HEPiX Trip Report, Fall Meeting, 2007

Genome Sequencing Center at Washington University, St. Louis, MO

November 5th to 9th

CERN-IT-Note-2007-046

This is probably the first HEPiX meeting not hosted by an HEP site. Due in part to strong attendance from the UK Sanger Institute for Genome research, the attendance was around 70, a healthy number for a US meeting, and this was further boosted during some sessions by some members of the local institute. The meeting had been planned with specific topics in mind, each topic with one or two conveners. As a result, there were a total of 61 talks submitted. It was well organized: the meeting room was well-equipped with power plugs and wireless capacity; the hotel was a few minutes walk away and the agenda still allowed lots of time for informal discussions. As usual, I have done my best to represent the contents of the talks, with some additional notes from Helge Meinhard which I gratefully ackinwledge, but readers are recommended to consult the overheads of topics of particular interest at HEPiX Fall 2007 (05-09 November 2007) and look under the Agenda link.

The welcome address was given by Gary Stiehr, the principle organiser. Why HEPiX at a Genome Center? Hehad got to know about HEPiX while working at Fermilab and now that he is here he believes it would be interesting to have the two communities interact, at least as far as IT support activities are concerned. The Center's Director, Dr. Rick Wilson then gave a talk about the work at the Center, considered one of the premier genome research centres in the world, created 17 years ago. He indicated some of the techniques used and their associated IT aspects. Increasing data samples and use of parallel processing are expanding greatly their IT needs and across the road from the Center is a large construction project which will house their new computer centre (described in a later talk).

Some random highlights

- Many sites are still reporting work in and around computer centre infrastructure, power, cooling and space and Wednesday morning was devoted to the subject
- SUN Thumpers are almost universally chosen for disc storage services
- Apart from CERN, almost all sites reported that they were "ready" with Vista but only CERN seemed to be actively promoting its use, even if only on new PCs. Offline, Rafal told me he hoped that the other sites had fully understood what being "ready" really meant.
- It is presumably just coincidence, but I lost count of how many sites reported new directors in the past or future 6 month period. Is there an epidemic?
- In an offline discussion with the OpenAFS gurus (see OpenAFS talk in the Storage Systems section), it became clear that they could be funded to implement HEP-desired features. It was agreed that Maslennikov would invite the AFS experts from the various labs to look at the OpenAFS wishlist and try to agree a package of features of interest. This would be given to the OpenAFS people who would price the implementation and the labs could discuss cost-sharing. However, I remain a little sceptical: even if such a package of desired features could be agreed and cost-sharing then agreed, the OpenAFS suppliers would need to adapt their pricing to academic environments. For US DoE-funded projects they charge over $300K per man-year of effort and HEP will simply not pay that much. We will see what happens.

- There was a somewhat heated Board discussion by e-mail on the suggestion (from me) on extending HEPiX to invite the 2 Genome sites attending this meeting and apparently interested in interacting with us given that there seems to be significant overlap in IT challenges, I suggested we invite both GSC in St. Louis and Sanger in Cambridge, England to join us on the Board and perhaps the Sanger may wish to host a future European meeting. Every Board member present in St. Louis agreed. But this provoked a mildly heated discussion on the Board mailing list where the Fermilab BoAard rep (not present at the meeting) was almost alone against this expansion, citing some non-existing HEPiX charter and suggesting we were stepping on to a "slippery slope", possibly leading to the "dissolution" of HEPiX. I therefore proposed and it was agreed to defer any discussion this week and allow time for reflection.
- Most CERN presentations were using the standard IT or EGEE templates but there were a couple of exceptions and perhaps GLs should check the overheads before submission.
- Several invited talks were given by speakers who flew in solely for their talk – the OpenAFS gurus, an Oracle speaker, and others. And there was a most successful presentation from Fermilab by Don Holmgren on Lattice QCD IT challenges. Such initiatives enrich the meeting.
- New European co-chair – Michel Jouvin of LAL, taking over from Wojciech Wojcik.
- Next meeting – CERN, May 5th to 9th. No site yet for meetings beyond that.

Site Reports

DAPNIA: new project – CARIOCAS, interconnect 4 French sites via 40Gbit links (actually 4x10Gbit) and demonstrate interworking. Another project is in astrophysics – COAST – which creates large data samples to be analysed in the DAPNIA high performance cluster, an IBM 24 node cluster interconnected by Infiniband. It should be doubled later this year. DAPNIA also participates in the GRIF LCG distributed Tier 2 spread across Paris; the Dell PCs have been decommissioned and they are borrowing some HP nodes from CARIOCAS until newly-ordered HP nodes of their own are installed; there is also a third disc server with 6TB of disc space. On the Windows side, they have decided to use Openoffice rather than Office 2007 and for the moment, new Windows PCs are re-installed with Windows XP.

SLAC: BaBar is about to enter its final physics run, hoping to double its data sample and this could cause serious bottlenecks if it happens. The GLAST experiment should launch in May 2008 and the target is to produce first results virtually online (to publish any interesting data within hours of download from the satellite). Being an ATLAS Tier 2 site, first data should also arrive at SLAC around May. Summer in SLAC could be quite busy. Plus there is the usual, and steadily increasing, activity from ILC studies, KIPAC and LCLS (see previous HEPiX reports for these projects). XLAM and PULSE are two new smaller projects featuring different sciences for SLAC. On the admin side, there is a large management and operations review in progress requested by the DoE. As previously reported, SLAC's computer centre is running out of capacity and they are expanding using Sun Black Boxes (see separate talk) and by replacing older nodes by more modern and more power-efficient servers; e.g. 11 Sun Thumpers replacing 40 old file servers, 100 quad-core PCs replacing 900 old servers. They are also planning modular new data centres. SLAC has suffered strange power problems recently (random node power offs) and after some searching have found high-order harmonics in many power supplies and they have installed filters to reduce the effect. There has been more capacity upgrade, mostly with Opterons but BaBar also installed 126 dual-dual Xeons. During an ssh upgrade, they dropped AFS token passing support, replacing it by GSSAPI ticket passing; they are working to eliminate the last vestiges of Kerberos 4. On the Windows side, they have used the appearance of Sharepoint 2007 to promote more collaboration activities, doc sharing, calendars, etc. No current plans for Vista.

LAL/GRIF: mostly for GRIF, LAL has added 250 quad-core PCs, funded by a new, non-HEP partner, Institut des Sytèmes Complexes. A new machine room has also been added next door to the existing one and the two may

later be merged. Also added 24TB of disc served by two SUN X4500s (Thumper). Networking is based on a 10Gbit crossbar switch and the external links between GRIF sites use 5Gbit links, with plans for a 10Gbit "GRIF OPN[1]" network in the pipeline. In the coming year, across the 5 sites (now 6), GRIF plans to add 1500kS12K of CPU and 500TB of disc space, more than doubling existing capacity. Among the scaling challenges this expansion creates are the possible need for a single SE[2] and which batch system should be used to support up to 2000 concurrent jobs. The EDGES project has been accepted under the EU FP7 programme for research into integrating institutional grids (EGEE) and desktop grids (BOINC, XtremWeb) and LAL will hire 2 staff to work on this. LAL also participates in CARIOCAS as described earlier.

TRIUMF: First post-Corrie Kost site report since his retirement in June after 37 years at TRIUMF. TRIUMF has a new director, Nigel Lockyer, ex-head of CDF. TRIUMF has had funding guaranteed for their dedicated Tier 1 centre for the next 5 years which includes major annual hardware upgrades to supply 7% of ATLAS computing resources and 9 new staff. Of course the available space is limited and cooling is a major issue and, like many sites, they have adopted alternate hot (40°C) and cool (20°C) aisles but they believe they have enough space until 2011. Their latest upgrade consisted of IBM 12 blade chassis using dual Woodcrest CPUs, 280 nodes in all. Storage is based on a dCache pool. They have a 5Gbit path to CERN with two backup links, one a 1Gbit path also to CERN and a new 1Gb path via BNL.

RAL: another new name – STFC – Science and Technology Facility Council where RAL is now an STFC site. RAL has celebrated its 50th anniversary. New computing building is under construction (see later talk) shared between UK Tier 1, RAL HPC and other STFC facilities. Should hold 300 racks and 3-4 tape silos. Now in the middle of the 2007-8 procurement for both disc (169 9TB servers) and CPU (1000S12K). The expected second tape silo should now arrive mid-late 2008. The other major STFC site (Daresbury) is trialing mail servers based on greylisting – initial email contact is rejected but accepted on retry, based on the principle that spammers don't retry. It appears to work (reduction of spam) but causes delay in delivery of legitimate mail. Other smaller STFC sites use a commercial anti-spam service which is also effective but fuzzy name matching is lost. Scientific Linux (SL) 3 is being steadily phased out in favour of SL4. The speaker (Martin Bly) appreciated the much improved support from CERN for CASTOR. RAL uses NAGIOS for node monitoring and continues to have some stability problems although it is improving. When tendering for new equipment, RAL do not test proposed configurations but use vendors' numbers, for example for power consumption. Comparisons were done with CERN procedures with the suggestion that there could be grounds for sharing such tests.

INFN: use of GPFS is increasing as their preferred cluster file system; release 3.2 scales better than before and an interesting feature allows to define external storage pools to expand pool data to tape. Disks and tapes could then both be managed by GPFS with different policies. Today, each INFN site handles its own mail service but a working group is looking at a single INFN mail service. Windows Vista is on hold but under study by the INFN Windows Group. Unlike most other sites reporting, there has been relatively little new equipment this year but much more is expected in 2008 for LHC startup. TRIP, a project reported on last year to provide a common wireless authentication infrastructure throughout INFN for roaming users, is in production and the current issue is to create a monitoring tool for when it occasionally fails.

GridKa: Apart from the array of HEP and other experiments supported, GridKa is now a host of the new D-Grid. All 980 dual-CPU, mostly dual core nodes are running SL4 in 32 bit mode. They have 3 10Gbit OPN links to CERN, CNAF and IN2P3. Procurement has started for LT04 devices. Last node expansion was in April with the addition of 168 worker nodes but they are not currently running because of local disc problems due to silent data corruption. Replacing the system disc in some nodes with another model did not improve the situation and the vendor is now

[1] Optical Private Network, as used between LCG Tier 0 and Tier 1 sites
[2] LCG Storage Element

expected to replace all system discs. The next acquisition is in progress, another 320 worker nodes and more nodes are planned for D-Grid.

DESY: with HERA shutting down, DESY faces a shift in user composition with the arrival of photon physics and laser physics with the XFEL project – X-Ray Free Electron Laser – and FLASH – Free Electron Laser in Hamburg. These produce new challenges for the computing department. DESY also participates in LCG and other grids for HERA, for ILC and also for some smaller local experiments. Between the Hamburg and Zeuthen sites, DESY is building a "National Analysis Facility" to provide additional resources to the LHC experiments. It will be independent of the existing facilities and will be by design flexible in order to cope with whatever data first arrives. The batch engine will be SUN Grid Engine and data access can be via AFS, dCache or Lustre, all of which are provided. It has a very ambitious timetable with a first working prototype by December. Compute nodes will be HP quad-core blade servers and the storage based on SUN Thumpers, all linked by Infiniband. On the Windows side, they are ready for Vista but feel under no pressure to roll it out. In addition their preferred install tool, Netinstall, is not yet Vista-ready. Similarly, no plans yet for Office 2007. The XFEL project reported many problems trying to use MS Project 2007 Enterprise. In preparing for the collaborations of their new user population, DESY is undertaking a notable investment in upgrading their audio-visual conferencing system.

NDGF: first site report from the Nordic Distributed Grid Facilities. He admitted that a year ago, NDGF consisted of a couple of managers but no staff! Things have clearly improved since with the hiring of Operations and Development Teams, spread across 8 production sites. There is a range of operating systems with AIX and Solaris in addition to various flavours of Linux. Apart from supporting LCG as a Tier 1, NDGF supports a Nordic Biogrid (under development) and a project on CO2 sequestration. The speaker ended by describing activities at his own site, HPC2N in northern Sweden. They are installing a new computer room and adopting the hot/cold aisle approach mentioned in the TRIUMF site report.

PDSF: PDSF, hosted at NERSC, is undergoing some restructuring with staff changes and changes in their client base. One new group is Planck for whom an Infiniband sub-cluster will be appended to their main cluster. Most of the main cluster is aging and fit for renewal, starting in December. But PDSF retains the same philosophy – user groups pay their hardware and share in a common cluster.

JLAB: The expected upgrade to 12 GEV has been delayed but it is still hoped for. The management of the site is now a collaboration of commercial and academic organizations and this has brought in more business-oriented thinking. Experience so far, one year in, is positive. There is a new, Infiniband-based, cluster for HPC consisting of 396 AMD processors, initially dual-dual nodes but 20% of the purchase was held back for trial upgrade to dual-quads. Since this upgrade worked, all nodes have since been successfully (so far) upgraded on site to a total of 3200 processors. The previous HPC nodes have been added to the Batch farm environment but this forced them for financial reasons to move from LSF[3] to PBS plus they are now faced with the familiar power efficiency challenge. Also dropped is support for Panasas and they have moved to SUN Thumpers for disc storage. Tape silos are almost 90% full but there are no funds for more until 2010 so they need to decide how to handle further growth. In general IT, they have adopted two-factor authentication for sys admins with Smartcards on Windows and CryptoCard tokens on Linux and Solaris. Vista is in the early support stage, especially for new purchases. For the future, they are reconsidering connecting to OSG or the ILC grid but they have been fully occupied for the past year coping with the above-mentioned reorganization.

GSI: new international Facility for Anti-proton and Ion Research – FAIR; 14 countries from all over the world, first beam 2012. The official kick-off ceremony is tomorrow afternoon as this is written, on 7th November. Windows group are promoting virtual servers. After the network problems reported at the last meeting, things have stabilized after the problem was traced to a faulty installation in the spanning tree of their new Foundry

[3] But in answer to my question, Sandy Philpott said that they made no attempt to negotiate with Platform because their batch service manager simply decided to move to PBS.

backbone. ALICE are running production jobs at GSI. GSI is a Linux Debian site but some glite services were hard to make work on this so they have implemented SL servers as virtual machines on top of Debian. They still use greylisting, concentrating on known spam-generating sources and they believe this is effective.

CERN: Helge summarized the structure changes planned for the physics side of IT. He noted the plans to replace CVS by Subversion and the infrastructure concerns of the Computer Centre where cold aisle confinement has started in the main room to improve the efficiency of the hot/cold aisle principle. A particularly pressing issue is to find enough back-up power for so-called critical services where requests already exceed available supply. He noted the recorded 100% efficiency of CERN's LCG production in September. All capable systems installed or re-installed currently are installed in SLC 4 64 bit mode but almost no apps are using this yet. He reported on the recent and planned hardware procurements. He has been impressed by the power efficiency of the latest acquisitions, even in comparison to blades although blades are occasionally tested for comparison. He listed various hardware problems met and solved, see slides for details. He noted the difficulty of supporting Linux on laptops where SLC 4 cannot work on most if not all modern models. He explained that the NICE password was now used for so many services that it is now called the CERN password.

Genome Sequencing Centre: part of the 12x12 block medical campus of Washington University in St. Louis. Some 75 Debian Linux servers are used for data input via barcode scanners but most data comes from DNA sequencing instruments via 235 systems, mostly Windows. Data is stored in Oracle databases, the raw data for older systems but for the newer devices which produce much more data the raw data is stored in files and only the tracking data is stored in Oracle. In order to qualify for Oracle support, the database RAC servers run Redhat Linux. They use LSF for batch scheduling and they add personal workstations to the batch service overnight. Some applications have huge memory needs and they have some systems with 4 Itanium chips and 96GB of memory.

Sanger Institute: also a Genome Centre, this one in the UK funded by the Wellcome Trust. He noted that the science may be different but the IT problems are the same and managing exponential growth is getting steadily more difficult. New data centre consisting of 4 250 sq.m rooms with 3.4MW power supply. One room is left empty so that as air conditioning becomes more efficient they can expand in there and refurbish other rooms in cycle. Big fan of blade systems – 3800 cores in 1500 nodes. Debian Linux is primary O/S but still some legacy Tru64 Alpha systems which still remain to migrate. Usual array of management tools – Ganga and Nagios, RequestTracker handling some 30,000 tickets per year. Also a major Oracle user: Sanger holds a public genome database, currently 60TB but doubling every 12 months and this has been moved to a 4 node Oracle RAC server. Despite promises by Oracle that they can handle terascale databases, Sanger has found that testing at this scale has been limited, even by Oracle. Storage is a SAN fabric with HP arrays. The latest sequencing machines produce 20-30TB per day so they have been forced, against their better judgment, to build a 320TB Lustre staging area (enough for 2 weeks raw data) and a 256 node HP blade compute farm to support these but they are unsure where and how to store the processed data in the long-term.

Windows Sessions

Windows Patching (Fermilab)

Previously, to connect to the Fermilab network a defined set of patches were required on a PC; patching was via SMS but individual groups could choose which patches to apply. Now, most systems use a central patching scheme, moving largely to WSUS although SMS is still used for inventory and reporting, mandatory patches and service packs and major O/S or application upgrades. Users may forgo patches for some time but for no more than a month (the Microsoft patch cycle). After patching, a user is not forced to reboot until the day before the next patch set becomes available. Patches can be applied only onsite or via VPN. Client groups have been created

to decide when and what to patch, for example a pilot group who test new patches or a "defer" group who cannot patch until a precise moment. Most users are in the General group. Experience shows that 80% of users update within 9 days of a new patch set becoming available.

Vista at CERN (Rafal Otto)

Rafal started with a most amusing story of why we should move to Vista in comparison with why cars can no longer be bought with cassette players as opposed to CD players; the longer we delay, the more effort it will be when we are eventually forced to do so – and we will be forced to do so eventually. Other reasons to adopt Vista include improved security (e.g. improved user account control, protected mode operation of IE and general running with reduced privileges), an improved graphics interface (Windows Aero), file shadowing (allowing access to the previous version of a file), improved instant search. On the downside, it requires notably more hardware, especially memory where CERN recommends at least 2GB (1.5GB if the graphics card of the PC has its own memory). Another issue is that an installation requires 6-monthly re-activation, done online. Various work-arounds are needed for laptops and for PCs on isolated networks (such as CERN's Technical Network). The default O/S remains XP and Vista is only supported on particular models, only those sold within the last year or so from the CERN Stores. A user can perform an automated check for Vista compatibility. The current Vista population is around 200 and growing slowly and at some point in the coming months, perhaps in conjunction with Service Pack 1, it will become the default for new PCs. Rafal ended with a short description of what's new with Office 2007. There was then a lively question and answer session which Rafal handled well, being able to answer all direct questions and not being provoked by a couple which questioned going to Vista at all.

Fermilab's Experience with the Vista Key Manager Service (KMS)

Also used for Windows 2008, this is the main licence activation mechanism for Vista. The speaker described its main features. It is easy to install but has a minimum population of 25 active activations and it needs a MOM server to obtain any useful reports although a site can build its own tool for this. In the questions session Siroli reported that INFN is planning a Vista test across all INFN sites and so would have to permit wide-area access via the port reserved for KMS which could have security implications.

Advanced Group Policy Management (DESY)

The speaker presented some details about the key features of GPM, illuminated (?) by almost unreadable screen shots of the tool. He noted its unforgiving nature – changes are immediate and there is no mode to evaluate the effect of a change before applying it. GPM permits delegation of the editing, reviewing or approval of policy changes but there can be race conditions if two people try to change policies at the same time.

Sharepoint Deployment at CERN (Alex Lossent)

It is described as a platform for collaborative applications and is integrated in Office 2007 but is usable with all modern web browsers. It allows to replace nntp, mail list archiving and Frontpage. It also offers other useful features wikis, web forms. RSS, support for external accounts, etc. He presented a number of use cases such as discussion forums, web questionnaires, web forms, calendars, documentation storage, community sites, etc. In particular he showed how CERN's HR, mainly James Purvis, are making substantial use of it for various tasks. He noted the remarkable success of IT's IS Group education campaign on Sharepoint. Interfaces they have added include one to permit editing from non-IE browsers, to CERN's choice of intranet search engine (FAST),to Indico (via RSS) and others. They have built a "CERN official theme" to be applied by default to new web sites with CERN's official colours and logo. He presented the technical details about Sharepoint's installation and operation. Future plans include links to web Single Sign On and WebDAV integration for file manipulation from any platform.

Storage Systems

Lustre at GSI

This talk was described by the speaker himself as a "graveyard of numbers" and indeed the audience spotted some obvious typos in his latter slides which he promised to correct before publication. Currently GSI is NFS-based but NFS has many problems with the scaling required for the future FAIR programme (see GSI site report), is poor at parallel I/O and has weak error recovery from network incidents. So they decided to evaluate Lustre as a cluster file system. On paper, Lustre appeared to reply to their requirements and had a number of good reports, including one from CEA at a previous HEPiX. Although open source, there is also commercial support available. The speaker listed the main features. The client does not need any kernel patches but the server does, at least until end-2008 when this is expected not to be required. Since SUN bought Lustre last month, it is expected that ZFS will appear at some time as the underlying file system but there is no confirmation of this. He showed the configuration of the test hardware/software used which included 26 multi-CPU client nodes. Performance tests show that XFS is much better in read and write than ext3 and that having 8 discs can show up to 30% more performance than 6 discs. Kernel parameter tuning does not affect write speeds much but can make read more than twice as fast; and RAID 5 is 30% faster than RAID 6. Also note that the cluster controller can be the bottleneck; the newest SATA controllers showed a dramatic improvement. In multiple client tests, he showed that Lustre can saturate a 1Gb network connection to a node and overall the setup is limited by the network interconnect. Lustre also passed disaster recovery tests in fail-out mode. Next they plan tests with real users.

HEPiX File System Working Group Progress Report (Andre Maslennikov**)**

The Working Group was setup in response to a request from IHEPCCC to study available file systems solutions and storage access methods. Members were appointed by IT managers at participating sites and the full list numbers 20 people. A total of 9 phone conferences were held and a first report was presented at a previous HEPiX. Work since then has concentrated on shared home directory systems such as AFS and NFS; and large-scale shared areas suitable for batch forms where examples include GPFS, Lustre, dCache, CASTOR, xrootd and HPSS. The collected information will be analysed and published in a technology tracking web site hosted at CASPUR and referenced from the HEPiX home page at FNAL. The site is intended as a storage reference site for HEP, not another Wikipedia. He encouraged more people to volunteer material and paid tribute to Charles Curran for his contribution on tapes.He then presented some general trends such as the growth of use of Lustre and GPFS for transparent file access and particular trends noted in LCG Tier 1 and Tier 2 sites. Before the next HEPIX when the final report is due, the working group will try to perform an assessment of some mixed disc/tape data access solutions for Tier 1 sites and various comparative analyses of various storage access methods for Tier 2 sites.

NERSC Storage Update

NERSC uses GPFS with and without local storage as the site's global file system; there is also limited AFS (client only) while NFS is being phased out. In the forthcoming roll-out of GPFS 3.2, it was noted that although this version can run in parallel with 3.1, 3.1 cannot see file systems created by 3.2! For new hardware, they are looking for a GPFS replacement and negotiations are ongoing. Any change of product would mean copying almost 200TB of data.

Lustre HSM Project (CEA)

Lustre's most relevent feature for this project was the fact that it is aimed at high performance systems. The intention is to add an HSM layer to Lustre but the "offline storage" will not be tape but an external storage system. Copy to external storage should be transparent and all files should always be visible. The speaker presented the design inside Lustre and the various Lustre HSM components such as initiators, agents, a space manager and so on. The project is a collaboration with SUN with a joint design of the architecture but all the coding to be done by CEA. So far, the architecture is made and the design documents should be ready soon.

Native Infiniband Storage (from the VP of Data Direct Networks**)**
No longer a fight between SCSI and iSCSI extensions to RDMA[4] protocol, both can co-exist. Most vendors promote both. SRP, SCSI over Infiniband is similar to FCP (SCSI over fibre channel) except that the get/put data address is included in the CMS information unit. iSER uses iSCSI which offers a more global solution. He showed, at rather high speed, a number of different storage solutions using one or other protocol. SRP has advantages for direct server connections; iSER is better for large switch-connected networks and supports device discovery. His company, he claims, is looking into supporting 1000+ disk arrays and parallel accessing. RAID 6 is becoming more popular. He is also aware that low-overhead parity checking is becoming necessary to avoid silent data corruption. For existing fibre channel storage networks, he says IB will mix and match and IB and fibre can live quite happily together.

Chimera (DESY**)**
Pnfs is the current name space and meta data provider for dCache. But it is becoming problematic as data volumes increase and Chimera is being looked at as a substitute. Pnfs only stores meta-data and is somehow independent of dCache and, using the current NFS-based protocol, there is a "long way" from dCache to Pnfs data and this is one of the bottlenecks today. Also, any file access locks the whole Pnfs database and that does not scale. Despite these and other drawbacks, US CMS uses it for a filebase of some 1PB. Chimera provides similar functionality towards dCache; it is only an API and a database table layout, but not a database itself; so it scales much better, in fact with the performance of the backend database. Whereas Pnfs spoke to dCache only through NFS, another bottleneck, Chimera speaks directly with dCache; also performance is independent of the number of files per directory unlike Pnfs; Chimera can distinguish between read and write so the locking problem is diminished; and Chimera can take advantage of any performance enhancement features offered by the backend database. In a direct comparison, Chimera showed a factor 10 improvement. It is available for dCache 1.8 and at least two Tier 2 sites are testing it since some months. Migration tools from Pnfs to Chimera are available.

The second part of the talk concerned NFS 4.1, an extension to NFS 4 which is aware of the fact that the back end storage system may have the same file stored on a set of different servers (pNFS, not to be confused with Pnfs). The specifications are in the final phase of definition and he discussed the major advantages such as awareness of distributed data (as dCache); it is faster; and it supports ACLs at a file level. NFS 4.1 would make data distribution from dCache easier without the user having to deploy a dCache client.

OpenAFS Status and Futures
Two AFS gatekeepers and the leaders of the OpenAFS open source project, Derrick Brashear and Jeffrey Altman, first presented their roadshow, what is OpenAFS, where can you use it, what makes it unique, etc. Future enhancements to the cache manager include read-write disconnected mode (for laptops), automatic tuning of the cache size and a native redirector client and UNICODE code on Windows. In passing they noted that the Windows version is already Vista and Server 2008 compliant. They then moved on to a roadmap of future plans for further developments, for example how they could perhaps further improve performance of the Windows implementation. Other planned work includes an improved human interface. See the link above for the full details. Some of the planned improvements are funded by Microsoft and/or Apple and they are trying to get the DoE to fund others. There was a long list of wished-for improvements. However, when Andras asked when a new read-write replication feature could be implemented and delivered, one of the speakers said it needed funding and when asked how much, he mentioned the figure of $100K. This discussion carried on into the evening and is reported on in the introduction to this report.

Data Corruption in the Enterprise (Oracle)
This talk covers non-malicious loss of data; possible causes are O/S bugs, hardware or firmware bugs, admin

[4] Remote Direct Memory Access, enables zero-copying memory access via the network, no CPU, cache or context-switching overhead, low latency

errors. After an incident, you need to find a good copy and this often happens under strain, so avoidance is clearly better. You could add protection metadata to the data such as CRC, parity tags, etc. Silent data corruption is where the ECC check on disc does not detect an error and leaves the application to deal with the corrupted data. Oracle's "checksum on the data" feature will permit to detect corruption but does not prevent it. However, if the storage device could understand the Oracle data block structure, it could prevent corrupted data being written and Oracle HARD is a scheme to licence this structure to hardware vendors. Another scheme is the T10 Protection Information Model which adds protection metadata to the data which then allows data integrity checks along the path to disc. But T10 does not span back to the application nor does it address host-oriented failures. On the other hand, Oracle HARD spans to the application but does not span to the drive and is Oracle only. The newly-formed Data Integrity Initiative is a group of vendors who joined forces recently to work on an end-to-end solution by extending the T10 standard back to the application. DII will also discuss how to pass the protection metadata through the O/S to the application. Plus it should also extend beyond Oracle. For further study, the speaker, from Oracle US, referred to work done by Bernd and others at CERN.

Data Centres

Introduction (Tony Cass, topic convener)
Tony presented the challenges of operating computer centre as power efficiency improvements lag behind performance improvements and the required growth in box numbers in preparation for LHC data. Tony estimates that CERN operates a power density of 30kW/sq.m; can this be cooled? Yes with water cooling and also yes if air cooling is carefully set up. But he noted that it requires some 400W of input power to a Centre to supply 100W for the servers and so any savings in cooling can have a multiplier effect. As alternatives to building a new centre, he sees 3 alternatives – hosting (too expensive), SUN-type black boxes (see SLAC talk) and virtualization to get more return on investment rather than duplication of under-used servers.

BNL (Tony Chan)
Currently hosting 9M S12K in CPU power, 3PB of disc space and 7PB of tape space and they have noted exponential growth recently. Floor space was actually maxed out in 2006 but replacement of less efficient servers by dual and quad core systems has avoided crisis currently but only for a couple of years, say until 2009. Similarly, UPS capacity is reaching limit. He admitted that in purchases early in this decade, they forgot to plan for more and better air conditioning and delivery times for improvements of these can be long and disruptive. Finally, the price BNL pays for electricity is gradually rising. As a result, recent purchases pay more attention to power efficiency. Other improvements include :-

- Some large part of the existing centre is being renovated and modernized during the next 12 months
- There are plans for a new and larger centre but it will arrive only in 2009. It will have a notably higher and reinforced raised floor (36 inches rather than 12 as now) for a higher air floor. The design will also include facilities for better cable management and consequently better air cooling.
- In the meantime, they are applying rack-top cooling units for some hot spots.
- And they are adding more modular UPS boxes.

Apart from deploying multi-core processors, they are looking at blade servers, DC-powered servers and the other alternatives mentioned by Tony Cass in his introduction. Dual-core AMDs showed a 20% power saving over previous generation of CPU and they expect further benefits from quad core processors. According to vendor specifications, blades should offer a 20% power saving per Specint compared to 1U servers but BNL are sceptical and plan tests with real applications. DC systems have steep up-front costs if the power supply must be DC and show insignificant savings if you use rectifiers. Use of xen for virtualization does offer some savings and initial deployment is starting after extensive tests but it does not work for all applications. BNL looked at mobile data

centres but do not consider it a solution for them – issues with sensitive data (BNL does more than HEP physics) and the power and cooling problems remain.

Longer term, beyond 2014 when they predict that even the new building described above will be full, they have plans for a 25,000 sq.ft. centre to serve all of BNL but it is not funded yet.

Genome Sequencing Centre Plans

The new Centre now under construction is targeted to satisfy their needs for the next 5 to 7 years, enough they hope, despite rapidly increasing data quantities being produced by the most recent and expected generation of DNA sequencer. The speaker showed the assumptions used in planning needs over the coming 5-7 years – two-thirds of racks containing discs at 8kW/rack and the remaining third with CPUs at 25kW/rack with a building capacity of at least 120 racks. Other considerations are to avoid single points of failure. For example, all electrical paths are duplicated, including dual UPS, one battery, one flywheel; the flywheel offers only 15 seconds of power until the generator switches in. And being US, there are earthquake precautions. Despite a total floor space of some 16,000 sq.ft, only some 3200 sq.ft are usable for servers, the rest being needed for electrical and cooling equipment.

NERSC Data Centre

The first part of the talk covered a comparison of AC and DC powered systems. The DC power was distributed at the facility and rack level. The results showed a 9% improvement at the point of distribution, but a lower percentage downstream. The result claimed a total of 10-20% (now clear how this number is arrived at) overall for the facility according to the author of the overheads (which were given by a different speaker). A fuller report is available from the original author.

The second part concerned plans for a new, green, data centre, being planned for up the hill in Berkeley. The "green" aspects are driven by something called LEED standards. They plan to use as much natural cooling, using external air temperatures for most of the year, since it gets quite cold in Berkeley in winter, and additional cooling in hot days. Final approval should be soon and commissioning and equipment moves are expected in 2011.

RAL's New Computer Centre

RAL had decided already last year that their existing rather old centres were not large enough for planned expansion. During the planning, they used fluid dynamic modeling to evaluate the effect of various layouts of cabling, raised floors, power distribution, etc. This modeling is expensive but judged worthwhile. The speaker, Martin Bly, noted the difficulty of finding consultants capable of understanding the issues in building computer centres and those who do, including a number of PC vendors, tend to be expensive. Building location had to take account of power feeds, various existing underground ducts and the existing centre(s). He then described the details of the planning RAL went through. RAL's requirements were for 300 racks and 4 tape silos, enough capacity for the RAL Tier 1 until at least 2012. They expect that eventually air cooling may not be sufficient so they have included facilities for water cooling. A part of the room will be separated for quiet running (tape silos). They are looking at a Combined Heat and Cooling unit which lowers the carbon footprint and could offer UPS support but so far it appears not to be economical so no decision has been taken yet.

SUN Black Boxes at SLAC (Chuck Boeheim**)**

BaBar needed 250 more servers at short notice but SLAC had no power or space available to satisfy this so chose to take advantage of SUN's new Black Box project – recycled shipping containers equipped as self-contained computer rooms (except for the power and water source) with up to around 250 servers. He took us through the features of the box and showed a web-cam series of photos of the installation process – in fact the very first installation of one of these units. First power was applied on 17th Sept and production running a week later. He then showed a series of photos of how to enter the box for some maintenance (non-trivial!) and how to extract a rack (weaklings need not apply!). After some months experience, SLAC are still considering their options regarding a second one, perhaps with SUN equipment but perhaps not.

Large Scale Remote Management via IPMI (Andras Horvath)
In 2003, CERN deployed serial consoles based on ideas from SLAC but this does not really scale. Today IPMI[5] seems to offer a better scheme and modern PCs come equipped for this. It includes power control so we could link them all to a single power control button to effect an orderly shutdown in emergencies. Programming the power control features could also permit the switching off of unneeded capacity or controlled sequencing of reboot after a general shutdown. We monitor various thermal and voltage alarms, ECC and parity errors, etc. IPMI is relatively new and not all implementations have proven to work; each vendor's version must be tested. Andras also listed some tips and tricks he had discovered. Access control is via randomly-generated username/passwords. Some vendors are working on similar schemes (Intel's Active Management Technology) but it appears too complex at this time so for now and the foreseeable future, CERN will use IPMI for remote management of its computer centre.

Virtualisation

Introduction
Veronique Lefebure, one of the co-conveners of this topic, introduced the subject by listing various benefits and reasons behind recent interest in the subject, especially in running various grid services but in many other environments also.

Operational Aspects (Alex Iribarren)
Standard scalability issues apply, the systems may be virtual but the problems are not and fall into three classes.

1. Scalability: for example, the more (virtual) machines, the more (real) management load; the more (virtual) machines on a particular node, the more (real) power that node needs; and so on.
2. Invalidated assumptions: having a non-physical node to run on, are there implications for the application? No MAC address or system serial number, no possibility of hardware keys, etc. For asset tracking, do you know where the physical node is? If an intervention is needed on a (physical) host, do you know which (virtual) services are going to be interrupted.
3. New features = new problems: debugging of virtual machines is more difficult. There will be new special virtualized configurations to manager and new host O/S to be supported (such as xen). Dynamic provisioning has security implications. Image management has many open questions about creation, security, storage, etc).

Alex's summary is that virtualization has definite benefits but must be carefully thought out in advance.

High Availability Cluster of VMs (INFN)
From a Tier 2 site in Rome, member of EGEE SA1 the speaker listed the various services required or encouraged to be run by Tier 2 sites. The classical configuration offers no consolidated failover and thus no high availability. There are 2 goals in this project – high availability of standard services and dynamic provisioning of worker nodes. He described how they had built virtual machines for the core services focusing on high availability as well as load balancing and showed 2 prototypes focusing on each of these. They found load balancing non-trivial in a grid environment and this needs more research. For the second goal, a scheme is built round a provisioning server. He showed the architecture and implementation details and some screenshots to illustrate its various features and benefits. They are quite pleased with the results so far and intend to conduct more tests and run benchmarks, hopefully eventually providing packaged solutions.

[5] Intelligent Platform Management Interface

Virtualisation at Karlsruhe
He started by listing some of the features of virtualisation. It was ironic that Alex had already mentioned most of them with their associated problems but this speaker only mentioned the positive aspects of each. He presented some features of the (open source) xen and how this is implemented at FZK and the University of Karlsruhe. At the latter, they have implemented a high availability scheme and he described that. As part of the University's Tier 3 site, they are implementing dynamic cluster partitioning using virtualization on a 200 node shared in-house cluster. Tests showed a loss in performance of only 3-5% on their test cluster. The boot time of a VM is long but considered acceptable to be able to participate in the shared cluster. Another project, this one at FZK, is to build a model of the grid workflow on virtual machines using GVE (Grid Virtualisation Engine) created in FZK. GVE can use xen or VMware as the host.

EGEE Services and WN[6]s using VMs (CESNET)
The MetaCenter in the Czech Republic participates in 2 grids, one of them EGEE, and virtualization makes this much easier as well as offering the usual benefits of VMs. They use xen and Vserver and he compared these: xen offers complete encapsulation, Vserver has a single kernel space; xen is perfect for service consolidation and support for complete linux distributions while Vserver has lower (in fact no) performance penalty and better memory management. Xen results were judged good on small SMP machines but bad on fast Ethernet (but good on Infiniband). He described the EGEE use case; job submission requires a small modification to PBS but no change to EGEE software.

HA at Fermilab
Various services on Fermilab associated with authorization/authentication (VOMS, GUMS, SAZ) are single points of failure and this led to a high availability project. They based the work on xen on SL5 (but not the xen which comes in the Redhat package). They debated between active-active HA or active-standby. The latter is easier to implement but risks lost transactions during failover creating inconsistencies so they chose the former where the risks are lower. Other challenges were DNS or LVS[7] and does MsSQL work (yes with the correct version (V5.0 or higher). See slides for how these choices were decided. He showed the initial results of stress tests which show good performance but also room and areas for improvement. This is continuing and initial deployment is planned for December. As targeted, they should be able to reduce the number of boxes needed for these services and they may extend this to other OSG and/or Fermigrid services.

Virtualisation at CERN (Jan Michael)
Once again, xen is the preferred tool. He started by listing operational issues associated with networking (providing virtual MAC addresses and hostnames), configurations (quattor components), monitoring (lemon sensor), etc. He showed the details of two use cases: gLite certification which has some 15 active users and has been heavily in use for about a year; and ETICS where 10% of the worker nodes are virtualized so far; ETICS uses VMware because Condor currently does not integrate to xen.

Globus Virtual Workspaces
They would like to offer advanced job scheduling features such as advanced reservations, periodic execution, and so on. Further, some potential applications are difficult to install. It is felt that a dynamic provisional environment through virtualization could help with both of these. Again xen is the tool of choice. The speaker described the workflow of deploying workspaces on a cluster via a workspace manager front-end. This technology has been integrated into STAR[8] production running. They are also working on a workspace pilot as a Condor glide-in to claim resources but not (yet) run a job. Having deployed a set of virtual workspaces, what next? Could they be

[6] Worker Nodes
[7] Linux Virtual Server
[8] BNL experiment

considered as a virtual cluster? From where come the images for these virtual workspaces (they assume the appliance suppliers). All these are under study and the pilot is currently happening on a couple of Teragrid sites.

HEP Applications with Globus Virtual Workspaces
Ian Gable of Victoria then presented a practical example of the previous talk. The motivation was potential access to nodes on a Canadian grid not otherwise available (wrong base O/S for example). The chosen application was the ATLAS distribution kit upon which they provide kits to permit users to build easily their application image. So far they have successfully run SL 4.5 images on SL 5 clusters but the image running on SuSE needs more work. They have done some security-related work such as sandboxing where the virtual workspace helps and image signing but more is needed in this area.

Grid Monitoring

Monitoring and Metrics at Fermigrid
Metrics collection takes place once per day, Monitoring is more frequent, typically hourly and used to check service availability or whether a site will accept grid jobs (does not guarantee they will run successfully). But the speaker did not explain why a given data item was considered a metric or a monitoring quantity in other than sampling frequency. He did however list what data was gathered under the "metrics" label and gave examples of the plots created from these. Similarly, he listed examples of data which is "monitored" and how this is used in problem discovery and resolution. Again he showed plots and dashboards created from these data. Among the lessons learned were that collecting enough but not too much data is a fine line, log formats differ and sometimes data must be extracted from multiple logs.

WLCG Monitoring Displays (James Casey)
Rule 1 of monitoring – you can't manage what you don't measure. James listed the various tools typically used for monitoring services at different levels – applications, middleware, local resources. However, getting these systems to talk to each other is difficult. In particular, grid site monitoring seemed to be lacking in many areas and there was little in common where it did exist. It was decided to build an easily-extensible scheme but not force it on sites with special plug-ins for grid services and make the resulting data readable by the standard grid monitors. The prototype was built round Nagios with the intention later to integrate it with Lemon. He showed the architecture, how they provide probes and interface them to Nagios, how they produced a Nagios configuration generator and how they published the data. To simplify visualization of the data, they have been working with EDS, a CERN openlab contributor. Graphic grid maps have been developed to allow different views of the monitoring data – geographic, VO, trends, site availability, etc. Within the grid maps, there are links allowing to drill down to investigate particular problems or issues.

LCG Application-level monitoring with the Experiment Dashboard (Benjamin Gaidioz)
This is an EGEE-funded project to create a framework to show the activities of the different VOs on the grid. He explained the framework. Data is retrieved via HTTP and it supports multiple output formats. There are API and CLI interfaces, both based on HTTP. There is a rather complete developers guide and all this makes it rather simple for users to write their own monitoring service. After the sessions the previous day, Benjamin now believes that dashboard applications could be good candidates for virtualization. He showed some examples of job monitoring but unfortunately the screen shots were unreadable. Like James's grid maps previously, one can drill down the dashboards to get more details right down to the user and job level.

Grid Security Update (Dave Kelsey)
Lots of work on grid authorization as opposed to authentication, for example trying to consolidate interoperable solutions and create longer-term standards, creating policies on who can run a VOMS, minimum trust requirements, etc. The EGEE Operational Security Coordination Team (OSCT) is now well established although

they are continually looking for additional experts to contribute. They have been performing some security service challenges. A major step is bi-lateral contacts between WLCG and OSG on security matters. The Joint Security Policy Group (JSPG) has been reviewing the various policy documents. The JSPG has expanded to include representatives of some newer, mostly smaller, grids. There is an EGEE group looking at grid vulnerability and risk assessments are underway. In the 2 years work so far, 122 issues have been identified, 62 were open at some point last month but only 1 was assessed as extremely critical.

Dave then moved to some current issues. For example, in order to run the pilot jobs demanded by some experiments, it is required to change the identity of the actual job to that of the user on execution. This is done by a module called glexec but it is seen as very contentious on some sites and has been, and still is, under discussion at all levels of EGEE.

OSG Centralised Logging

Unified logging requires a log file collection mechanism and a standard log file format or convertors for legacy formats. An existing open source product, syslog-ng, was already available for log collection and the speaker listed its features and showed the deployment used for OSG. They have defined some bes practices such all logs should contain a unique event name, an ISO-standard time stamp; the start and end times of any transfer and so on. Errors should be reported as the end event in the format STATUS=N where N is negative for failure.

Benchmarking

Introduction (Helge Meinhard, topic convener)

Helge listed some of the items we may consider under benchmarking. These now extend to power consumption and even perhaps air flow of systems. He cited the ongoing debate between the use of industry standard or home-grown tools. Currently there is a lot of emphasis on worker node performance and most people use SPEC benchmarks.

INFN's Migration to SPECint2006

Like most sites, INFN currently uses SPECint2000, usually represented as SI2K, but this is now declared obsolete by SPEC and replaced by CPI int 2006. The problem is that, for example, the LCG funding agencies have made promises based on SI2K but the vendors no longer publish these figures and there is no single conversion between the old and new scales, the range is 137 to 172. CERN and FZK have started to use a "new currency", SI2K as measured by gcc with various levels of optimization but this will only work for some time. This is because with an increasing number of cores the results become confusing because of the different effects in different processors when running fully loaded. Chip generations and chip vendors also affect the scaling between the old and new numbers. INFN performed some tests and compared the results and found that the CERN modified SI2K tests most closely resembled the SPECint2006 numbers published by the vendors. The result was to propose to INFN to stop using SPECint2000 tests; for the next year, use the CERN modified version but next year move to SPECint2006 and the best solution is SPEC INT 2006 RATE test measured by the gcc compiler; the RATE test is claimed to take account of scalability of multi-core architectures. This however must be agreed with the LCG funding agencies to convert their promises in SI2K units to the new units.

Benchmarking at FZK

Switched to SPECint2006, the RATE metric specifically, for the same reasons as expressed by the previous speaker. One advantage of the new suite is that the inner loops of some tests no longer fit in cache memory and any memory bottleneck will now show up. On the other hand, the new benchmark takes a full day rather than a couple of hours for SPECint2000 – is this an advantage or not? He then showed a series of graphs of tests against various models of Intel and AMD chips under various conditions. One set of plots showed clearly the effect of the

SPECint2006 tests not running in cache. This for the speaker would be yet another argument from moving away from the SPECint2000 tests which give too optimistic numbers for modern chips with large caches.

Benchmarking at CERN (Helge)
Regarding the move to SPEC2006, we are some way behind INFN and FZK; recent tests still use SPEC2000 and we have not made a decision for the next round of acquisition. For multi-core systems, we run as many parallel streams as cores as that is how these systems are run in production. Recent adjudications are based on specifying a SPECint2000 target and a penalty per box for power as well as charges representing infrastructure. Helge described some aspects of the formula used for adjudication in detail. Experience shows that some vendors lack experience in power rating, using estimates or specs rather than real measurements. Turning to disc servers, a similar power factor was added to the adjudication calculation. After our last acquisition, we submitted a LINPACK test to Supercomputing in June 20007 and achieved number 115 and we will submit a new set after the next forthcoming acquisition. Helge ended with a description of work being done now to measure SPEC power. Ideally at some time in the future we could demand that potential suppliers to run SPEC power with a defined workload.

Performance of Lattice QCD Codes
Presentation via audiolink from Fermilab. Lattice QCD is the numerical simulation of the QCD theory of the strong force. LQCD therefore makes heavy use of Monte Carlo simulations. First one generates gauge configurations which requires an extremely long single-stream job; they use a 10TFLOP system and it takes many months. Then the further simulation is embarrassingly parallel and can use smaller systems. The US has a national programme for QCDOC spread across a number of labs and using custom hardware for the parallel work. The current programme funds clusters at FNAL and JLAB. On a given platform the bottleneck could be memory, FP performance, etc.; currently it is memory bandwidth or network bandwidth so Infiniband clusters are popular. For memory bandwidth, the STREAMS copy benchmark is used and the results were shown in a table in the slides. Then followed some single node performance slides. The architecture of multi-core AMDs is that memory is attached to specific cores and if you access memory on another core, there is an extra transfer to do[9]; the effect of this was shown on a plot; with AMD they must use numactl to make sure to only do local memory accesses. The Intel shared memory linked by a single front-end bus has a better profile for their application, also shown graphically. Nevertheless, various tests show that the new AMD Barcelona quad core chip easily beats other current offerings although they hear that Intel will have interesting competitors coming round the corner (next week!).

Miscellaneous Sessions

Scientific Linux (Troy Dawson, FNAL)
Continual steady growth of user sites and installations, especially 4.x version but installations of V5, released in May, have started to appear. And although the number of Itanium sites are negligible, there is a definite growth in 64 bit sites. Version 4.5 was released in June with better support for xen among other improvements. They are preparing for the appearance from Redhat of 5.1 which actually happened this week so the first build is in progress as Troy speaks. He hopes for an Alpha release next week and a formal release shortly afterwards. Other current work is to move away from version 3 with a target of mid-Feb to “obsolete” versions of V3.1 through 3.8, leaving only 3.9 in legacy mode. Further out, there are plans for SL 6 when Redhat releases their V6 but there is no timetable for that. And of course they will track 4.x and 5.x.

High Density Visualisatons (Chuck Boeheim)
Chuck reminded us that a “cluster is a very large error amplifier”, how to scale the visualisation of monitoring data to quickly recognize and drill down to a problem? He showed examples from Ganglia where some white

[9] NUMA – non-uniform memory access

space is wasted and the relevant data is over-concentrated in small charts. According to some literature he hasd read good practice should allow the eye to find patterns, use all space, avoid bar charts and so on. An interesting tool he had found is Magnaview, a commercialization of Sequoiaview from the University of Eindhoven. Apparently a young firm but early experience is positive. It has a variety of display types – tree maps, colour maps, grids, pie charts and so and so on. A data set can be viewed in many different ways. Chuck then demonstrated with some screen shots which proved once again that art is a very personal thing – I found the screens probably more confusing than the original version but perhaps a learning cycle is needed to appreciate the new formats. One nice feature was that as you run the mouse over parts of the display with the display tool, you see the details of the next level down. Some display features work with common browsers but very dense screens overload most browsers. His conclusion that GUIs can scale and you can use all 18 million bits (pixels and colours) even on small screens. The product costs less than 2K CHF and after the meeting I learned that Peter van der Reest of DESY has already started negotiations to try to get a HEP licence.

ISSEG (Alan Silverman on behalf of ISSEG)
I presented the goals and deliverables of ISSEG based on a presentation put together by the ISSEG scheme. One question posed was why the project had used Excel with macros for the risk assessment questionnaire. After the talk, Siroli (CNAF) expressed interest for courses he has to give to a physics class and he would look in more detail at the site.

CERN Alerter (Rafal Otto)
CERN wanted a scheme build on modern standards to send messages to logged-in users. RSS was chosen as the transport protocol. He described the attributes of a message – content, behaviour (message urgency or validity periods for example) and scope (target audience). Urgent messages are popped-up immediately, important ones only at login or next morning, information only by a bubble to inform the user of its presence. Scope can be by building, group, etc. The architecture is based on a Sharepoint backend where the Manager on Duty enters the messages, the RSS protocol, Active Directory as the front-end, the CMF agent as the Windows client and any RSS-enabled browsers to display the messages. Non-Windows systems (Linux, Mac) have no special clients so using the browser, a user needs to subscribe to the appropriate RSS feed.

Single Sign On Update (Alex Lossent)
From a multitude of username/passwords, we are almost down to 2 – Windows which is used for many other services than simply NICE, and AFS. The target is still central identity and access management where the first defines who the user is, to what logical groups he/she belongs and the second defines what access he/she has. There are three components of IAM (Identity and Access Management) – AAA - authentication, authorization and accounting. A central authentication scheme has now been deployed and more and more applications are joining this (CERN's admin procedures, twiki service, etc). It offers single sign-on for web applications already and it support for smartcards. Linked to this there are plans to implement eGroups next year. There are also plans for an authorization scheme, the second pole of AAA. Alex then presented the technical background in a couple of overheads. For non-web applications, they provide a SOAP web service and the user needs to write a SOAP client for which they have built an interface based on the SOAP standard.

Alan Silverman

10th November 2007