HEPiX Spring 2004 Trip Report

National E-Science Centre, Edinburgh

24-28 May 2004

Introduction

The meeting was hosted by NESC who provided a considerable infrastructure plus the coffees, lunches and the mid-week dinner at no cost (no registration fee). It was attended by over 100 delegates (a new HEPiX record) although not all attended all week. Dave Berry, the Research Manager of the Centre described some of NESC's activities. UK e-Science programme is funded to the tune of £213M, distributed across various UK research councils of which PPARC has 27%, all for staff costs, grid resources are funded separately. Much of this is spent in dedicated centres such as NESC. His presentation (which is on the web, as are almost all of the sessions during the week) gives more details of the funding and where it is spent. NESC's specific role is to help coordinate and lead e-Science in the UK as well as to manage the E-Science Institute which hosts a number of research visitors and events, training sessions and workshops (hence the justification for holding HEPiX here). NESC also has a number of industrial partners including some of our favourite suppliers such as IBM, Oracle, Microsoft, HP, SUN, Apple, etc. There are some 35 projects underway at NESC, many focusing on data management (as demonstrated later in the week during the Mass Storage Workshop).

Summary

As usual, this report is probably too long to be read by all but devoted HEPiX people so I will make a summary of the high points as I saw them. Also as usual all the overheads (without exception I think this time) are on the web, thanks to Dave Kelsey who insisted that the dedicated portable be used to show the talk, thus ensuring he had a copy for the web.

- First and foremost, it was a well-organised meeting, disproving the theory that there is no free lunch - there was no registration fee and we not only got 5 free lunches but free coffees and a free conference dinner as well. It is good to see that this was accompanied by a record attendance of just over 100.
- Redhat: certainly the theme of the week. On Monday, the Redhat Sales Director responsible for the contract with the US DoE Labs gave an overview of the Redhat offer for **all HEP sites** worldwide, effectively confirming the offer we had negotiated between CERN and Redhat, with help from SLAC. Among the audience there was some measure of acceptance of both the principle of paying a licence fee and the prices on offer. But there was also a request for Redhat to consider more seriously a site licence scheme, rather than the cluster scheme they are offering, and there remains the question of whether we will get value for money. SLAC have been using the licensed version and a Technical Account Manager for only a month; early signs are positive but it is too soon to judge.
- On Tuesday there were presentations by Fermilab and CERN of their schemes based on rebuilding sources, including an open invitation by Fermilab for other sites to participate in their "scientific Linux" initiative.
- On Wednesday I chaired an interactive panel on the subject which demonstrated a very positive and constructive willingness to work together. Under the prompting of Les Robertson, I summarised that it appeared that Fermilab and CERN's policies are not very different and that perhaps could be brought even closer. Happily Jan Iven and Mark Kaletka of FNAL agree and by

the end of the week they had prepared and distributed for comment a draft resolution for a common scientific Linux, except for a very few essential differences. And because this originates from the Redhat-provided GPL source, it should be binary compatible with the SLAC distribution, acknowledged by the BaBar representative present.

- There was an interesting summary of the very successful AFS workshop at SLAC in March.
- In recognition of the apparent lack of serious interest by Windows support teams, Dave Kelsey proposed that we drop the HEPNT part of the name of the meeting and revert to HEPiX alone. Windows talks would be still very welcome and encouraged.
- The last day and a half were devoted to Mass Storage issues, organised by Olof Barring. Many of the talks were interesting but possibly the most important session was a two-part panel on establishing an agreed plan for service challenges for LCG. Most LCG Tier 1 sites were present and many represented by decision makers. By the end there was a good understand by all parties of the tests to be performed, the timetable (starting in the coming weeks), which sites will participate and what resources are required. Les would need to confirm but I believe the LCG fully achieved their objectives in staging this meeting in the context of HEPiX.

The next HEPiX meetings are being scheduled for BNL for the week of October 18th and Karlsruhe (FZK) next spring, probably in May, either the week of the 9th or 23rd. In BNL we will try to stage Large System SIG sessions on "platform technology" – MacOS, Itanium, AMD versus Intel.

Alan Silverman
30 May 2004

Site Reports

BNL (Tony Chan) – after a quiet year in 2003, RHIC has taken off again this year with results surprising all expectations and this has been matched with 4 new hires for a total of 29 staff. 4 HPSS STK silos with 37 9940B drives and 4.5 PB capacity. There is a 41 node SUN-hosted NFS-based SAN disc farm for users files, aggregating 600 MBps on average. In their AFS cell, also for user home directories, they are still running Transarc AFS on the server side and OpenAFS on the client side but looking at moving to an entirely OpenAFS service. After a new acquisition of some 250 servers, the RACF (RHIC and Atlas Compute Farm) is now 1359 dual rack mounted batch Intel servers. Looking to replace the current batch system (PBS?) with Condor; they are also using LSF but this may also be replaced by Condor if the first exercise is successful. Peak usage today is by ATLAS, not RHIC. They use Ganglia for monitoring. BNL Security was covered in previous meetings (e.g. two layers of firewall and limited interactive use through secure gateways; they are moving to Kerberos 5 single sign-on. Also looking to move away from NFS-based SAN to a commercial solution as well as to dCache for distributed storage management.

PSI – the Villigen lab are not often present but Juergen Baschnagel gave an update on their activities. They use Redhat 9.0 but are unhappy with their most recent offer from Redhat, hence their presence at this meeting. Their OpenAFS cell is hosted on IBM X series 345 nodes and they will move to Kerberos 5 authentication (Heimdal) in Q3.

SLAC (Chuck Boeheim) – BaBar is achieving record luminosities; and data is being reprocessed into root format. Two other projects (accelerator physics and the KAVLI experiment) are getting interested in MPI clusters. All this adds up to significant demand. Online storage now 350TB. They are using Redhat's RHEL 3.0 under their bulk licence scheme, the WS edition, on most serves and desktops. They hold regyular weekly service meetings with their Redhat TAM (Technical Account Manager) who they share with Livermore; it looks productive, opening and subsequently working on some 20 issues. All this is an evaluation, as CERN is planning. SLAC does not use the Redhat satellite service, relying on mirroring their update repository, using a script they jointly developed. They have not yet decided how to modify offsite clients but they have developed a desktop installation scheme (ALDI). Now running Solaris 2.9 on most systems, beta testing version 10. No support for MacOS X but the first requests are coming in for this, inclusding some for file serving; the cost of this support is estimated at 3-5 FTE. [Similar pressure was reported to me by FNAL managers, but again with no support being offered at this stage.] SLAC is currently negotiating with Platform to improve their licence terms, including getting them to extend these to world-wide HEP. Minor changes (compared to the above at least) include

- migrated to Citrix XPe
- finishing rolling out Exchange 2003
- using a home-built web site (XWeb) to allow users to manage and track Windows installations
- switched to Symantic anti-virus
- and migrated whole lab to Office XP.

DESY (Stephan Wiesand) – DESY is consolidating on Windows XP, SPARC/Solaris 9 and Linux; other platforms being removed or already gone. There are a few other unsupported platforms, including MacOS X. They are currently rolling out SuSE 8.2, but its support life ends April 2004 and patch availability before the end of the year In fact this is likely to be the last DESY Linux based on SuSE, they are now looking at Redhat, depending on its pricing or basing themselves on the sources *a la* FNAL. First analysis of the FNAL so-called "scientific Linux" is rather favourable, it is largely compatible with their current SuSE distribution. DESY is also looking at Linux support of AMD 64 chips. They are participating now in their first grid projects, EGEE and D-Grid with a testbed in Hamburg running Redhat 7.3 systems. Now performing regular security scans, forcing rules for individually-maintained systems and forcing automatic patch updates. During the sasser alert, they manually checked all portables being brought on site for several days. Their firewall forbids outgoing smtp except for approved mail servers. Evaluating Ximian mail connector to work with Exchange but some problems remain to be solved. Developing a set of Computing Rules which users will have to sign.

LAL (Michel Jouvin) – lots of emphasis on mail service, including anti-spam measures. The project to roll out an Active Domain forest across the IN2P3 sites is late but now underway. Getting involved with EGEE and getting 2 funded FTEs. Planning expansion of their Linux capacity with the addition of 25 dual Opteron servers (on order).

FNAL (Mark Kaletka) – head of Core Support Services. Major trends are a steady growth in production capacity and moving to 1U systems; they expect to reach some 4500 systems by FY07. Need to provide more space, power and cooling; so they need to re-use some available space and renovating another building to cope. Some 1.7PB in online tape libraries, transferring 10s of TB/day for analysis. Installing dark fibre to Starlight in Chicago. Mandatory node registration (including a mandatory security scan) and deploying VPN. AFS cell now completely OpenAFS. D0 reprocessed 600M events in fall 2003 and rolled out a Monte Carlo production form with grid-style job submission based on SAM. US CMS is continuing to build CPU and disc capacity.

JLAB (Kevin Edwards) – recently undergone a DoE security audit; reported to be ok when attacked from the outside but not so good from the inside. Completed upgrade to Solaris 8; may skip V9. Gone with RHEL V3 licensed from Redhat, installing a satellite server. Found there is no support for XFS but it does support a Linux 2.6 kernel if this is preferred. Adding new 9940B tape units. Hardware problems on their file servers, put down to heavy use; so evaluating StorageTek B280 system. Added 24 systems to their batch system, all running RHEL 3 WS with the Redhat-supplied 2.4 kernel. Big push on the pentaquark search (hints of which are appearing) and thus pressure on resources. Also a fourth experimental hall is almost approved which will increase load. Developing standard Windows builds but held up by licence negotiations; for email upgrade, looking at Oracle Collaboration Suite, MS Exchange, Cyrus and open source solutions. The first was almost impossible to install and get working and was dropped; no decision yet on the others.

CASPUR (Andrei Maslennikov) – CASPUR operates their usual collection of SMP servers and clusters, from IBM, HP and a NEC vector system; includes now an Itanium-2 SMP server. Running Heimdal KDC on 3 of their 10 AFS cells without problems. Experimenting with Heimdal single sign-on between AFS and Windows and moving to production based on this, although probably with the MIT variant of authentication. Using the Clavister Firewall and very impressed by it. They have established an Internal Exchange Point with 20 major customers.

NERSC (Cary Whitney) – as everywhere, more emphasis on security in the sense of being more pro-active, closing loopholes and increasing communications with other sites. More work on their home-written network analyser, which has gained DoE funding.

CERN (Helge Meinhard) – Helge described the new CERN hierarchy and that of the new IT Department. He then covered major IT changes such as AFS password expiration, the banning of off-site ftp and the large effort expended on security. He described

- the deployment of quattor and Lemon
- the work on LEAF state management system
- the status of the computer centre system administration in-sourcing
- the refurbishment of the computer centre
- confirmation that CERN will adopt the SLAC (Chuck Boeheim's) scheme for handling serial consoles; also following SLAC, CERN has developed a stress test scheme for new acquisitions
- the coming evaluation of Oracle 10g and its main features of interest to CERN
- LCG deployment
- The completion of mail migration to Exchange and the deployment of the Windows Terminal Service
- The internet land speed record
- The introduction of the new GSM operator, the imminent ending of ACB and the forced registration of portable PCs.

GSI (Walter Schoen) – he described the FAIR proposal which would enlarge GSI with ion and anti-proton work and hence increase the scope and load of the IT support at GSI. Meanwhile, GSI continues to run 350 Debian Linux boxes along with a few SuSE systems for special purposes and 800 Windows nodes. Like several labs reporting, they are installing disc servers with serial ATA (SATA) Raid discs. They will finally phase out Windows NT, moving to XP on clients and Windows 2003 server. They are discussing groupware (2 candidates, Exchange and Oracle Collaboration tools) and a new batch system (several options, the present one, LSF, or an open source product such as OpenPBS).

TRIUMF (Corrie Kost) – Proposing 3 flavours of Linux – Redhat 9, Fedora Core 1 and "scientific Linux" (the one distributed from FNAL). Taking a very active part in WestGrid (a Canadian project centred on UBC and TRIUMF) and they host a 504 dual node cluster for this. Suffered lots of disc failures, which are the prime cause of a failure rate of about 1 node per day. Participating in LCG in a minor way. Again more SATA drives reported here.

Manchester – a small site today but interesting for their plans to build a farm of Apple G5 nodes as well as a major expansion in the light of LCG participation.

RAL (Andrew Sansum) – Assured funding from GRIDPP2 out to 2007 to host a Tier 1 service for LCG. Current hardware has not changed recently (350 dual CPU systems and "HEP-standard" disc and tape services). But 256 more dual Xeons coming (looked at Opteron but unimpressed by price-performance). Also awaiting another 140TB of disc capacity. Andrew feels it is now time for RAL to explore other storage architectures as limitations are beginning to appear on the horizon. RAL is performing Grid operations for LCG but also for many other projects. There is a new Opteron cluster planned for other scientific services (such as ISIS).

Redhat Linux

Nathan Jones, Redhat's sales director responsible for DoE labs, described the Redhat Linux offering and strategy for HEP labs. He described both Fedora and RHEL (Enterprise Linux) and which market each was targeted at. For RHEL, hardware support (for approved hardware suppliers) is guaranteed for 2.5 years; software support for 5 years. RHEL is available for Intel 32 ands 64 bit, AMD 32 and 64 bit, Itanium and IBM i, p and z series. The itanium price is about 1.3 times that of Intel 32 but he expects it to come down. Lots of Linux 2.6 kernel features are included already in RHEL 3 but not all, the rest will come in RHEL 4. Redhat recently announced a Desktop package but it is limited to 1 CPU systems whereas WS and ES support up to 2 CPUs and AS supports more.

He confirmed that the price offer for CERN would be applied to HEP sites world-wide and offered to intervene on demand if a local subsidiary was unaware of this. Redhat also offer a scheme to ease management of the licences: there are two variants, a satellite server and a proxy server. The first costs $13K for the installation at the user site plus about $20 per connected node (which thus receives the kits and updates as scheduled by the local administrator). In the second case (more expensive), the licences are stored on a Redhat site and the local proxy server access them from there as scheduled. JLab have signed for a satellite server, SLAC have neither and handle the distribution themselves. In his experience, satellite servers are more commonly used for managing desktops than clusters. He agreed that for large configurations the overall price could be negotiated.

He noted that the above prices are for annual "subscriptions" and not licences because of the GPL nature of the software. He said Redhat are not planning to audit sites but expect them to be honest, take as many subscriptions as needed and not distribute these anonymously to other sites. He also confirmed that he understood that we need to modify some modules. This would not in principle nullify the support although Redhat could not be expected to support local modifications nor defects caused by such modifications.

Day 2

ELFms

German Cancio described the management system for Extremely Large Farms in the CERN Computer Centre. He described in some detail the three sub-components, the node configuration tool quattor, the Lemon monitoring tool and LEAF for high level management tasks based on the two previous tools.

Today, ELFms manages almost all of the 2100 nodes in the Computer Centre and has been ported for use with Solaris 9, now under certification. EGEE has decided to use quattor and some LCG Tier 1 and 2 sites are evaluating it, as well as some LHC experiments. More information at http://cern.ch/elfms.

Lemon
Miroslav Siket described the web interface to Lemon in much more detail.

CDF Hardware Management
Paul Miller of Glasgow University spoke on experiences gained in trying to manage a small CDF analysis farm in Glasgow. They experienced a 2 month down time on a RAID cluster and the service was not staffed to manage system administration which contributed to the length of the down time. He demonstrated a formula to justify a certain team size for managing clusters. But at very small cluster sizes there still needs to be a minimum number of system administrators which he estimated at 0.5 FTE even for the smallest cluster. He listed some lessons learned from the experience.

LCG User Registration
Maria Dimou talked about user registration and Virtual Organisation management. She explained why this topic is so important and how it is done today, both from the point of view of the user and behind the scenes, including authorisation from the relevant VO manager for the new account and storage of the account information in one or more VO databases. After a short delay to generate and distribute the updated grid-map file, the user can then generate a proxy certificate and submit jobs to the grid. A new scheme is being evaluated to optimise this process and avoid errors by making use of available existing account information. There will also be expiry dates and checks on any security incidents associated with the requested account. A task force has been created to investigate this topic in collaboration with similar activities in the US.

LCG Test Suites
Michel Jouvin presented these slides on behalf of Gilbert Grosdidier. Various tests were described ranging from installation and configuration testing to tests of functionality and stress testing.

LCG Project Status
Oliver Keeble described the current status of LCG. He presented the timeline up to first LHC data taking in 2007 and highlighted the major LCG milestones along this timeline including the migration from LCG-1 to LCG-2 and the various data challenges. He gave more detail on LCG-2 and its deployment. There are currently 53 active sites (as of today) with over 3300 CPUs but it is changing (upwards) on an almost daily basis. He continued with some of the lessons learned and some plans for the near-term future and closed with a brief overview of EGEE and its interface to LCG.

GridPP Status
Tony Doyle of Glasgow University presented this project and updates on its status. This is part of the UK contribution of LCG. It is a collaboration of 19 UK universities, plus RAL and Daresbury labs. GridPP1 is coming to the end of its life and will be superseded by GridPP2 from 2005 onwards. It covers UK Tier 1 and Tier 2 sites and links to LCG as Tier 0 but also links to US experiments (BaBar at SLAC and D0 and CDF at FNAL) and to other sciences. He showed the financial breakdown and internal organisation with its collection of the usual boards associated with all grid projects. GridPP1 contributed greatly to EDG and this will continue with GridPP2/EGEE collaborations. As of April 2004 there were some 700 dual CPUs, 80TB of disc and 60 TB of tape capacity available at RAL from GridPP1 and 0.5PB and some 2000 CPUs UK-wide.

He described the coming challenges as we build towards LHC startup and the UK's promised contribution to this, the scale of the resources needed and how to identify and access individual interesting events distributed across the grid.

Fermilab Linux

Connie Sieh presented the recent Fermilab Linux release. They have called it Scientific Linux; it is built from Redhat RHEL 3 AS sources recompiled at FNAL with some added packages from Redhat and local modules such as openafs, yum and so on. It also includes features which permit local site customisations. They are proposing an open-source development where anyone can contribute but she noted that her task was related to Fermilab Linux support, implying that there is no guaranteed support from Fermilab for this release for other labs. In particular they are looking for someone to supply missing RPMs such as those for Root and the various grid tools. They also lack for the moment support for AMD, Itanium and Intel64 chips. She invites people to use it, join the mailing list and help make it better. She had brought with her some free samples of CD kits ready for installation.

Using YUM to Distribute Linux

Connie continued by describing how she uses yum to distribute the Linux releases. She compared the different choices available such as autorpm, Debian's apt, etc. She covered the main features of yum, its origins and its features.

CERN Linux

Jarek Polok then described the CERN Linux release. Also recompiled from Redhat RHEL sources in a very similar process as Fermilab described (could have been the same except for timing constraints on the support teams). CERN has also compiled a release for Itanium and the CERN additions include not only openafs (but with Kerberos 4 while Fermilab use Kerberos 5) but also the xfs file system. A 2.6 kernel is ready for testing but openafs support is missing for the moment. Unlike Fermilab, CERN base their distribution mechanism on apt.

He also presented work he is doing on a Linux Printer Wizard; while xprint relies on lprng, his wizard is based on CUPS (the default print protocol in Redhat Linux) and has a GUI similar to that of the Windows Printer Wizard.

Regarding the different Linux releases, they should all be binary compatible with respect to applications since they are all based on the same sources. If someone finds an exception to this, he invites them to inform him. The actual CERN release should be fully certified in a few weeks.

Fermilab Computing Infrastructure

Stephen Timms described Fermilab's current computer facility and how space, available power and, especially in the summer, cooling capacity are approaching limits. Plus new acquisitions are on the way and will continue for several years. Upgrading the present Feynman Computer Centre would result in a long down time. Hence the need for a new facility, to be known as the Fermilab High Density Computing Facility which should host some 3000 systems over the next 5 years. The new building will only have enough UPS for a graceful shutdown and no generator backup (unlike the present Feynman building). Thus it should be dedicated to high density compute nodes along with a few tape silos (probably with their own generator backup). Will be lights-out since it will be several kilometres from the existing computer building and with no offices. Construction began in May and the first computer racks should be installed later this year (Nov or Dec).

For the power and console infrastructure, they will use the Cyclades Alterpath series of console servers, switches and power strips. Each controller box has 48 ports and there are many interesting features. He will also use NPACI ROCKS, see previous HEPiX reports.

CVS at CERN

Manuel Guijarro explained the history and development of the two CVS services we have set up, one dedicated to LCG and a more general one hosting 71 software projects; the second one is using AFS to store the repositories. Both services provide data integrity (by regular mirroring), various access methods and various tools for administrators such as automatic detection of CVS locks. The architecture of the general service offers automatic fail-over because the files are stored in AFS rather than attached to a particular host. Since the LCG service was specifically requested as not wishing to depend on AFS, performance under normal conditions may be better but failover is not automatic and must be done by hand including renaming of the DNS pseudonym of the service. Also the need for more frequent mirroring swallows up some of the performance advantage coming from the use of local discs for storing the repositories. Will look in the future at moving to Kerberos 5 and somehow automating the failover also of the LCG service if possible. We also need tools to detect inactive projects to avoid the build-up of obsolete deadwood.

UNIX Application Software at DESY

Stephan Wiesand described how DESY handle distribution of the application software on UNIX. The previous scheme depended on symlinks which is unsuitable in some circumstances (portables for example). They have moved to RPM packages and automated tools for installing these, either on local file systems or AFS as desired, with options for control of individual packages and nodes.

Citrix at SLAC

Brian Scott presented a review of the various remote access methods to SLAC including limited VPN, Citrix or RDP (remote desktop protocol) and the features and restrictions on each, some of which are technical restrictions and others forced by security concerns. Outlook web access is also available for e-mail. The Citrix service has recently migrated from version 1.8 to XPe. There are some 900 accounts on this service. The Citrix farm is based on 128 bit SSL encryption. He described the failover nature of the farm. Different departments can buy and have installed their own servers. One silo gives access to the usual Office and desktop applications and a second to more specialised and/or licensed software.

Security Survey

Bob Cowles of SLAC performed his now-traditional survey of recent hacks and attacks. It was as scary and as entertaining as usual. SLAC's recent conversion to Active Directory and SUS meant that all SLAC systems (except some visitors') were patched within 80 hours of the most recent Microsoft patch release and thus they were not affected by the resulting attack 2 weeks later. He is impressed by an early look at the new firewall coming in XP SP2 (due Q4 but not expected until 2005). As usual he covered Windows weaknesses followed by some in UNIX and Linux, then Cisco, MAC and a few others and he confirmed that many attacks are becoming more "professional". He also covered the recent (the day before) CVS vulnerability which hit both SLAC and CERN the previous day and several other sites over the following days.

Day 3

BQS Update

Yves Fouilhe presented some modifications being implemented for BQS, their local batch scheme, updating the architecture in order to have better scalability, to handle quasi-interactive jobs for biologists and to offer more control for operation and administration. BQS now supports parallel jobs, arborescent job sets (that is, jobs with inter-dependencies) and jobs submitted from the grid, eg LCG-2.

PDSF Activities

Cary Whitney described some recent work at PDSF. This includes

- A framework to permit multiple Linux distributions to co-exist concurrently on a single node (CHOS or chroot OS)
- A process to associate Distinguished Names to processes running in a grid environment (ProcDN)
- One-Wire serial interface for hardware monitoring (temperature, power management and remote system resets for example)
- St.Michael and Patchfinder to perform kernel integrity checks and a rootkit checker respectively
- Using a Linux virtual server to mirror a BNL mysql database
- Consolidating a variety of event monitoring tools used by the operations team into a single user interface
- Comparing LSF and SGE (Sun Grid Engine) as batch schedulers; SGE (Enterprise edition) looks like a good alternative to LSF in their opinion
- Investigations on Lustre, described as a mixture of Linux and cluster: mixed results and unfavourable price conditions but hints of improvements on both horizons
- Improvements on Linux accounting and auditing

Many of the tools and procedures are, or should soon be, freely available; check the PDSF web site.

CERN Solaris Update

Manuel Guijarro reviewed the current state of Solaris at CERN. Quattor was ported to Solaris by a SUN-funded visitor and this will be used to replace ASIS and SUE. Unfortunately the amount of work engendered by this change (e.g. translating SUE features into NCM components) has delayed the certification of Solaris 9 but this should start very soon now. In combination with this they are developing a GUI to allow desktop users to prepare their own installation procedure (quattor was initially targeted at computer centre systems). Also porting Lemon to Solaris with a view to using this to monitor Solaris nodes which today are monitored by a locally-written tool.

Tests have been carried on a SUN Blade 16 node server and in particular the N1 management system which comes with it. Unfortunately the most interesting part of this, the N1 Provisioning Server, which could be an alternative to quattor was not a success – many bugs in the early release, requires dedicated extra resources required, not usable outside the Blade. Another test on the Blade was for use with the Oracle Application Server but we did not have the correct version. Finally decided that if we could not make this useful without contracting specialised SUN consultants (which is what SUN proposed) we are not interested in it. However, we are interested in the N1 *Service* Provisioning Server which is a different product designed to manage objects across a loose cluster of systems but we are (still) waiting for SUN to deliver something we can test.

Redhat Linux Discussion

Alan Silverman led a discussion on various issues around the different Linux releases from CERN, FNAL and SLAC. First each of those sites reported on its current status and future plans. These were described earlier in the week. SLAC start from binaries using the paid subscription model – note, it is formally a subscription, *not* a licence fee; Fermilab start from the open-source sources; CERN also start from source for the coming distribution but are buying 200 subscriptions from Redhat and will evaluate both solutions for price/performance before deciding for next year's distribution.

Among the points made in the discussion were

- Fermilab emphasised that their release is available world-wide but with no promise of support; they had included a site customisation module with documentation
- CERN will try to measure value for money when decided what to certify next year, comparing FTE costs of an in-house solution with (hopefully) reduced FTE costs plus subscriptions fees for a Redhat binary-based distribution
- CERN will support the coming, free distribution, until the end of 2005, again only promising full support inside the lab.
- SLAC estimate their costs this year starting from the Redhat binaries to be less than 1 FTE and noted that there will need to be negotiation for next year's model
- Some people considered that Redhat's pricing model does not take into account the true value of having large clusters built from identical hardware
- We should encourage Redhat to introduce a model based price per incident.

The session ended with a plea to the various providers to get together and work out a scheme to make a HEP release of Linux which all sites could make use of, perhaps with some straightforward local tailoring, and which could be proposed to the experiments participating in the major grid projects. A straw poll of the audience revealed no objection to the subscription price being proposed by Redhat to all HEP sites, as long as experience is positive with respect to the value for money of these subscriptions and the various support options.

Anti-Spam at LAL

Michel Jouvin described anti-spam measures taken at LAL recently. Sendmail is used on the message router with mail stored on an IMAP server. All mail clients can perform some level of filtering. Spam filtering is performed at server level according to a local policy decision but the user should be given the choice of whether to delete or deliver spam. Rather than a rigorous anti-virus campaign at the server level, LAL ban certain file extensions using MIMEdefang, passing the message on with an appropriate warning replacing the forbidden attachment. MIMEdefang calls Spamassassin to scan for spam. This scanning involved rules-based checks, Bayesian analysis and black and grey lists. Spamassassin has been tuned and the various methods have resulted in a success rate of 90% in identifying spam.

Arc version 2

Wolfgang Friebel described a replacement for Rainer Toebbicke's arc programme to submit batch applications needing AFS tokens. The main tasks were to replace Kerberos 4 by version 5 and to solve some load problems. The various features and desired functionality were listed.

AFS Best Practices Workshop

Alf Wachsmann summarised the results of this workshop, held at SLAC in March. It included a tutorial on best practices and some invited speakers. It had been much more popular than expected with some 100 attendees including a few from Europe and a few commercial sites. About 70 signed for the tutorial. There were excellent quality talks about AFS in many different environments (including at Intel and at Morgan Stanley); and interesting talks on some performance issues. There was a description of the new openafs for Windows. Discussions included the possibility of one of the principle maintainers creating Openafs Inc. to support it. Another maintainer is unemployed and has created a method for people to pay him for assistance. But the open-source nature seems not to be in danger. Alf suggested that the larger HEP sites get together and contract one or other of the people in question to develop needed features, for example for Windows/AFS.

InDiCo

Mick Draper described the integrated digital conferencing tool being developed and pioneered for use for CHEP. It grew out of CDS (CERN Document Server) and in particular from CDS Agenda. InDiCo was a European-funded project to develop Agenda into a fully-fledged conference handling scheme including long-term archiving, aimed at scientific conference content and able to cope with multimedia. CERN, one of three partners, concentrated on the conference management part. It is being prototyped by CHEP and he described its main features, how they relate to the various steps in conference planning and the underlying technology. Like CDS, it will be open source and be freely available.

SUS Features for SMS

Michel Jouvin, in one of the very few Windows talks this week, described in great detail how SUS (Software Update Scheme) works with SMS (Software Management Scheme).

CERN Data Update Service

Tim Smith described his new activities in data services. He showed the target data rates of the coming LHC experiment data challenges and graphs of installed capacity. There are 370 disc servers with 544TB capacity spread over 6700 spinning discs. Hardware Raid 1 is used but there is a multiplicity of management environments in use and quattor is currently being modified in order to take over management of this installation. They are also porting Lemon for monitoring. There was an unacceptable disc failure rate. This was eventually traced to head instabilities and 1224 discs have been exchanged out of the 6700. For the future, the question is if hardware Raid 1 EIDE in a box is still the preferred answer, perhaps it is time to move to Hardware Raid 5 plus Software RAID 0 and perhaps we should look more at SATA (Serial ATA), both in a box and in disc arrays. Based on the trip reports on Day 1, SATA may be coming of age.

There are 70 Linux tape servers split across 2 installations of 5 STK Powderhorn 9310 silos. There is a move to 200GB STK 9940B drives which decreases both error rates (more modern technology) and the number of tapes required. By tracking tape errors, they are identifying problem tapes and retiring them, copying the data to new tapes. They are also retiring the most heavily-used tapes, thus again reducing the error rates. Investigating, with CASPUR, LTO-2 commodity drives and new high end drives from IBM and STK as well as new STK robotics.

Disc Storage Performance

Jan Iven presented some technical activities being performed in his group. These include evaluating disc server performance with respect to transfer rates and power consumption in various configurations, comparing USB 2 and Firewire. There were investigations with Infiniband where results are promising but it is still a new market with consequent high pricing and little practical expertise. Lastly he presented a review of disc server performance evolution with comparisons, recommendations and possible improvements.

Mass Storage Workshop

This day and a half meeting was organised and run by Olof Barring. Apart from most of the HEPiX attendees, it attracted a number of people, from LCG and elsewhere, who do not normally attend these meetings.

IBM SAN File System

The first talk was by Paul Bradshaw of IBM Almaden, speaking about Storage Tank, formally known as IBM TotalStorage SAN File System. It is part of the TotalStorage Open Software Family between storage virtualisation and the file systems. It is intended to shield the user from the need to access individual file systems by virtualising them across all the servers. Supports almost any attached storage, from IBM or other vendor, FC or iSCSI, all managed by a cluster of load-balancing metadata servers.

SAN FS puts the actual data in one or other storage pool with the individual pools defined according to local policy (customer, access speed, file type, etc) All this is declared as part of a global name space *a la* AFS and files can be easily moved between pools without changing the name space. Systems can be added or removed transparently. The key to all this is the virtualisation feature. He then covered individual elements of the components in detail.

Clients exist for most architectures and the Linux client is published as open source code. To the levels tested so far, benchmarks show the limits are those of the SAN bandwidth and it appears to be scalable since there are no architectural limits in number of files, file sizes, etc. They are currently working on features expected of hierarchical storage management, for example transparent staging to and from tertiary storage. It works with a variety of backup tools including TSM, Legato and Veritas. He finished by quoting the results of tests done at CERN.

Information Life-Cycle Management

The second invited speaker, Gordon Kennedy of STK, explained the application of information life cycle strategy to large scale storage requirements. ILM refers to managing data repositories in the most cost-effective manner based on the value of each piece of information. ILM depends on understanding the business value of the data. He explained STK's offering of different storage infrastructure components which are the bottom layer, where the data is stored. Above this, one must classify the data objects according to required access patterns, usage frequency or data integrity for example. This may demonstrate that particular data objects may be stored on particular storage components which match these requirements, at which a data mover may be needed. This process is ongoing and there is an aging scheme similar to an HSM (files unused for a certain period are moved towards offline storage). The whole is packaged as a service offered by STK; a reference customer is the Wellcome Trust Sanger Institute working on the Genome Project.

New Results from CASPUR

Andrei Maslennikov updated us on the recent storage activities at CASPUR. Much of the work is funded by industrial sponsors. Various configurations were tested including SATA/FC discs, Lustre clusters and so on. For each configuration he presented the most interesting features. Regarding the SATA/FC tests, he presented typical array features and some comments on the two vendors whose equipment was tested. He reported a failure rate of 2-3% per annum for the discs, excepting effects of power cuts. He presented the parameters for the tests, done in collaboration with CERN. The very first tests showed problems with one supplier's equipment and this has resulted in new firmware from them, expected shortly. Other tests show possible loss of performance depending on the file system in use (EXT3 as compared to XFS for example). For full details on the results, including performance graphs. Overall he is impressed by

SATA-to-FC disc arrays and believes they give good price performance; but the choice of the local file system is important and he strongly recommends XFS.

Turning to SAN file systems, based largely on current prices, he does not believe that these are yet a viable option for large farms but should perhaps be re-evaluated in light of the new iSCSI-SATA disk arrays. Rather they are better adapted to high performance or high-availability computing needs. Again his overheads have plenty of details about the tests and the results. The bottom line is that you can achieve speeds close to the raw disc hardware values and that the IBM solution (see above) looks very promising.

He also quickly showed a few methods to speed up AFS performance and one slide on the performance of a Lustre cluster and some about linear tapes. As usual, he ran out of time but his overheads are worth checking if this is your area of interest.

SDSC SRM
Wayne Schroeder of San Diego presented an overview of the Storage Resource Broker. This a piece of middleware that provides a uniform interface for connecting to heterogeneous data resources over a network accessing unique or replicated data objects. It is based on a logical name space via a metadata catalogue. It is a distributed solution and it has a rich set of APIs and GUI interfaces for file replication and management in general. It is tuned to be performant over a WAN and it scales well to millions of files and terabytes of data. Authentication management includes Grid Security Infrastructure (GSI). SRB uses containers to hold physical groupings of objects which eases management of sets of objects.

It is used in a large number of projects across many sciences. There is also a commercial version available from Nirvana, an offshoot of General Dynamics. Heavily used in various grid projects. It is not entirely open source but it is available to academic sites. Among its recognised weaknesses are that it can be difficult to explain and understand and its semi-open source nature. On the other hand, it is a sound and mature architecture while still being actively developed and it is supported by a highly coordinated team.

Supporting Multiple Mass Storage Interfaces
Jens Jensen of RAL described experiences gained by EDG WP2 on how to achieve a unique interface to different storage resource management (SRM) schemes (get, put, etc). The SRM protocol should be web-based such as SOAP interfaces via HTTP. Data transfer should be via GridFTP. EDG has developed a Storage Element (SE) with a uniform interface to mass storage and disc based on a simplified SRM. Some lessons learned including looking for opportunities for software reuse and that you should realise that prototypes often last longer then expected. In EDG, and many grids, grid files must co-exist with files belonging to the local site in the same mass store and this must be provided for. To get round incompatibility between GSI (HTTPG) and HTTPS, for security reasons, there is a proposal for something called G-HTTPS to bridge the gap. He closed with some of the plans for SRM 2.1 and the challenges ahead.

GFAL and LCG Data Management
Jean-Philippe Baud presented the tools being promoted for LCG data management to meet the requirements of the coming Data Challenges, especially for the LHC experiments. They require a common interface and they want reliable and performant tools. The LHC experiments need to access 3 grids around the world and they do not want to have to integrate 3 sets of tools to access the many different and widely-spread storage elements holding the data. The 3 tools used are SRM (previous talk), the Grid File Access Library and Replication and Registration Service (RRS).

SRM is widely deployed, there are interfaces to EDG-SE and CASTOR and interfaces are being prepared for LCG as well as an effort to create a standard to be approved at GGF.

The goal of GFAL was to provide a Posix I/O interface to heterogeneous mass storage systems in a grid environment. Various interfaces are or soon will be supported including RFIO and (coming) Root. As well as a library, which requires users to modify their code to call the open, read, write, etc commands from the library, work is going on to offer a GFAL File System which users could access without code change. GFAL is deployed in LCG-2.

The RRS is necessary to handle different SEs, different replication catalogues and to optimise the replication of files and handle failures. This is work in progress and will be reported on in due course.

There are various options for a disk pool manager including CASTOR, HRM/DRM and dCache and the latter, a joint DESY/FNAL collaboration, is the choice for LCG-2 and the reasons for this were given.

During the recent CMS data challenge, various problems were seen, although mostly on performance issues rather than stability, and work is going on to fix these for the next round. A lot of this involves reshuffling and working on the replication catalogues. Jean-Philippe explained how Oracle had been licensed by CERN for external sites to store these catalogues but, in the limit, mySQL would be tolerated instead for sites which refuse to install Oracle.

LCG Data Management Panel
Bernd Panzer started by setting the scene of the service challenges, what does is required to set up and operate a robust and reliable data management service for LCG? When LHC starts, there will considerable data copy from Tier 0 (CERN) to the various (today 7) Tier 1 centres, both raw data (one way from Tier 0) and ESD (Event Summary Data, both ways); estimated at 10PB per year in total. All this wanted in near real time. This implies a 70Gb/s connection from CERN. Hence the need to build up service data challenges to test the whole chain, from disc at Tier 0 to disc at the Tier 1 centres. Bernd listed some of the specific aspects to be tested. Who is willing to invest and participate on this? FNAL (US CMS) has committed but more Tier 1 sites are needed now to implement the first tests already this year.

Kors Bos asked why mass storage to mass storage, why not from the experiments directly: because it will be initially stored on Tier 0 mass storage before distribution. Karlsruhe felt the sites should be involved before the experiments perform their own data challenges. Bernd agreed and that was why he wanted data challenges independent of the experiments.

TRIUMF said that they are preparing tests as recommended by CERN. Fermilab had participated in the CMS data challenges but felt there should be some less work-intensive method for participating if such tests are to be prolonged or repeated. Nevertheless, they are interested in participating. RAL would also be prepared to participate although the data link is only 1 Gb/s and it needs to be scheduled carefully because of experimental data challenges; a link upgrade is planned later this year. Finding dedicated staff effort is more difficult, as Fermilab stated. Karlsruhe is in general on track installing hardware needed for such tests. Would like to define a week when all Tier 1 can get unique access to the facilities and that the experiments agree to give way to such tests. NIKHEF has the line speed needed and are busy now with experiment data challenges; so they believe that they could participate. Kors asked for more concrete details on the requirements. BNL is close to having the network and hardware capacity needed and although it is shared there may have sufficient unused capacity. Lyon is not yet ready to interface the required mass storage and there is doubt about the networking.

Is random access needed to tape to read back the data into users' programs? For raw data, random access is not common and anyway most if not all sites will use disc caching to access data. We should perhaps organise a workshop on this topic, data pre-staging, when there is more information about the experiments' data access patterns. Another question is what is the real network and mass storage access speeds required? Taking the raw network speeds is not sufficient, we need practical experience, hence the need for such service challenges. To the argument that we don't have all the pieces, the answer is that we should start with the simplest test and add components such as replication, as we understand the base and are ready to add more components.

The plan for this year is 10Gb/s end-to-end where the minimum is disc to disc; if the mass storage at Tier 1 sites cannot absorb this rate this is not a problem. The goal is to actually to use a 10Gb path and part of the test is to see what fraction of the 10Gb can be used. CERN will start with a single partner (FNAL) but would like to add more, with each Tier 1 site using a 10Gb line; CERN could then test Tier 0 distribution to different sites, first serially and later to several in parallel. In all cases, to fully exploit a 10Gb line, there needs to be multiple, perhaps hundreds or thousands, of streams. CERN should make a list of prerequisites to participates and then, when they have ascertained who is interested, to define the actual tests which should take place.

There could be a timetable for ramping up the tests from the simplest, copying files with simple scripts from disc to disc using GridFTP, overwriting the files at the receiver end; it would be the responsibility of the sites with the best aggregate network and disc server capacity to tune this transfer from their side. Suitable sites this year are FNAL, NIKHEF/SARA, Karlsruhe and TRIUMF. This would run for a week 24/7 at full speed without intervention after 2-3 weeks preparation. Along the same principle, a set of more complex tests, adding SRM for example, and detailed parameters (duration, suitable sites, etc) were listed and agreed.

There was a discussion about whether all sites need to have the same middleware installed; there may be conflicts with local requirements. The minimum is to have components which interface correctly.

Mass Storage - Day 2

Panel Wrapup

Les Robertson summarised the discussions of the previous day and the goals of the exercise: the scope is networking and data management, storage management and inter-operability and a fully functioning storage element. We need to establish a permanent infrastructure by simulating real-life data transfers and develop long term fixes rather than temporary workarounds to problems. This implies these frequent performance limit tests and feeding the results into the production LCG service. The focus should be service operability, establishing a reliable data transfer service stressing end-to-end performance.

Short term targets – by the end of next week - are to agree the participating sites; by the end of June, to agree a ramp up plan with milestones on a 2 year horizon. The end 2004 targets include being able to perform SRM to SRM, disc to disc transfers on 10Gbit links at a sustained 500 MB/s. Also there should a permanent service with mixed user and generated workload across at least 10 sites where a key target should be reliability. The first step is for sites to confirm that they can devote the required resources over the duration of the tests and identify a contact to represent them.

Olof Barring then continued summarising the problem of distributing the raw and ESD data among the sites, a total of 10PB of data per year to export from CERN; full-scale tests are planned for 2006. To this end, a sequence of tests are planned, adding complexity at each stage and he listed some of these with a proposed timetable for the Tier 1 sites to participate. Various meetings are now planned to get this on underway.

Storage System Performance
Jon Bakken of FNAL reviewed the results of integrating storage systems into high performance networking. The goal is to improve the data distribution from FNAL across the wide area to remote client sites of the FNAL experiments. He first listed the salient characteristics of wide area networking such as the effects of bandwidth and delays. He then described how storage systems should be configured to match these characteristics and how some investigations were ongoing at FNAL along these lines.

Integrating dCache at GridKA
Doris Ressmann described the steps performed in integrating dCache at the Karlsruhe LCG Tier 1 centre. Originally, all tape access was via TSM library management but this is not really an archive tool so dCache has been adapted which has more appropriate features for this. She showed how it fits in to TSM and the user interface.

CASTOR SRM 1.1 Experiences
Olof Barring described the evolution of SRM (Storage Resource Manager) 1.1 (a joint collaboration between Jefferson, FNAL and LBNL) and its integration into CASTOR. He described a number of interoperability tests, the various problems found at CERN and elsewhere, for example with the SRM specification, and how these were resolved. CASTOR SRM has now been introduced into production at CERN and has also been adopted by some other LCG sites.

Final Remarks
The meeting ended with a summary of the Mass Storage sessions and the participants were liberated into the sunshine of Edinburgh – just in time for the first rain of the week!

Alan Silverman
4 June 2004