It is expected that by 2008 we will have 2500-3000 boxes for physics (25,000K SI2K, doubling by 2010; 6,800 TB of disk data in 1200-1500 boxes, almost doubling by 2010; 15PB of tape data per year needing 30,000 500GB cartridges and 5 6000 slot new robots per year. Among the issues are

- Multiple vendors, even among equipment of the same type
- The operation of such large scale equipment
- Cooling capacity and power backup, at least for the vital systems since we could not afford to provide backup for all of it

For system admin CERN built a full set of tools known collectively as ELFms (Extremely Large Facility management system) which includes the Quattor, Lemon and Leaf tools[8]. He showed screenshots of some of the interesting output displays of Lemon.

Alan Silverman
15th October 2005

[8] The first 2 have been covered earlier and in previous HEPiX meetings; Leaf manages state changes.