HEPiX Trip Report

Brookhaven National Laboratory

18-22 October 2004

Summary

The meeting was attended by some 70 people from HENP sites across US and Europe. As usual almost all of the overheads as well as video (coming) of the talks are available on the web at http://www.rhic.bnl.gov/hepix/agenda.shtml. The following notes were taken online and supplemented with notes from Helge Meinhard. As usual, as this report is primarily targeted for CERN, I will not describe most CERN talks in great detail. Any errors are mine and I apologise to speakers if I have misunderstood or misrepresented their talks.

Among the highlights were:

- The adoption of Scientific Linux is increasing rapidly; it was mentioned in the site reports of NERSC, BNL, TRIUMF, NIKHEF, CCIN2P3, INFN (at least some sites), DESY (yes, DESY moves to Redhat-based Linux), LAL and RAL. In summary the HEP world is going to SL (qualified by SLAC and CERN…); it is being taken up very rapidly everywhere which is a major success for FNAL (and HEPiX)
- Other software packages which are being reused by various labs, especially in Europe, are quattor from EDG/CERN for node management and dCache from DESY; it is good to see such software sharing among the sites.
- The introduction or building-up of support of MacOS is still pretty slow except in some labs, for example in SLAC for a particular batch farm and in JLab for desktop use
- Still very few Windows-based talks and most of those were from CERN. Probably explained by the fact that CERN is one of the few labs building other services, for example groupware, web services, mail lists, on Windows. But before we re-think IS attendance at HEPiX, I should pass on the message expressed by a senior HEPiX colleague, in relation in this case to the relatively low attendance in man-days by Fermilab, that he considers that the Tier 0 sites (CERN, FNAL, SLAC) have a moral obligation to support HEPiX and tell us what they are working on because other labs often look to them for guidance.
- NERSC is migrating from LSF to SUN Grid Engine, starting with newly-installed nodes; the reasons are mainly cost.
- More and more OpenAFS sites are moving to Kerberos 5
- As usual there was a number of interesting performance tests reported.. I am considering establishing a web site under which we can store performance presentations and perhaps avoid repeating those already done.
- Tests at DESY in comparisons of Opteron against Xeon seem to indicate that (a) Opteron almost always came out on top (b) they are binary compatible so it matters little if they are mixed in a cluster and (c) it is worth adopting application code for 64 bit operation, there is as much as a 25% performance boost.
- IHEPCCC has asked HEPiX if we could provide technical input and feedback on specific issues. After an open discussion by the delegates present and a long discussion among the HEPiX Board, a positive reply has been drafted and is in circulation. In passing, HEPiX is encouraging IHEPCCC, especially certain sites and regions represented in IHEPCCC, to send more representatives to HEPiX events.

- BNL security rules – all visitors need special passes which are limited to 3 days at a time and there are sometimes problems getting first access on a Sunday when the administration offices are closed. A point to be considered when deciding whether to choose the inexpensive onsite accommodation or an offsite hotel! Despite these measures, the meeting was well organised by Tom Throwe and, apart from a small glitch with a beamer on Day 4, ran very smoothly with lots of time for discussion, good wireless access and a nicely-structured agenda.

Introduction (Peter Bond)

The meeting was opened by BNL's Deputy Director for Science & Technology, who gave a brief overview of the Lab. The total staff is 2800, the annual budget is $450M, mostly for nuclear physics and basic sciences. The main physics projects are the RHIC/AGS complex and the NSLA (national synchrotron light source) but various other sciences, environmental, chemistry, medicine and others, are represented in the programme. There have been 6 Nobel prices related with BNL work

SITE Reports

GSI (Walter Schön)

Tested the IMAP server and groupware from the Oracle Collaboration Suite but it needed a lot of work including help from Oracle to get it to work without bugs. Even then it was judged to be too complex, was missing functionality and was not quite ready for production use so they switched to Exchange 2003 and a Linux-based discussion forum (FUD) tool. Much work has been done in the war against SPAM, using Spam Assassin with a Bayes filter and Amavis; they estimate a success rate of 95% spam identification with less than 1% of false positives. Their treatment of SPAM, in particular the deletion of SPAM mails, is restricted by German law so the user must be involved, no default deletion is possible.

NERSC (Shane Canon)

PDSF at Berkeley has had a significant hardware upgrade, both dual Xeon and some dual Opteron systems, all nodes using native SATA technology and all on Gigabit Ethernet. They have adopted Scientific Linux but they are using CHOS (change root, see later talk) to run multiple Redhat releases. All new nodes run under SUN Grid Engine (SGE) rather than LSF. Indeed they expect to move all existing LSF nodes to SGE. They are still upgrading their 6000+ processor IBM SP2 which is still the largest system at NERSC. The networking team is building a 10Gigabit infrastructure and a 10G upgrade to WAN is in the works, waiting on Bay Area Metropolitan Area Network deployment by ESNet. Currently procuring a 320 node system with IB interconnects, running Lustre. Visualisation system procured as well (SGI Altix). Good experience with Apple RAID array

BNL (Ch.Hollowell)

They have 37 NFS file systems running under Solaris 9 and Veritas file systems but have experienced problems under the new release, VxFS 4.0, and there are scalability issues with NFS. Hence they tested both the IBRIX and Panasis products to resolve these NFS problems. There were initial reliability issues with Panasis but they have now worked them out and both products seem currently stable although Panasis seems more promising. They are also running AFS, currently the Transarc flavour but they are moving to OpenAFS to get Kerberos 5 support. They have added 335 dual Intel chip PCs to the 1400+ system Central Analysis and Reconstruction Farm (CAS/CRS) for RHIC. They are moving to Scientific Linux on this cluster and they use both LSF and Condor there. Disc storage on the farm has nearly quadrupled in the last year. Like everywhere, security is a major concern with plans for more hardening and monitoring of internal systems. Currently ATLAS data challenge jobs are being submitted via Grid3. They have been doing some research with dCache.

TRIUMF (Corrie Kost)
Westgrid is running well, a 504 node dual-Xeon blade system running Redhat Linux version 9 (for GPFS). They also use Redhat Fedora on some desktops but recommend and offer support for Scientific Linux. As usual, TRIUMF takes great care to optimize the various hardware and software parameters affecting throughput, both disc I/O and network bandwidth. Emphasis over the past 6 months has been on disc I/O and there is great detail in the overheads (see http://www.rhic.bnl.gov/hepix/agenda.shtml). In the near future they expect to acquire dual-CPU 2GHz Opteron systems which they believe out-perform 3.4GHz dual-CPU Xeons for I/O throughput (but he was unable during the talk to give details of the tests performed or the hardware configurations tested; he will try to get us access to them).

NIKHEF (Paul Kuipers)
A major activity recently has been the migration of their Windows domain from one based on NT to one based on Windows 2003. They use third-party tools for disc quota management (Quota Server from Northern Parklife) but it gives problems at this time. They use Windows Terminal Server with high security connections. Ironically, this became the source of a major security break-in and it needed a lot of effort to investigate, understand and recover from. On the Linux side, they too are moving to Scientific Linux, the release distributed by CERN (SLC3), on both servers and desktops.
.

CCIN2P3 (Wojciech Wojcik)
They are planning to migrate from Redhat 7.2 to Scientific Linux. 45TB of disc storage is being installed, almost half for BaBar. HPSS is still in production use with lots of local developments, interfacing to RFIO, bbftp and SRB. Also Objectivity is still in use for BaBar with an interface from HPSS.

INFN (Roberto Gomezel)
Desktops are 95% PCs (Linux and Windows) with a few MACs and some UNIX boxes; heavy use of VPN. On Linux boxes they use Citrix or VMware to access Microsoft applications rather than install Windows Terminal Server. A large number of sites have no anti-virus checking on mail but there is rising use of SPAM Assassin, now on 75% of INFN sites; some sites use other products for this and a few use nothing! On the Tier 1 farm in CNAF Bologna they have chosen to install commercial tools for remote console management and he described the Tier 1 farm in some detail, see the overheads.

SLAC (Len Moss)
.BaBar restart has been delayed as a result of a serious accident last week and a DoE investigation has just started and is due to run for around 3 weeks. BaBar has decided to adopt xrootd format rather than Objectivity and will gradually run down the latter. SLAC has just purchased ~300 Opteron systems from SUN (SUN Fire 20Z); they will run 32 bit Redhat Enterprise 3 on these. All power supplies on a recent purchase of 384 rackable Xeon systems were replaced by the vendor. They are happy with their Redhat support via their TAM; there are weekly phone discussions, bug fixes are provided as needed as well as useful information on new fixes and patches. Patch update is done via a nightly cron job which collects any updates from Redhat and then uses yum to distribute these to the SLAC nodes. Under pressure from BaBar, they are beginning to look at Scientific Linux and will evaluate it for build and interactive servers. Most SUNs now run Solaris 9 and they will look at Solaris 10 soon. OpenAFS support on Linux and Solaris will be provided by a small company specializing in this and they are talking with a senior OpenAFS supporter on possible work on Windows support. They now use Heimdal Kerberos 5 on the AFS DB servers and the switch went smoothly. They are investigating expanding LSF usage to manage

Windows and MacOS X clusters; the latter is a new project described in a separate talk later in the week. Another project, being promoted by Richard Mount, is a huge memory system, also described in a separate talk.. They have introduced Request Tracker, an open source tool but with optional commercial support, for ticket tracking for system administration.

DESY (Stephen Wiesand)
On Windows, they have decided to delay introduction of XP SP2 because of problems with some applications; they hope to introduce it by the end of 2004. They have introduced blade systems among the Windows servers farm. Although they are rolling out the latest SuSE version of Linux, they are actively preparing to move to Scientific Linux, based of course on Redhat Linux. Evaluations have shown up the first incompatibilities between the two distributions but they do not expect major problems. They also run (SuSE) 64 bit Linux on Opterons, chosen because of problems of AFS/Kerberos 5 problems on the Redhat 64 bit release; these are now thought to be fixed and future Opterons should have Scientific Linux installed. There is growing interest to run Solaris 10 on SPARC systems but no plans for a central MacOS X support nor HP-UX although they had to revive one HP node for a user whose code could not compile anywhere else. They participate in EGEE and other grid projects and they have a production grid cluster in operation since August running LCG-2. There is a plan for a d2Cache project to handle the very large number of files on the newest generation of tapes by introducing a middle layer between the current dCache and HSM. They are seriously considering adopting CERN's quattor for managing Linux and Solaris systems. Their new user registry propagates password changes to both Linux and Windows domains. Zeuthen uses the latest release of SPAM Assassin but Hamburg will go to a commercial product from Crocodile.

Karlsruhe GridKa (Manfred Alef)
GridKa is the German LCG Tier 1 centre and also used by BaBar, CDF, D0 and Compass. There are more than 1000 CPUs, all in dual-CPU systems, mostly Intel but including 36 Opteron systems.. All run Redhat 7.3. They are suffering some serious problems these days including frequent crashes of the PBS server and scheduler which the supplier is unable to resolve. Another problem is the cooling of PC clusters and water-cooling is being (re-)introduced, see later talk.

LAL (Michel Jouvin)
Main change since last meeting is the addition of 25 dual Opteron systems and the move towards Scientific Linux. They have added SASL to sendmail to get authenticated SMTP and thus permit roaming and home users to relay mail through LAL. SPAM filtering is now at 95% efficiency but it has taken 3 months to tune the Bayesian filters to reach this level. The IN2P3 Windows forest project (see previous HEPIX meetings) is now in production with 11 participating sites but there has not been time to introduce LAL. In a collaboration with Ivan Deloose, he has developed a multi-site Windows Printer Wizard. He also mentioned interest in adopting quattor for Linux installation and the serial console management tool from SLAC which has been installed also at CERN. Installation of InDiCo is in progress.

CERN (Helge Meinhard)
Helge gave a *tour de force* of activities in the different IT Groups.

Jefferson Lab (Kelvin Edwards)
Purchased MXlogic to short-circuit detailed investigations of various open source products; all mail dealt with off-site for checking and filtering before delivery to users on-site. On Windows, like DESY, XP SP2 has been delayed because of third-party CAD products. They have purchased 2 STK B280 storage systems with 14TB of disc space controlled by 4 SUN nodes and they are evaluating a 10TB Panasis storage system but this has been less stable and reliable that the STK although it is getting better now and they expect to purchase this soon as its performance is far above that of the STK systems. There is no change in their use of LSF on their main batch farm. Among future projects they include an evaluation of new hardware and software for e-mail and the introduction of support of MacOS as a desktop (the speaker stated that, after a period of self-denial, they now accepted that "Mac's do exist"). They are building a new computer centre due to open in January 2006.

RAL (Martin Bly)
The LCG Tier 1 cluster has now become the largest single CPU resource at RAL. They are now preparing for their next big CPU upgrade and discussing if it should be Intel or AMD, 32 or 64 bit. Various SUN-based systems as well as some CPU farms are moving towards Scientific Linux or Redhat 7.3, even 7.2 in some cases, although a few SUN servers will be retained. They have been disappointed in the disk I/O performance of early releases of Scientific Linux when used as a disc server but they have not tested the latest kernels. Quattor is being prepared for implementation on farms.

HEPiX Talks - Monday

INFN TRIP Project (Mirko Corosu)
The aims are the authentication and authorisation of roaming users requiring previous MAC address registration. It supports different authentication methods including passwords, certificates, LDAP, mySQL, etc. This works fine for captive web portal identification but to make it work for MAC address authentication, there must be a second logically-isolated local LAN to which unknown MAC addresses are directed. They are currently adding support for encryption.

Ranger (Chuck Boeheim)
This was an update on the latest developments on the SLAC monitoring tool, in this case to gather information from system log files, replacing the simple pattern-matching tool currently used for this, SWATCH. He used various Ranger objects such as Collectors, linked by rulesets to perform certain actions based on the state and context of the information being treated. Different rulesets can be applied to the same Collector depending on the state of the system or its context. He has built a number of standard actions including creating a "dup" action to take care of message floods, for example discarding repeat messages until some threshold is reached. As usual, the PERL source is freely available.

INFN Kerberos 5 Project (Enrico Fasanelli)
An update of a similar talk given in Vancouver at Fall HEPiX 2003. Kerberis 5 is needed to properly integrate the 7 AFS cells in INFN, one of which is already using Kerberos 5. The MIT flavour of Kerberos 5 was chosen, partly because it is closest to the Microsoft implementation of Kerberos, the latter being derived from the MIT version. Since more and more HEP labs are moving or have moved to Kerberos 5 (although perhaps not at CERN!), the speaker asked if it was time to introduce a common

trust relationship between them? [Two talks later, Bob Cowles pointed out the risks in inter-site trust, for example when hackers have gotten so good at grabbing users' passwords and even RSA certificates.]

LCG/EGEE Security Update (Dave Kelsey)
Security activities in EGEE are concentrated in the Joint Security Policy Group, JSPG) to define policies and procedure for security operations and to feed the input requirements into the Middleware Security Group. And all of this must be coordinated with cooperating grids such as Nordugrid and Grid3 in the US. He described the LCG Security Policy agreed by the JSPG and adopted by EGEE as well as by some national grids. The Open Science Grid in the US has been leading the operations for incident response (see report on Friday's sessions). He then described in some detail the plans for the user registration procedures agreed by and with the 4 LHC experiments. He included a brief description of the gLite security architecture. He emphasized the importance of site feedback on the proposed policies to ensure that the eventual policy will be acceptable to all sites which today or in the future may participate in a grid and he urged all sites represented at HEPiX to take note of the plans being described.

Security Update (Bob Cowles)
Bob entertained (and scared?) the audience with his now-traditional list of recent hacks, viruses and other attacks on systems commonly used across HEP. As usual, most of the talk concerned Microsoft Windows and products, from Microsoft and third-party, which run on Windows. He described the dangers of mails containing HTML addresses (but is that not how we are supposed to access our CERN pay slips in the future?). Spyware is a growing concern and it appears that when XP SP2 is applied to systems running certain spyware, the blue screen of death is the result.

Grid User Management Systems (GUMS) (Gabriele Carcassi)
GUMS is a tool developed at BNL which translates a grid identity to a local identity, similar to a grid map but centralized for a site, allowing that site to apply a consistent policy across its users. [Perhaps to be compared to the VOMRS tool being worked on at FNAL.] He described how it works, what a typical policy may look like, the main features and future plans. The speaker said that it should be used in the next major release of software from the Open Science Grid, due in Feb 2005.

HEPiX Talks - Tuesday

Using HEPiX as a Technical Advisory Body for IHEPCCC
There was a discussion on the request of Guy Wormser, chair of the newly-constituted IHEPCCC, for HEPiX to agree to provide on-demand feedback on technical issues to IHEPCCC. This might take the form of discussions of a particular nature at regular HEPiX meetings (as done today via the Large System SIG) or via specially-created technical sub-groups of short duration and limited scope. Some of the topics requested may be beyond HEPiX's current area of interest but this is not expected to be a problem, rather an opportunity. IHEPCCC would be prepared to add outside expertise as needed. The meeting attendees were somewhat reluctant to establish sub-groups but rather to discuss the requested topics in open session where possible. Based on the views expressed, for the rest of the week there were heavy e-mail discussions among the HEPiX board members in order to formulate an official reply. This reply, which has now been sent to the general HEPiX membership for approval prior to being sent to IHEPCCC, is generally positive to the request and offers access to the HEPiX mail list or the use of regular meetings like this one as a means of obtaining HEPiX feedback on specific issues. Specific-topic task forces could be established in cooperation with IHEPCCC but the latter should help provide

members for these. In addition, HEPiX will request IHEPCCC to use its influence to increase attendance at HEPiX meetings from HENP sites worldwide, especially those which today do not send delegates.

SATA File Server Performance (Walter Schön, GSI)
The results of some experiences and performance tests were presented. He started with the pros and cons of SCSI and (S)ATA discs and controllers, in particular the relative unreliability of early generation IDE discs. But the newer SATA servers have triple redundant power supplies with hot swap and the 250GB discs specified for the GSI tests are described as good for 24x7 service. The discs are attached to 2 8-port SATA RAID controllers and they could attach 14TB of discs. As shown in the overheads, he described in great detail the hardware and software configuration and the tests performed and he showed graphs of the results in read and write performance, the influence of cache for different file sizes, the effect of tuning the cache size (good tuning can increase read performance by factor of up to 3), total read throughput for different number of concurrent processes, and so on. The presentation was very thorough and anyone interested in this topic is recommended to read the overheads or watch the video. After bypassing an initial, rare, problem, no disc hardware failures occurred and overall performance of the discs was deemed good if combined with high performance controllers and well-tuned system (especially cache) parameters.

Panasas Evaluation (Robert Petkus, BNL)
Currently they use a SAN-based NFS service with 37 SUN controllers. The reasons for looking beyond this include the lack of load balancing, limitations of Veritas (scaling issues and poor support), an upper quota limit of 1TB and the significant cost. There are a range of implementations available to bypass some or all of these disadvantages. After a survey of the market, Panasis was chosen for a detailed evaluation and he showed some reasons why – integrated hardware and software solution, dynamic load balancing, seamless expansion, etc. It is a blade server with director and storage blades. Both use Intel chips and run Free BSD. Initial testing was disappointing with bottlenecks appearing quickly before system collapse and at least one case of server data corruption. Once Panasis had understood the problem and fixed it (very quickly), the tests were much more successful and they were able to saturate the network at 2Gb/s with low CPU overhead under NFS and no crashes or corruption. Some minor drawbacks were reported to Panasis and some of these have already been fixed. The next stage is production testing by an experiment and the speaker is optimistic of eventual success.

Managing Managed Storage (Jan van Eldik)
Jan described CERN's current data services with 350 disc servers, 6700 discs, 550TB plus 10 STK robots and 70 tape drives. He listed the applications, the "players" and the configurations. He described some of the problems encountered and how the challenges were met. The result is a more stable, more reliable and automated service better integrated into standard operational work flows.
.

Condor Batch Submission (Tomasz Wlodek, BNL)
The challenge is to submit jobs which stage HPSS files, both input and output files, and capable of mass production. Condor was chosen, partly because it is becoming a standard and partly because of DAGMAN, allowing to create sub-jobs for data staging and combine these with jobs for processing the data. He described the models of the job submission scheme.

LCG POOL (Dirk Duellmann)
Dirk described POOL, the persistency framework for storing and accessing the multi-PB experiment data and metadata in LCG. He also described a proposal for an LCG Distributed Database Deployment (3D) Project on shipping databases around the grid. He explained how he has broken down the needs and what the architecture could look like and what is being discussed with the LHC experiments, the first users of the proposed scheme.

CERN Software Licensing (Matthias Schröder)
Matthias described the management and service issues around the licensing of commercial software in CERN as implemented by IT/PS Group. NERSC has expressed interest in our interface to FlexLM monitoring data.

HEPiX Talks - Wednesday

AFS Monitoring (Alf Wachsmann, SLAC)
Alf described some work done by himself and a summer student on writing PERL APIs to AFS monitoring and debugging tools. As much source code as possible was taken from OpenAFS but much simplified and cleaned-up; bugs were found and fixed and the result will be fed back into the next OpenAFS release. In the meantime, the package is available from him and the release package includes, for each function, an example of a PERL script calling sequence. He ended by showing some examples of using these APIs.

Use of Quattor outside CERN (Rafael Garcel Leiva, Uni Autonoma de Madrid)
His first reaction on faced with installing Grid middleware was its complexity and how hard it was to understand how to install it. The hope was that adopting quattor would ease these tasks on the grid nodes, as well as helping to support desktop nodes so he went to work with the CERN quattor team. He first described the use of quattor at LAL where they have dropped use of the software repository (do small sites really need this?) and they use only part of the configuration database (CDB). LAL's experience was largely positive but they believe that the CDB should be able to manage multiple test beds with one instance, that the pan language should be extended to simplify configuration management and there should be a mechanism for secure distribution of machine credentials. In Madrid, the software repository and full configuration database were retained but only part of the automated software installation package. Their feedback is that they need help to build new or extend existing configuration components and they would like to see more support for disk partitioning. But again they were largely pleased to have quattor. He then summarised the results of a questionnaire he had sent out. He listed more than half a dozen sites or organizations using or planning to adopt quattor. General experience was that it is highly modular, flexible, portable (at least within Redhat Linux), it is easy to install and no major design problems have been reported so far. [Take a bow, members of the quattor development team.]

UIMON to LEMON (Matthias Schröder)
UIMON is a tool developed some time ago inside PS Group for monitoring their servers and now this tool is being replaced by the new IT monitoring tool LEMON. The main work was to adapt LEMON, especially the actual monitoring sensors, to take care of Solaris UNIX since it was originally targeted only at Linux. The port identified some internal bugs (e.g. memory leaks) and non-POSIX code. The major missing feature in LEMON which is present in UIMON is that of issuing recovery actions and work is taking place in this area currently. LEMON's main features are its modularity, ease of addition

of new components and new sensors for any O/S, its central repository of statistics and the exception or alarm metrics. The disadvantages are that it has only basic graphs, only a small set of metrics built-in and the fact that it does not use system calls to collect data.

Real-Life Problems and Solutions on a Large Farm (Thorsten Kleinwort)
Thorsten showed how the ELFms tools (Quattor, Lemon and Leaf) fit together. Using Lemon's web interface, he showed some live monitoring data from LXBATCH. Current work on quattor at CERN is concentrating in overcoming some scalability issues and on developing some utility scripts around it. He described the two main components of LEAF, the State Management System (SMS) (a system has various possible states such as production or stand-by) and the Hardware Management System (HMS) to track where a system is physically located (for example, newly-arrived, in repair, etc). He then moved to some use cases, for example, how they upgraded the kernel of 1500 LXBATCH nodes, how they configure LSF batch resources automatically every day (with a cluster of 1500 nodes, there are changes practically every day), how they could easily shutdown nodes prior to moving them to the other side of the Computer Centre and how they configure batches of new nodes as they are delivered to CERN.

High Availability Linux (Karin Miers, GSI)
This was a scheme to design a scheme which relied on more than heartbeat to provide a highly available NFSservice. The speaker explained why heartbeat monitoring alone was not enough and how she had added some open source modules to overcome heartbeat's defects. This consisted of DRDB, a distributed replicated block device – basically a kernel patch implementing RAID 1 over the network.

Chroot OS (Shane Canon, NERSC)
CHOS was developed partly in response to NERSC's need to support multiple user groups simultaneously. CHOS is a framework to support multiple operating systems concurrently on a single system. CHOS should be as transparent as possible, recognizing which O/S to run for a particular submitted job. It is based on using chroot on a Redhat 8 base into an alternate O/S and then setting up the path links appropriately. The talk explained the internals, how it solves scalability via the use of links, how a .chos file in the user's home directory is used to decide the target O/S (with system or batch queue defaults) and some of the batch and security aspects. He ended with a live demo from his laptop.

Windows Terminal Server (Ruben Gaspar)
Ruben explained why CERN has setup the WTS server, what problems it solves and what services it offers. To avoid incompatibilities with standard desktop applications settings, the users' WTS profiles are kept separate to his/her normal roaming profiles. He showed the configuration and how load balancing was achieved. He performed a live demo of the service. There are currently some 460 active users from a registered number of 782. Average sessions per day is 35 with peaks of 45. The service is popular but the support model is unclear if it continues to grow.

DESY's Windows 2003 Domain (Reinhard Baltrusch)
Since some time DESY has been working towards a single Windows domain based on Windows 2000, later 2003, across both sites. Currently some 50% of the Windows nodes have migrated, the rest are still on the old Windows NT domain. The servers of the new domain include a 14 node HP Blade server. As experienced by all sites which have gone through this, such major migrations from NT are long-drawn out and some user communities migrate faster and with fewer problems than others. A great deal of effort by the support team is essential. Indirectly associated with the Windows 2003 project are an

Exchange 2003 server (currently they are using Exchange 5.5), Windows Terminal Server and others. The main problems are or have been interruptions to the home directory service which has taken some time to debug and understand, problems with their third-party quota manager (from Northern Parklife), confusion over the use (or not) of WINS, and synchronization issues in the online/offline use of laptops.

SMS 2003 Deployment for Managing Windows Security (Rafal Otto)
Rafal described what SMS (Systems Management Server) is, its architecture and how it is integrated with Active Directory. It is very useful for application deployment and licensing monitoring (although by itself it is not a monitoring tool) and its use by the central Helpdesk for remote debugging users' Windows problems is becoming more common in CERN. He explained the model for assigning SMS access rights to different support teams and end-users. In order to get SMS to reflect immediately any changes to the group policies and Active Directory, a special mechanism had to be applied otherwise it the user would have to wait until the following day. The SUS package is used with SMS to check the Microsoft update web site for new patches and decide which client nodes on the site need to be updated. But not all patches can be applied by this means. Server updates must be carefully handled because most server patches need to be followed by a reboot after patching and we need to schedule carefully reboots of servers. For desktops, non-urgent patches are not applied immediately but rather pre-advertised to those users interested. Other patches are forced immediately. For those patches not supported by SUS, an SMS packaging is added locally and the result is added to either the non-urgent patch packages or the update-now set. Current work is in the preparation of the deployment of XP SP2 and in tightening the rules around the use of administrator privilege.

Exchange 2003 and SPAM Fighting (Rafal Otto)
Rafal summarized on behalf of Emmanuel Ormancey various activities at CERN around e-mail. The team is currently in the middle of a server software upgrade which should be completed shortly. Migration for the users is transparent except for the need to close and re-open the client when the user is moved to the new server overnight. He showed the new mail gateway structure, noting how it was simplified and better suited to fighting incoming virus attacks and SPAM mail. He then described in some detail recent measures to identify SPAM including reverse DNS and reverse SMTP checks. Such measures identify 84% of incoming mails (92% at weekends) as SPAM. The latter test, added recently, had blocked a small number of valid incoming mail, eg from a listbox server at Fermilab, and some investigation was needed to understand why (configuration issue at source). On the other hand, it alone is responsible for identifying 25% of mails as SPAM.

New Mail List Infrastructure (Ruban Gaspar)
This was given on behalf of the IS Group Listbox team. Compared to the previous implementation of Listbox, the new service should have a lower maintenance cost and is better integrated with the Exchange mail server and anti-SPAM fighting. From the user's perspective, the list management interface is little changed and the functionality is very similar. Special user accounts with restricted access rights will permit access by people with no CERN mail account.

Large System SIG

Platforms for Physics

The aim of this part of the meeting was to present talks referring to different architectures used in HENP today. Unfortunately, promised talks on the MAC cluster at Virginia Tech and on using AMD chips in Pisa had to be cancelled due to travel budget restrictions and we were unable to get a speaker on work done recently on PS2 chips in Urbana.

Intel 64 Bit Processors (Phil King, Intel)
Intel intends to maintain their current 2 chip families for the foreseeable future and he showed the near futures for these. Xeon should follow Moore's Law but Itanium will scale up faster. Itanium should always beat Xeon in processor power but for the time being not in terms of price-performance, although Intel's intention is to narrow that gap over the coming years. Itanium will be the first Intel chip to get dual-cores as it is a less complex chip than Xeon; by 2007 the target is 8 cores.. He explained the EM64T extensions to Xeon but he warned that the 64 bit extension is quite new and some necessary software drivers, compilers, etc may still be missing. He presented a slide comparing the architectures of both families. Among his backup slides (see the meeting web site, www.rhic.bnl.gov/hepix/agenda.shtml) was a slide comparing Intel and other chips, as seen from Intel's perspective.

AMD64 / Xeon EM64T Comparison (Stephen Wiesand, DESY)
DESY chose to use "extended 32 bit" platforms rather than Itanium to avoid users having to modify their codes to take advantage of the 64 bits. In fact both AMD and Xeon "64 bit" families are limited to 40 bits physical memory and 48 bits virtual memory. He detailed some architectures features of the new chips and the effect they have, positive or negative, on programming for them. He also noted some differences. He then moved to comparisons between the chips and showed the results of a number of tests, for example those coming with the Root package. The results should be consulted on the web but for the Root tests, for example, showed that the Opteron performed better, possibly because the second CPU of the Xeon had trouble getting sufficient access to memory. Both chips run fast in both 32 and 64 bit mode but significantly faster in the latter. And commercial compilers perform better than those traditionally used in HEP. 64 bit Linux looks very similar to 32 bit but there are a few problems, for example in porting physics applications to 64 bit mode to make the best use of the 64 bit extensions. He ended with some 32 bit compatibility issues and some problems found with AFS Kerberos 5.

IA64 Storage Servers (Jan Iven)
Jan presented the results of some work done in openlab by Andreas Hirstius. We need to optimize I/O throughput for the LHC experiments to make best use of the grid, with 2005 disc to disc targets of 500MB/s sustained. Various network interconnects could be adopted to achieve these rates and some of these are the subject of these tests, using Itaniums under the banner of the CERN openlab project. The results using 10Gb Ethernet NICs were shown. Similar tests are starting with Infiniband. Jan invited interested people to contact the author at CERN.

Price Performance AMD/Intel (Maxim Potekhin, BNL)
A collaboration of CASPUR, BNL and CERN (Eric McIntosh) performed some tests comparing Intel Nocona and AMD Opteron chips based on a few standard Fortran benchmarks (eg Sixtrack). As on the DESY talk, the Intel chip showed a notable performance drop on the second CPU, again blamed on

memory access. The suggestion was made to pursue such tests but on codes more relevant for the LHC experiments.

Water-Cooled Clusters (Manfred Alef, Karlsruhe)
GridKA is physically-constrained to build compact clusters consisting of pizza boxes and blades and this results in heating problems. They have decided to revert to water cooled heat exchangers in 19 inch cabinets and closed air-cooled racks and they claim to have the world's first completely water-cooled PC cluster. They have currently 30 such cabinets with a further 25 on order with each rack costing some $12K. There have been no significant problems and a notable reduction in noise levels in the computer room.

Huge Memory System for Data Intensive Science (Alf Wachsmann, SLAC)
This was a repeat of the recent CHEP presentation by Richard Mount based on a proposal to the DoE. A major challenge in today's science is sparse access to objects in huge databases with the characteristics of these objects differing widely among the different scientific disciplines. The proposal is to revolutionise the query and analysis of scientific databases with complex structures. Key to solving this problem is memory, preferably physical memory as opposed to disc memory in order to radically improve access speeds. But the raw cost of memory is a factor of 100 higher than disc. SLAC is creating interest among commercial suppliers, first for a development phase, using BaBar to show proof of concept and building interest and support to build a leadership-class facility. They will begin with AMD server mainboards, 2GB DIMMS, probably with Solaris as the O/S (because of its maturity in handling 64 bits). In the first year the target is to have 650 2-CPU systems with 16GB memory each. They currently have funding for between 64 and 128 SUN Fire V20z nodes as data servers and a total of around 2TB of memory. The leadership-class facility, a proposal for which should be developed by 2006, should have around 256TB of data cache.

Installing a Mac G5 Cluster (Chuck Boeheim, SLAC)
At least for now, MacOS support at SLAC is limited in scope, aimed solely at a new Stanford/SLAC department devoted to particle astrophysics. Their choice was between Opterons or Macs and they preferred Macs, partly because of their familiarity as desktops and partly because of their visualization features. The cluster consists of 2 file servers, 2 interactive servers and 10 compute nodes, all G5 servers. Early experience of installing the hardware is positive. It supports serial console and serial BIOS but no remote BIOS management and no power management - you need to press a button to boot. Network installation works and subsequent installs should be able to be automated. For users, it looks like UNIX (including most gnu tools from Fink) but for system administrators this is only partially true; it has a strong BSD flavour of course but many tools are in a different place, some remote configuration tools are missing, directory access seems not to be scriptable. Worst, startup is significantly different and will change again with the next release (10.4, Tiger). It supports standard authentication via NIS, LDAP, Kerberos and Active Directory. Other minor drawbacks relate to the presence but non-use of /etc/passwd, the non-standard implementation of shadow passwords, the choice of AFS default permissions and non-case sensitive file names on the default "HFS" file system (Makefile and makefile are the same!). Chuck's summary is that it is a good hardware package, ready for network installation, with a few wrinkles, and it has good configuration management tools although these have unfortunate differences from other UNIX flavours.

.

DESY Production Grid (Patrick Fuhrmann)
DESY is a member of EGEE, D-GRID (German e-science R&D, proposal due to be submitted this week), ILDG (Lattice QCD), LCG (trying to become a Tier-2 Centre for CMS). Related to these, dCache developers are working with the LCG team to resolve the final dCache issues affecting its wider use in LCG. The DESY Production Grid is based on the LCG-2 release. Installation has moved away from "clumsy" SuSE tools towards the friendlier LCFng and they hope to move eventually to Scientific Linux and quattor. The main users are HERA, H1, ZEUS and the ILC (International Linear Collider).

Conclusions of Platforms for Physics (Alan Silverman, Chair)
After an informal, interactive discussion, the following consensus views emerged:

- At the present time, AMD's Opteron systems out-perform Intel's price-comparable Xeon systems; the percentage depends on the application but can be as much as 30%
- Since they appear binary compatible, it is not a significant risk to purchase whichever system is ahead at the moment of purchase
- Adding price/performance to the comparison appears to make little difference
- More and more purchases will be of 64-bit capable systems and running applications in 64 bit mode can add 25% performance; however physics groups are likely to need some persuasion (or incentive) and some people complained that they require a 64 bit CERNLIB!
- The Tier 0 sites should lead by offering such 64 bit platforms
- Only a few sites represented are seriously considering Itanium; SLAC have a request from their astrophysicists; Ian Bird states that several LCG sites not represented are seriously considering Itanium
- There is some pressure for MacOS support but mostly as a desktop, the major exception being the astrophysicists (again) at SLAC
- It is unlikely that MacOS will overtake or replace PCs in our environment in the foreseeable future although they may become the preferred second platform for physicists.

Linux

Scientific Linux at INFN/Trieste (Roberto Gomezel)
After investigating several alternatives, and being forced to move on from Redhat 7.x and version 9, Trieste selected SL because of the promise of support by the "HEP community". They now have 30 clients and 3 servers running SL. They use mostly a variety of locally-developed tools to assist in managing their systems but the more common yum is also used for managing RPM packages.

BNL Experience with SL (Christopher Hollowell)
They looked at RHEL (not cost effective) and Fedora (too short support times, questions on stability), other distributions (need to port management tools and methods). They are now in the middle of upgrading the main compute farm to SL. Successfully tested various commercial packages against SL.. Overall they are impressed with the distribution.

Linux Panel (Alan Silverman, Chair)
Chuck Boeheim (SLAC), Jack Schwarz (FNAL) and Jan Iven (CERN) gave status reports on their respective Linux support activities. Chuck repeated what he said in his site report – they have regular contact with their Redhat technical support and are happy with the communication and the subscription model in general; hence they have recently renewed this for the next 12 months. SLAC have evaluated

the cost as approximately half an FTE and they believe this is value for money. They share the TAM with Livermore and this sharing has its own value in addition. Such sharing, which includes seeing open and fixed problems, could perhaps be extended to other HEP sites which possess or join the subscription model. Questions need not be limited to only Redhat software, one example was a request for help with AFS support under the 2.6 kernel, another was for help in identifying the source of an LSF problem – was it in LSF code or in the kernel? Although they would like to sell their update scheme, Redhat have helped SLAC to access the Redhat updates with their own (SLAC) code and then re-distribute them inside SLAC with yum. SLAC has opened some 50 problems or requests in 8 months and are generally satisfied with the response times.

Jack reported on efforts to create and maintain a web site for Scientific Linux (www.scientificlinux.org) which covers mail lists, access to distribution, a bug tracker and so on. They are happy with the uptake of SL and feedback on its use and help in its support. They have started looking at Release 4 and should have a beta release out two weeks after RHEL 4 official release. A suggestion from NERSC that other sites participate actively in SL support, for example taking responsibility for part of it, was welcomed. But I insisted that such offers must be for serious long-term commitments.

CERN's Redhat support contract, based on a subscription of 200 nodes and a Technical Account Manager, has only been alive for 3 months but the situation is less satisfactory than that of SLAC and Jan and Redhat are discussing how to improve this. Nevertheless, CERN has already submitted some 40 calls of many different types (requests for help, requests for changes or advice, etc). Initial response times are good for acknowledgement but less impressive for actual fixes or replies, especially where Redhat support believes the question is outside the terms of support for RHEL 3. Weekly phone conversations have not been so useful and are no longer regular. Despite this, CERN will continue with this support for the rest of the 12 month period and re-evaluate it then.

Regarding SL, CERN are completing certification of SLC3 where C standards for CERN tailoring. There has been general acceptance by most of the CERN Linux population (and a few outside sites). They are sharing debugging of problems with FNAL but admit to being guilty of not feeding back every issue to FNAL. They decided to model the distribution layout to the Redhat style rather than that of FNAL SL. External sites can take the distribution from CERN and either accept the CERN modules or roll them out. Or they can take the FNAL kit and use the "handles" included by FNAL to insert local tailoring.

Other sites, for example NERSC and TRIUMF, expressed satisfaction with SL and many other HEP sites have or will adopt it. BaBar seem to insist on having SL build servers as well as RHEL as they have lingering doubts on 100.0000% binary compatibility. DESY would like to help support it but do not feel they have available resources at this time.

Summary: the HEP world is going to SL (qualified by SLAC and CERN…); it is being taken up very rapidly everywhere which is a major success for FNAL (and HEPiX).

Grid Operations Experience

This was a series of talks organized and led by Ian Bird to discuss operational issues discovered in the recent data challenges in both LCG and Grid3. Many of the issues which will be covered will be explored in more depth at a grid operations workshop in CERN in November. Ian's intention was to use the opportunity of this HEPiX meeting to get input from some sites which may not be represented at the forthcoming workshop.

OSG Incident Response Plan

OSG stands for the Open Science Grid and Bob Cowles presented the current state of their plans for incident responses. With little or no control over the physical resources, local security personnel; must feel comfortable with the grid's use of "their" resources. OSG should provide centrally a list of security contacts, secure e-mail, an incident tracking scheme and other similar functions of a Grid Operations Centre (GOC). In return, local sites must have a site incident plan, they must promise to remove compromised servers from the grid (if not the network) and provide evidence (logs for example) of incidents at that site. There are likely to be Response Teams created to react to serious incidents and already some broad guidelines for incident investigations and handling have been drawn up. The plans are being discussed with EGEE and LCG.

iVDGL Incident Response

Leigh Grundhoefer actually covered incident response in a number of US Grid projects such as Grid3, iVDGL, etc. Their iGOC (iVDGL Grid Operations Centre) is staffed by 2 FTEs to look after a wide range of operational matters. They established collaborations with the component grid units. They perform lots of resource monitoring using the usual tools, Ganglia, Mona Lisa, etc; they spend 60% of their time on this, making the data available via the web to the grid users. The rest of their time is split between Virtual Organisation support, support for application developers and workflow environments (portals) and support to Grid users. Having no direct control over a site, they operate (rather successfully it would appear) in a cooperative manner. They operate a trouble ticket system and try to perform end-to-end troubleshooting for resources. She explained how the current activities relate to future OSG plans. In the discussion, Markus Schulz noted the similarities with the LCG GOC model, in spite of having not been previously coordinated the two GOCs.

LCG Grid Monitoring and Accounting

Dave Kant presented this remotely from RAL via a telephone link. LCG needs to monitor over 8000 CPUs across 83 sites (today) and be ready to react to problems. LCG has gathered a number of monitoring tools together from all over the world and integrated them into a coherent set. There are lots of local aspects of Grid sites which are not centrally known and which complicates monitoring. Plus there is a need to develop tests of grid functionalities. Dave went through a number of the tools in use, some of which were the same as those in the iGOC for iVDGL and he showed some status graphs and reports produced by the different tools. The LCG GOC has established a configuration database of people, resources, contacts, local maintenance schedules, etc. This allows to determine which monitoring tools should be applied to which sites. Site certification of new sites is performed by the CERN deployment team, certifying the correct installation and configuration of the Grid middleware and since one must continually hunt for problems this service is performed on a daily basis by running a series of tests via the replica manager. All of the tools have been developed independently and do not interface to each other, nor share their data. So an R-GMA archiver has been developed to save monitoring data from different places. Packages have been developed and tailored for the various users of the GOC (regional GOCs, individual sites, grid managers, etc) and he demonstrated as a use example the GOC accounting package for LCG. Further on-demand tools are planned.

LCG/EGEE Security Operations

Dave Kelsey presented the status of security operations as of today in LCG/EGEE. First of all, security is not (yet) part of the LCG GOC activities although the GOC has written 3 security guides for the LCG Security Policy. Steps are being taken to integrate LCG and EGEE security activities. While O/S security patches are a local site responsibility, the Grid Deployment team is responsible for security

patches of Grid middleware. LCG has agreed that sites should keep security-related log files for 90 days for audit purposes. There is also agreement on an incident response policy, including what constitutes an "incident". The Joint Security Policy Group is developing formal site registration procedures for joining LCG/EGEE. He then showed some overheads as presented recently by Ian Neilsen, LCG's Security Officer, to the EGEE Regional Operations Centre managers. These covered the objectives of security coordination (ownership of problems, active follow-up of incidents, etc), coordination of security policies across the sites of EGEE and with similar grids elsewhere, incident response procedures and in general the need to work together, with both the local sites and central grid deployment having a role to play.

LCG Operations

Ian Bird explained some of the terms referring to Grid Operations, for example the difference between LCG the project and LCG the release packages and where LCG interacts with other grids such as EGEE, Nordugrid, Grid3, etc. He explained the hierarchy which was introduced to handle operational aspects between Tier 0 (CERN), Tier 1 and Tier 2 sites, especially in regions where there are no Tier 1 sites (Russia, Canada, etc). One issue is the perceived need for a single trouble ticket scheme across all sites. LCG covers many small and a number of large sites. The former generally want a simple packaged installation, the larger sites often must "fit" LCG middleware to their local environment. He explained briefly what EGEE is and what is (will be) LCG's relationship to it. Although the sites and the partners may not be the same (there are US sites in LCG), LCG-2 and EGEE share the same infrastructure, hence there is a single operational coordination activity (at CERN). He listed some issues which need to be addressed in order to move forward as the experiments begin to use grids as production facilities, for example feedback to middleware developers, fabric training for managers of small grid sites.

Operational Models

Markus Schulz then continued with some experiences from recent Data Challenges performed on LCG. During these data challenges, it was noted that large sites integrated their resources into LCG (rather than dedicate them). Since May of this year, LCG release monthly incremental updates to LCG-2 although not all of these are distributed to external sites. The LHC experiments have made significant use of LCG as shown by the graphs in Markus's slides. On the downside, all experiments met similar problems such as some sites with inadequate human resources, sometimes problematic load balancing between the sites, identification and location of problems, the need for more tools to handle thousands of jobs in bulk mode and the performance and scalability of services. The current status is that the middleware, while not perfect, is quite stable, so work is ongoing on the other issues (lack of staff, wrong configurations, better and easier-to-read documentation, firewalls, etc). He explained the daily site certification scheme. Looking to the future, how much of operations can be delegated to different players, can one model fit all sites and regions? The first model considered is a strict hierarchical model where different regions may have different operational procedures and the GOC never contacts sites directly. The flexibility to have different operational policies is probably a plus but the need for involving the Regional Operational Centres in all cases adds latency to incident handing. A second model is direct communication with sites but local control of the resource centres. This has a high cost for sites in the need to be contactable at all times and doubts over the "maturity" of all local administrators across hundreds of sites. The obvious, third variation of this is direct communication with direct control of grid resources but this requires a set of secure remote operational tools, and trust.

Discussion

How acceptable would be remote operations? Is a grid-sudo ever going to be acceptable? Certainly not among larger sites but this is not where such tools would be needed, rather at "one-man" system

administration sites. Perhaps the Cluster Building Guide being worked on (at a low level) by the Large System SIG (ie. me in my “spare time”), or some cut-down version of this, could be useful for small sites, some of which appear ready to receive assistance in setting up and running a cluster in a production manner which may be a new way of working to many of them.

Alan Silverman
(with help from Helge Meinhard)
24 Oct 2004