HEPiX Spring 2006 Meeting

CASPUR, Rome

The meeting was held in the Italian National Research Council (CNR) Aula, a very comfortable auditorium although the networking should have been more stable than it was all week. Initially there were hardware problems but by mid-week it was the presence of locally broadcasting nodes in the room which created much instability and which could not be traced. A good argument in favour of MAC[1] address registration for such events.

Alongside the traditional HEPiX sessions, there were a number of special meetings such as the LCG[2] GDB[3], the OPN[4] working group and others so the total registration count was over 120 although not all were present all week. Unlike previous meetings, this one was mostly separated into topics with a convenor appointed for each topic. This new formula attracted many speakers and attendees for whom this was their first HEPiX. Also, as in the past two HEPiX meetings in Europe, this meeting attracted a noticeable number of representatives of LCG Tier 2 sites, from across Europe especially.

As usual this report is a personal set of notes and readers interested in particular points are referred to the overheads, all (or almost all) of which are online at http://hepix.caspur.it/spring2006/agenda.php. In the notes which follow I have tried to insert footnotes to expand abbreviations and links to web pages which explain more of the details.

Highlights

- Computer room cooling and air conditioning systems were mentioned in a majority of site reports, either at the rack or building level. Several sites are having to build or equip new computer rooms to get round capacity restrictions in existing facilities.
- As usual at recent HEPiX meetings, there were a number of benchmarks presented with very detailed overheads – well worth a look if you are interested in performance or costs.
- New format for HEPiX with half-day sessions on dedicated topics; such as networking, performance optimisation and databases were new to HEPiX, with corresponding invited speakers. The HEPiX board like this new scheme and will try to continue with it at future meetings.
- Collaboration on and re-use of HEP-developed tools was not particularly emphasised at this event although many examples were mentioned in passing as “normal business”. On the other hand, there were, as often the case, a few examples of wheels being re-invented for no obvious reason.
- Also some random tools CERN/IT might want to look at – Imperia for web page content management (PSI site report); Subversion, mentioned several times by DES Group as a possible replacement for CVS for code management, seems to have arrived on at least a couple of HEP sites.
- What to do about Bird flu – see Bob Cowles’s security talk at the end of this report
- Next meeting – Jefferson Lab, 9th October. Followed by DESY Hamburg in Spring 2007.

[1] Media Access Control address, the device’s hardware address
[2] LHC Computing Grid
[3] Grid Deployment Board
[4] Optical Private Network

The meeting was opened by Prof Romano Bizzari, the CASPUR Director. He explained very briefly the make-up and *raison-d'être* of CASPUR.

[5]

Site Reports

TRIUMF (Corrie Kost) – Since the last meeting, they have moved from acquiring white box PC to rack-mounted pizza boxes. They are also building up their connection to ATLAS towards 10Gb networking although they have already surpassed the LCG Service Challenge 3 (SC3) target of 100Mb/sec throughput sustained over 24 hours. They report problems with air conditioning and partly for this reason they are considering blades for ATLAS as well as possible water cooling such as the Modular Cooling System offering (MCS) from HP.

CASPUR (Andrei Maslennikov) – still running their usual collection of SMP[6] mainframes from IBM (new 21 node Power 5 cluster), HP (32 CPU Alpha EV 7-based), NEC and several Opteron clusters. The new IBM cluster was hard to install – many hardware problems and buggy AIX software. The Opteron clusters are popular and likely to grow in number and size. He also mentioned air condition and cooling worries. For batch scheduling they have moved from SUN Grid Engine (SGE) to PBS on the Opteron clusters and are looking at doing the same on the IBM system; reasons include better MPI support, more flexibility and a better fit with their accounting scheme. Storage – they have decommissioned IBM's Storage Tank and are moving to GPFS[7] which uses lower cost hardware plus they had a recent fatal crash of Storage Tank when they lost meta-data, luckily backed-up. There are two new NFS servers running Raid 6; two new Infortrend disc systems are under evaluation; and they have migrated from Brocade switches to QLogic. He described some ongoing collaborations with CERN and some other European labs. CASPUR is offering an AFS cell for HEPiX which is already maintaining an archive of past HEPiX meetings.

RAL (Martin Bly) – 21 new Opteron storage servers with 8TB of usable space per server. Added more than 200K SI2K[8] CPU capacity, again Opteron. Considerable interest in Oracle, testing it for the LCG 3D and FTS[9] applications. They have a monitoring project using NAGIOS with the intention to replace SURE later this spring. New STK 8500 robot was delivered in Dec and is equipped with some 9940B drives and 10 new T10000 drives under evaluation. CASTOR 2 is being deployed and will be built up as we approach LHC startup. They have a 10Gb backbone for the Tier 1 LAN and another to UKLight but will only have 10Gb link to CERN this summer. Similar to LCG Service Challenges, RAL has been doing UK GridPP service challenges in conjunction with their Tier 2 sites. The first of these severely disturbed the RAL network and was abandoned after only a few hours. Since then, their use of the network has been temporarily restricted and subsequent tests to individual Tier 2 sites have been somewhat more successful.

CERN (Helge Meinhard) – major activities have been LCG Service Challenges, in particular the successful second phase of SC3 where most of the target sustained rates were achieved, helped by a successful CASTOR 2 rollout. Lots of hardware acquisitions including 1200 farm PCs, 65 disc servers, 70 tape servers and IBM and STK robots for evaluation and more servers are due shortly, including the Computer Centre's first Opteron servers. Another first is that current tenders specify not a number of boxes but a number of Specints. Refurbishment of the Centre is nearing completion but not quite fast enough to cope with an unexpected rise in temperature just a week ago; we also suffered 2 power failures in January. These are perhaps to be expected when the refurbishment had to take place around a running

[5] Message Passing Interface
[6] Symmetric MultiProcessor
[7] General Parallel File System, from IBM
[8] SpecInt2000
[9] File Transfer Service

production centre. RedHat 7.3 Linux has been decommissioned and we plan to move from SLC3 to SLC4 later this year. He noted the recycling of old farm PCs for use with BOINC and listed several new possible applications which could use this architecture. Meanwhile, at last count LHC@HOME had 15K clients and had provided 163 CPU years for LHC beam simulation. Commodity servers are now heavily used for Oracle RAC[10] servers for physics and this service is growing rapidly. Two new 10Gb links to the US and the central LAN switches have been updated. Significant efforts on security have helped to reduce the number of incidents on site but it is an ongoing battle. The EGEE phase 2 project started this week; openlab phase II started in January and more partners are expected to join. Finally, CERN has introduced a new user administration scheme (CRA).

DESY (Dirk Jahne-Zumbusch) – planning a 10Gb link to FZK and a 1Gb link to Zeuthen. Adopting Indico from CERN for conference handling. Collaborating with INFN and Orsay on the apeNEXT processor. After a number of problems, they are moving to “higher quality” SATA discs. SuSE Linux is being replaced by SLD3 (based of course on Scientific Linux, SL). Quattor is being adopted and DESY will host the next Quattor workshop. Due to thermal limitations in their current centre, they are building a new computer centre to host expansion, with room for 1000 boxes. Deploying Oracle RAC servers. DESY now has a significant number of Grid-based activities.

FZK Gridka (Manfred Alef) – New server room coming on line and 153 new Opteron nodes will be installed, replacing old Pentium systems. 4.5PB of disc space will be added soon but the tender operation is still in progress.

GSI (Walter Schön) – in the area of spam fighting, they use “smart” greylisting – adding domains as suspect after some threshold of spamming is recognised even before downloading the daily update of this list. Experiments are in the process of moving to 64 bit operation. More Opteron systems have been added for their CPU farm and Linux boxes are now also in use for disc and tape servers. Yet another speaker who mentioned having to work around air cooling problems by investing in new air-cooled racks. And he described a number of load problems found on the latest series of Maxtor discs.

CNAF (Andrea Chierici) – CNAF is the INFN Tier 1 site and one of the main nodes of the Italian university and research network GARR. The speaker noted that it is “in real trouble with the electrical and cooling systems” for their 600 node Xeon and 150 node Opteron farms and may be forced to buy only low-power processors in the future. Migrating to CASTOR 2. There is active collaboration in the 3D project with a 4 node Oracle RAC cluster installed in a test environment.

Jefferson Lab (Sandy Philpott) – evaluating 64 bit X86 environments in preparation of moving in that direction. They still want to keep a second platform and are guarding Solaris for that reason. Still using Panasas for disk servers and adding another shelf (8TB in addition to their current 5TB shelf). StorageTek Flex 680 demo equipment returned to vendor after 2 instances of data loss. They are trying to wait before buying more tape capacity until the second generation of STK’s Titanium drives (2007-8?) in order to get maximum return on investment on their current installation. They are upgrading their LAN and WAN to 10GigE. New 11,000 sq.ft. Data Centre is operational but not equipped with a generator, only UPS by design, so they are relying on autonomous shutdown but they do not yet know how long the UPS can supply power so they do not know how long until the shutdown decision can be delayed. Working with IS group to install CERN’s NICEadmin tool at JLAB for Windows management. Installing Subversion for code management.

LAL (Michel Jouvin) – 12 new quad Opteron systems (dual CPU, dual core) but problems to install the correct network driver to permit network-based SL3 loads. SL3 everywhere and all re-installed with Quattor; Quattor is also used to monitor the Paris-wide GRIF grid sites. A workshop was recently held for

[10] Real Application Cluster

Quattor at LAL; it is thought to be a very valuable tool but has a long learning curve. Work is ongoing to simplify the templates. He recommends that CERN considers more use of standard components now being developed outside CERN where most current Quattor developments are now carried out. LAL is investigating Lemon for monitoring the GRIF sites. They have just started looking at SL4, initially on dedicated servers. They are completing migration from CERN's Agenda to Indico but, as Michel usually does when he adopts CERN tools, he found some CERN dependencies. LAL, like JLab, has installed Subversion for code management.

NIKHEF (Wim Heubers) – Like all LCG Tier 1 sites, they are ramping up for LHC startup. Favourite server supplier is Supermicro while Dell is most prevalent on the desktop. They have also noticed a distinct Apple revival although officially not supported. Phasing out a Solaris home directory service in favour of one based on EMC NS502 filer system and they will use this migration to unify the Unix and Windows home directories since the device offers native NFS and CIFS access without the need for Samba. The neighbouring SARA lab offers "storage on demand" at a fixed price per TB for a 3 year contract; so NIKHEF have started a pilot project with them for NFS and CIFS space using dedicated fibres and based again on EMC. Their SURFnet connection has been upgraded to 10Gb and now offers OPN[11] services. Together with other Dutch partners they have received a €25M 5 year grant to build a "BigGrid" for Dutch e-Science which will incorporate the Dutch Tier 1 site for LCG.

PSI (Urs Beyerle) – PSI has made their own tailoring of SL, including 64 bit support; it is also enhanced with the latest releases of KDE and OpenOffice. Configuration is done by cfengine which is not always easy. In addition, they sometimes use their own kernel! They also provide a diskless distribution but mainly for hardware testing and system rescue. There is also a thin client version to offer access to more powerful Linux or Windows servers. For high performance computing they access a large Cray XT3 (1100 Opteron 2.6GHz processors, 4.5TFlops) at the Swiss National Supercomputing Centre in the Ticino. At PSI a new SUNFire X4100 cluster will come online in the Spring, also based on Opteron chips and equipped with 12TB of SATA FC RAID discs for which GFS[12] is under evaluation. After evaluation of Plone, Typo3 and Imperia (the last a commercial product from Germany) for web content management, they chose the commercial option, Imperia, to the extent of justifying its cost in terms of effort saved.

RZG (Hartmut Reuter) – RZG (the Max Planck Computer Centre in Garching) is a supercomputer site offering computer services and computer hosting for the various Max Planck Institutes around Germany and partners in Italy and Holland. They run a large IBM Power 4/5 system and are currently evaluating their next generation system. There are also many Linux clusters, all IBM, on a mixture of chips (Power, Xeon and Opteron). The batch on the Linux clusters is based on SUN Grid Engine and all clusters use AFS although many (non-HEP) applications also use NFS. GPFS was tested 2 years ago but was found then to be too unstable. For HSM, archive and backup they use IBM's TSM[13]; TSM was chosen as the cheapest option. RZG will become a Tier 2 centre for ATLAS (although they use SuSE on their Linux clusters).

SLAC (Chuck Boeheim) - BaBar is still running although they have recently been battling a vacuum leak, solved in the past few days. Among SLAC's new activities. the new Kavli building for astrophysics (see presentation from previous meeting)has been commissioned; GLAST[14] is getting ready to announce a launch date; and the LCLS[15] is under construction. [For details of each of these, see the trip report from the previous meeting.] Their main computer centre cooling system was replaced in January and they were running fans at full speed and keeping all doors open in order to keep the centre operational. 350 new

[11] Optical Private Network
[12] Global File System
[13] Tivoli Storage Manager
[14] Gamma Ray Large Area Space Telescope
[15] Linac Coherent Light Source

Opteron systems were added to their main cluster and some older systems were decommissioned, thereby saving heat and power needs somewhat.

BNL (Ofer Rind) – Recent RHIC experiment running was disturbed earlier this year but has been fully operational since early March. The RHIC computer centre is running 2 STK 8500 robots very recently equipped by LTO3 drives. 100TB of NFS space will be phased out later this year and will not be replaced. Their Panasas disk service has given some concerns, especially on the service administration side and the Panasas-based NFS service does not scale as they would expect. There are now over 4000 processors in the combined RHIC/ATLAS installation and more are on their way, probably Opteron-based. Power and cooling is becoming a factor in purchasing choice although the centre's power capacity was recently upgraded. Almost all batch scheduling is now based on Condor with only a little LSF left in certain places. NAGIOS is used for system monitoring but much of it had to be re-engineered because it did not scale to their clusters (obviously the team who had proposed NAGIOS had not read the discussion of its scaling limitations at the Karlsruhe meeting a year ago).

Plenary Talk

Les Robertson

On the Wednesday, the LCG GDB met in an adjoining room and Les took the opportunity to address the combined audience on the challenges of the WLCG, as the project has been renamed, the W signifying Worldwide to denote that it now spans the globe using multiple grids, in particular EGEE[16] in Europe and OSG[17] in the US. The WLCG is a collaboration of currently ~100 computer centres spread over 20 countries; its funding was agreed in a Memorandum of Understanding (MoU) which maps out the coming 5 years operations. He also explained the various bodies of the project, their composition and their functions. He showed overheads from the LHC experiments' Technical Design Reports (TDR) illustrating the computing models of the LHC experiments, showing how they will distribute, store and access experimental data across the particular Tier 1 and Tier 2 sites associated with each experiment. He pointed the audience to the LCG web site for more documentation including all the TDRs of the experiments and of WLCG itself. In showing the resources dedicated to WLCG, it is clear that the majority of CPU power is outside CERN and this has implications on these sites. While reliable high-speed networking is clearly a vital requirement the MoU defines availability and problem response targets for the sites and mechanisms are being put in place to measure these. He then briefly explained the LCG Service Challenges, why they are so important, how they are progressing and future targets as we move towards LHC startup. Currently WLCG accounting shows a daily load of over 20,000 jobs per day with peaks of 15,000 simultaneous jobs but the main challenge in the coming months is to raise the overall reliability of the service and how all sites must contribute to this.

CPU Technologies

(Convener Bernd Panzer-Steindel)

Introduction - Bernd introduced the topic by covering some "random" aspects including :-

- Intel and AMD roadmaps of decreasing chip size while introducing 64 bit, emphasis on virtualisation
- Multi-core issues: dual-core is here, quad is coming but probably not 8-core until 2009 or later; how can HEP use this, if at all? Also can we use specialised heterogeneous systems such as the Cell processor with a PowerPC surrounded by 8 DSP cores

[16] Enabling Grids for E-sciencE
[17] OpenScience Grid

- Game processors, can we use these? Is there a usable Linux? What about their limited memory?
- Market trends: AMD chips heavily dominate the desktop in recent years and are making serious inroads in the server market. Another trend is the rapid growth in the number of notebooks
- Looking at costs, the major issue is the amount of memory required for HEP applications which has particular impact on multi-core systems
- Benchmarks: need to consider real applications when comparing measurements; CERN openlab have developed low-level monitoring tools to understand what the CPU is doing at instruction level (see Sverre's talk later in the week)..

Power Consumption Issues (Yannick Perret, CC-IN2P3) – heating problems during the summer and a steady cooling and power requirement has forced IN2P3 to understand more deeply the power needs of current and future processors. Direct measurements of electric power used are compared to a representative set of CPU instructions. This creates a unit of power and an (old) system is defined as the base unit; thereafter, all measurements are compared to this. He showed the list of systems tested and some of the results in tabular and graphical form (see overheads). Opteron systems are more efficient than Xeon at this time but the most efficient systems measured (CPU power per unit of electric power) were IBM blades. But this evaluation did not include purchase cost as a factor. In this area, he noted that when calculating the costs of acquisition, you need to include the costs of electric power, cooling, racking, networking, etc. While there was support in the audience for blade futures, it was noted that a single incident could make quite a large amount of capacity unavailable; TRIUMF and IN2P3 however both report good experience of blade systems at this time.

Dual-Core Batch Nodes (Manfred Alef) – Having recently acquired 153 of these, he explained first the procurement. Gridka is bound by EU public tendering rules. The tendering was based not on a number of boxes but on a total Specint performance target. Either Intel or AMD is acceptable. The first parameter in the adjudication was system price but also the costs of electric power, rack space, cooling and admin costs (€200 per box) were included. From the variety of offers received, they chose dual-socket, dual-core Opteron 270 (2.0GHz) systems. Based on these systems, Gridka has updated the benchmarks presented at the last HEPiX (see overheads). He also showed power consumption and thermal efficiency charts of different processors. Here again, Opteron systems were the winners. He confirmed that all benchmarks were run in 32 bit mode but 64 bit mode did not change the relative performances, only the absolute numbers (by some 5% - a figure more or less confirmed by GSI) Conclusion – buy dual-core Opterons.

Benchmarking AMD64 and EMT64 (Ian Fisk, CMS and FNAL) – Tests were done in true 32 bit mode and both true 64 and 64 bit compatibility mode. He listed the processors tested and the applications used (see overheads). Even in 64 bit compatibility mode, having the operating system running in 64 bits gives a performance boost to the applications throughput. Pythia in recompiled 64 bit mode gives a 20% performance boost over compatibility mode running and no performance drop when running 4 processes in parallel. The ROOT stress test shows a very small (1.6%) performance drop when running 4 processes in parallel but is 6% faster in 64/32 mode than in 32/32. OSCAR also showed a very small drop in 4 processor operation but is 4% slower in 64/32 than in 32/32. ORCA however shows a more noticeable performance drop when running 4 processes in parallel, possibly due to increased I/O, and is 15% slower in 64/32 than in 32/32, possibly a memory limitation. In power consumption tests, dual core systems draw almost the same electric power for double the CPU power. [This has saved several US Tier 2 sites from having to upgrade their facilities when acquiring new systems.] Summary

- Today's 32 bit applications run just fine on 64 bit processors without porting in 32 or 64 bit compatibility mode
- don't be afraid of running in 64 bit compatibility but there is a boost in performance if you can recompile to true 64 bit
- dual core processors offer twice the compute power for the same electrical and cooling power.

Networking Technologies

The Convenor of the session, Enzo Valente, noted the paradox of talking about networking in a room with no networking! [It was due to a memory hardware fault in a switch resulting from major power cut over the weekend and was fixed by the following day although other influences took 2 more days to resolve.].

GEANT 2 Status and Plans (Marco Marletta, GARR) – GEANT2 was not simply an upgrade of GEANT but also incorporated support for new services such as high capacity P2P[18] services; there is also more emphasis on research and development and there is extended coverage compared to the first phase. The topology is mainly dark fibre in the centre and leased line towards the periphery. There is improved performance monitoring including the establishment of a Performance Enhancement Response Team (PERT). The speaker displayed the variety of circuit services available and projects supported.

Using E2E Technologies for LHC (Marco Marletta, GARR) - today we have a multi-layer network topology but there is a need for many interconnections at layers 1 and 2. However, IP was designed as an end-to-end (E2E) layer 3 protocol and these facts are in conflict. A case study was made in 2004 by FZK and CNAF (KarBol) to establish an L2 (layer 2) E2E VPN network using an E2E layer 3 path across 5 competence domains. The challenges in this were listed, for example including possible security concerns, platfor inter-operability and cross-border fibres. He then explained the LHCOPN – an optical private network linking the LCG Tier 1 sites. Like the previous example, a major concern is security - it bypasses the usual firewall and other site security systems.

TCP Performance Optimisation (Tiziana Ferrari, CNAF) – these results come from a CNAF/CERN openlab collaboration to identify the software components needed and to understand and compare different TCP stacks and different 10Gb Ethernet NICs. She presented the performance metrics and the parameters which can affect performance. For example, block size can greatly affect application throughput, but it flattens out around 12,000 byte blocks and at some point increasing it further causes system instability. She then explained the effect of adopting a new TCP stack known as BIC – Binary Increase Congestion control – comparing it to the traditional Reno TCP stack. In tests, BIC appears better in handling congestion up to some level but a closer examination shows that the different Linux kernels used for the different stacks probably also contributes to the effects thus making reliable interpretation of the results more difficult. Summary – tuning of both system and applications can give significant improvements in TCP performance but a good kernel understanding is very important and she noted 2 references where such knowledge may be acquired.

Discussion - Although the talks were fairly network-technical and this is the first time HEPiX has covered networks in some detail, there was a lively question and answer session at the end. Enzo's summary was that modern performance is good enough to start looking beyond pure bandwidth-on-demand numbers to what alternative protocols such as E2E can offer. Another issue is good connectivity to countries such as India, Japan, China and Taiwan to enable their physicists to participate fully in HEP. And beyond that to third-world countries – the Digital Divide which was a recent hot topic at CHEP06 in Mumbai.

Batch Systems

(Convener Tony Cass)

Passing Information to the gLite CE (Francesco Prelz) – this talk followed a very interactive presentation by Francesco at the previous meeting. The problem then was to how to make a local batch system access and understand the global information submitted with the job. Options range from full-fledged match-making to a "do it yourself" approach, translating the the job's requirements to what is

[18] Peer to Peer

availably locally. This talk was a status report for a translation effort with CERN for the LSF batch scheduler carried out as an implementation of the second approach, chosen because this leaves the local batch administrator more flexibility, for example able to select which attributes are relevant locally. It requires Prelz's team to develop an algorithm to extract the global attributes and map them to LSF attributes. A generic extraction method has been developed using a standard Condor ClassAd feature and the result can be forwarded to a gLite CE[19]. It will be available in gLite version 3.1 which "is coming soon"; an RPM is available for version 3.0. The translation invokes a script-based callout to the local batch system. Prelz then explained the various steps included in the BLAHP[20] job submission interface in some detail. He says he is still seeking volunteers to work on other batch systems. He was followed by Ulrich Schwickerath who described in more detail how BLAHP is being interfaced to LSF using a PERL script.

ATLAS Batch (Laura Perini) – as all subsequent experiment representatives, she started with the computing model for her experiment – 1000 events per 2GB file using a data taking trigger rate of 200Hz independent of luminosity. She also described the various tasks to be performed at each Tier level. Most production so far been done on the grid with several 10s of millions of events fully reconstructed and she said it had not been particularly easy. In production running they expect users to submit short jobs to optimize their turnaround. Data for analysis will be at Tier 1 and Tier 2 sites and users are expected to send jobs to the data. All simulation should be at Tier 2 sites (shared with analysis, 50% each) and Tier 1 sites will be shared by analysis, reconstruction and calibration jobs. Most ATLAS users have not yet been exposed to Grid middleware and expect a simple extension of their current batch environments. Therefore ATLAS expects tools to simplify the grid interface as much as possible and they will start in LCG SC4 to see how this can be achieved, first on a small scale but later expanding, adding more complexity step by step.

CMS Batch (Stefano Belforte) – CMS will have a few groups mapped to individual VOMS groups, each with allocated CPU resources. CMS expects good monitoring of jobs and jobs must have access to experiment software distributions and Posix-style access to the data. CMS has 7 large Tier 1 sites. They have around 30 smaller Tier 2 sites where most of the simulation and analysis jobs will run. They will require several batch queues at these sites of differing time limits and expect sites to offer this in order that the grid can map grid queues to local queues although at a site CPU resources may be split by design between different activities (simulation or production by given groups). CMS believe that sites should have more control of the use of their resources and that Grids should merely exploit these resources rather than invasively enforce their policies. Sites need to set up shares for around 3 CMS groups mapped to VOMS[21] groups and implement fair shares between them with job queues with different time limits for each group. John Gordon made a plea not to have to establish X queues for ATLAS, Y queues for CMS, etc, but rather state how long the job is and let the local scheduler assign the job as appropriate according to local resources.

LHCb Batch (Andrei Tsaregorodtsev) – in LHCb, Tier 2 sites are only supposed to run simulations. Their overall software framework, DIRAC, consists of modules providing services controlling VO[22] mapping, workload management, etc. They use a central task queue which allows easy prioritizing of all LHCb jobs and application of VO policies; thus sites are not required to manage shares or priorities. Pilot agents are deployed on worker nodes and, if run successfully, call down user jobs from the central queue. Normally it will be one pilot agent per user with jobs in the queue but if a high priority task does not get scheduled in a given time, one or more extra pilot agents may get scheduled for different sites. [In the discussion Harry Renshall pointed out that this gives a bad mix of short and long jobs on any given node.] This

[19] Computing Element
[20] Batch Local Ascii Helper Protocol
[21] Virtual Organisation Management System
[22] Virtual Organisation

scheme was successfully demonstrated in the recent LHCb Monte Carlo simulation run when over 5000 simultaneous jobs were run in parallel. They intend to use this scheme again in the planned LHCb data challenge in June this year.

ALICE Batch (Federico Carminati) – In ALICE, Tier 0 is used for first pass reconstructions, Tier 1 sites are for subsequent reconstruction jobs and scheduled analysis jobs, Tier 2 sites for simulation and user analysis jobs. An ALICE job optimizer will split jobs according to data locality. On VO Boxes, ALICE has a computing agent to collect jobs, load ALICE software if needed and then submit job agents to the appropriate resource broker. Job agents are only sent when needed; job location is determined by data location. The scheme was used for 22,500 production jobs with only 2.5% inefficiency. With this scheme, fewer users now submit jobs locally but rather "to the grid" and they do not really care under which job scheduler the job is eventually run. Currently they require a single long ALICE queue expressed in KSpecint units mapped to local nodes by the grid. One important restriction is the need for 2GB of memory per core.

DataBase Technologies

(Convener Dirk Duellmann)

Introduction (Dirk Duellmann) – Dirk described the history of how databases became common in HEP environments, starting with RDBMS's in the early 90s. We have learned that databases should be used only for critical data which needs DB features. He described how LCG databases are kept up to date via asynchronously replication via Streams. He compared the concerns of local and central site managers and how these must be reconciled to provide an overall reliable service.

Database Service for Physics at CERN (Luca Canali) – At CERN, DB services for physics are deployed on a cluster of 2 and 4 node which can be expanded as required, both in terms of CPUs and storage. These are called RACs – Real Application Clusters – and they are a feature of the Oracle RDBMS engine. He listed the features of the latest release of this, version 10g. These include failover and load-sharing and applications do not need to be built specifically for RACs. Another fundamental piece of software is ASM, Automatic Storage Manager, which for example implements mirrors and RAID arrays for storing Oracle DBs. He presented some performance numbers, showing for example, linear performance as the number of discs is increased, at least up to the limit of the 64 discs used for the test.

Database Deployment at CNAF (Barbara Martelli) – they are actively collaborating with the LCG 3D[23] team at CERN and follow their guidelines. CNAF offer a production service as well as a development one. The first executes on dedicated hardware, configured by agreement with the users, but the development service must be shared by all its users. Currently there are 2 2-node clusters, used by ATLAS and LHCb, and a third one is used for CASTOR, FTS[24], etc.

Database Deployment at RAL (Gordon Brown) – They have been involved with 3D since its start and operate a test setup for it. They have also 2 2-node RAC clusters running Oracle 10g, Data Guard and ASM. These are still in test but should enter production over the summer. They also run the SRB[25] from San Diego and CASTOR AS well as for LCG, they run DB services for the UK national grid service.

[23] Distributed Deployment of Databases for LCG
[24] File Transfer Service
[25] Storage Resource Broker

Authentication Technologies

(Convener Wolfgang Friebel)

Introduction (Wolfgang Friebel) –Wolfgang has started a collection of useful information on these topics which he will link to the official HEPiX web site. He listed major areas of work such as Kerberos, PKI[26], Single Sign-On (SSO) and password synchronization across platforms. PKI is becoming the favoured technology. He then listed some of the issues he has come up against, such as how to synchronous passwords and the "dream" of single sign-on. 0

One Time Password Integration at BNL (Robert Petkus) – they offer SSO via Kerberos 5 authentication, generating AFS tokens via klog for example; via ssh-based authentication; or via LDAP credentials. But due to attacks such as stolen passwords, impersonation or copies of ssh keys, they felt that they needed something stronger. Issues around one-time passwords include user education, the need for a user possibly having to carry round a set of multiple hardware tokens for different sites – unless related sites can adopt a common system or fabric. Another major drawback is the loss of SSO possibilities unless the one-time password can be linked to Kerberos tickets. BNL has chosen a Radius server, Cryptocards for users (he did not know the cost), gsi[27]-enabled ssh, Myproxy credential management services, a BNL Certificate Server and BNL's GUMS (Grid User Management Server). He then listed some pros and cons of using one-time passwords to connect to the RCF facility at BNL and he described how one could implement closer integration with Kerberos. He ended with an open question about how easy it would be to establish cross-realm authentication. His scheme is in pilot phase with some users accessing some services.

Single Sign-On at RAL (Jens Jensen) – SSO is more than simply typing your password once; it includes identity management and user management. If you are on-site, you effectively use Active Directory (with Kerberos behind). If you are offsite, you may have a grid certificate. Otherwise (from an Internet café for example) you must rely on username and password. RAL's SSO scheme relies largely on sshterm, written in Java, which can take the form of a stand-alone application or an applet. RAL has integrated it with Myproxy for grid integration. The second part of Jens's talk was how to cope with 14-15000 user accounts for all the experiments based at RAL. One of these, the Diamond Light Source, has bought a commercial system, Vintela, which manages accounts across Linux and Windows. It has a "lost password" scheme and user creation is scriptable. RAL are trying to integrate this with work on other authentication activities such as Shibboleth.

Integratiung PKI and Kerberos Authentication (Alberto Pace) – Today at CERN, we use 2 technologies, PKI and Kerberos. Both have strong and weak points but both are here to stay so we need to integrate them. He explained in some detail the PKI and Kerberos technologies. For Kerberos there are Windows and AFS schemes and a CA[28] for LCG. The new service should maintain the facilities offered by the LCG CA but also it should grant Kerberos tickets needed for AFS and for Windows access. Initially it will still be based on Kerberos passwords and user authentication. The future scheme could be based on Smartcards. He then discussed issues around managing user certificates and smartcards

.

Optimisation and Bottlenecks

(Convener Wojciech Wocjik)

Performance and Bottleneck Analysis (Sverre Jarp) – this is work done in the framework of CERN's openlab collaboration with industry. One of the first choices to make is which compiler gets the best

[26] Public Key Infrastructure
[27] Grid Security Infrastructure
[28] Certificate Authority

performance from your chip; then which compiler parameters have which effect? Use shared libraries at runtime or not? The choice of chips itself can greatly affect performance, especially the choice of 32 or 64 bit architecture and the level of multi-core in the chip. In starting an analysis, every change must be tested. There are a number of performance analysis toolkits available and he listed some for different platforms (see overheads). He showed some recent measurements of real experimental codes using some of these tools. One of the major obstacles in analyzing HEP code is its complexity – for example its modular nature creates a lot of time navigating between modules or routines. He presented some code optimization ideas ranging from modifying compiler options to modifying your code taking account of the instruction-level operation of your target chip. Having explained the methodology and emphasized the importance of selecting good tools, knowing the chip architecture and how your algorithm maps to this, he then presented some results obtained from the openlab collaboration with Intel.

Code/Compiler Problems (Rene Brun) – Compilation time is becoming a problem in both ATLAS and CMS codes; compiling the complete ATLAS suite takes 12 hours! On examination this was traced to a number of causes including excessive use of in-lined classes in the code; the casual use of Include statements (which expand to 20,000 lines of code each!) combined with a total of 9000 classes in the ATLAS code has created a huge hidden cost. He showed how the use of pre-compilers can result in significant savings. Another problem with shared libraries is increased startup times. Although shared libraries have many advantages, they can seriously disturb large interactive applications. Again detailed examination shows that frequently only a tiny fraction of the code in shared libraries is actually used by the applications linking to them. A serious re-think is required. He also recommended finding hot spots in programmes, referring to methods such as described in the previous talk. Lastly, he spoke of multi-threading and the importance of making programmes thread-safe in order to take full advantage of multi-core chips.

Controlling Bottlenecks with BQS (Julien Deveny) – a BQS[29] resource is a semaphore representing a service to be controlled and contains information about that service. Using these semaphores, BQS knows when to and when not to schedule a job depending on which resources the job needs. Unfortunately the scheme relies on users accurately and honestly describing the resources they need and also there is no interface to the grid since this is a local concept at CC-IN2P3.

Optimising dCache and the DPM (Greg Cowan) – Each Tier 2 site has unique policies and constraints. This leads to various combinations of middleware components. The University of Edinburgh chose dCache and LCG's DPM (Disc Pool Manager). They have tried to optimize their use and offer recommendations to similar sites. He described the configurations of hardware and software modules for the tests. He presented the full results in some detail. Using XFS in the DPM tests showed noticeably better performance but not on the dCache tests. Transferring more than 10 files in parallel caused some level of instability but this was traced to simultaneous start-up of the transfers by FTS[30]; perhaps a staggered startup of the transfers would help.

Miscellanea

ATLAS Data Management over Grid (Alexei Klimentov, BNL) – He started by showing, for the third time this week, the ATLAS data flow then moving on to illustrate the problem of managing huge quantities of data over very many sites from a variety of sources. "Don Quijote" is a second generation suite of data management tools in ATLAS. It is based on global repositories and catalogues, local stores,

[29] BQS is the Batch Queuing System at CC-IN2P3
[30] The LCG File Transfer Service

the mapping of logical and physical file names and the tracking of data movements. It is coming into use by ATLAS production programmes.

Monitoring Services in a Grid Environment (Paul Millar, Uni Glasgow) – he spent the first half of his talk reviewing the concept of monitoring, illustrating each aspect with an example from outside the world of HEP, before eventually referring to tools, such as NAGIOS, familiar to a HEPiX audience. He has developed a tool called MonAMI to unify the sensors which a site might wish to monitor and send to their favourite monitoring tool. MonAMI is (or will be) configurable according to local requirements and he displayed a log of what it is can currently monitor and to which monitoring tools the data can be sent. He has placed the tool on sourceforge and invites people to use it and feedback comments to him.

Accounting on Large Farms (Andrea Guarise and Felice Rosso, INFN) – they have developed a software tool, DGAS[31], which gathers accounting information from the batch job workers and forwards this to a distributed database which is managed by a network of servers. From thi database, the servers will extract data and present reports from a user point of view, from a cost view, etc. In order to have site-wide accounting independent of any particular batch scheduler, this tool was integrated into LSF accounting by parsing the LSF accounting data and transferring it to the above database. They have built privacy into the tool but certain people can have access to all data belonging to a particular VO[32]. The tool also distinguishes between local and grid jobs when reporting its data.

SYMPA for Mailing Lists (Dirk Jahnke-Zumbusch, DESY) – DESY has 7000 registered addresses, 800 central mailing lists with 30,000 subscribers. Those at Hamburg are controlled by PMDF (now no longer maintained), those at Zeuthen by Majordomo. As well as merging both into a common system, they would like better support for German language mails and modern mailing list features such as self-service, archiving, etc. He then listed the main features of SYMPA and how these fulfilled their requirements. Each list has 50 parameters which can be set but for most lists the default settings will suffice with only a few changes. It allows to manage families of lists as one unit. It has a 3 tier architecture with a backend database (choice of DBs includes Oracle. MySQL, etc; even LDAP is possible), a middle layer consisting of an Apache web server and a client module for the user interface. It supports X.509 certificates. It appears to be a stable product, in development by the "Comité Réseau des Universités" (CRU) in France since 1997, now at version 5 and used by over 4000 sites.

SLAC Request Tracker (John Bartelt) - RT, an open source product in active development, has been adopted by SLAC for tracking requests made by users towards the system administrators. It supports queues of tickets and the workflow is configured to deliver these to the appropriate person. The queues are monitored by a "hot seat" person during business hours. To date, they have created 45,000 tickets but he admitted it has some defects, including poor interaction with the Remedy trouble tracking product which is still used for the SLAC Helpdesk and RT also requires its own user IDs. They would like to expand it to the Windows side. In answer to the obvious question (why a second trouble tracking scheme) they rule out using Remedy because they claim it had no e-mail interface when they started this (incorrect, CERN has had this since many years for example) and was too heavy for their use.

Integrating JASMine and Auger (Sandy Philpott, JLab) – JASMine is the storage management system and Auger is the interface batch system on the farm. JASMine is one of the clients of Auger but it also receives raw data storage requests from the experiments. It was felt that a better integration of the two could make more efficient use of the farm nodes which sometimes had to wait for data from JASMine. The first change was to add intelligence to Auger to pre-stage the required data. But they quickly realised that they were pre-staging too much, for example if a user submitted a large number of jobs. Auger was then modified to only pre-stage data just ahead of what LSF (the batch scheduler) requires to schedule the

[31] Distributed Gas Accounting System

[32] Virtual Organisation

jobs. To this they then added an adaptive cache where the space used by a person or group depends on that person or group's activity and the overall throughput of the system. Recent graphs of CPU usage shows definite improvements.

IHEPCCC (Randy Sobie) – Randy was elected chairman of IHEPCCC earlier this year. He explained what it is, what it does and its relationship to HEPiX. It is a sub-committee of ICFA[33] and its mandate is to provide a forum for the exchange of information on computing issues of a global nature. He listed some recent activities such as participating in the Digital Divide discussions and promotions at CHEP06. One major issue is the support of travelling physicists where VoIP[34] is the current hot topic, especially the resistance of labs such as CERN to offering Skype access. Other areas under discussion includethe use of commercial software and whether HEP-wide licences are worth pursuing. IHEPCCC would like to closer links with HEPiX, starting with a presentation by a HEPiX person after each HEPiX meeting to IHEPCCC.

Data Centre Cooling (Bill Watts, Intel) – the speaker is responsible for 150 data centres globally and he knows very well the sort of cooling problems we have in HEP. Air flow efficiency is measured as the amount of heat which can be removed. As chips get smaller and space is more efficiently used, the air flow efficiency has dropped and this has fuelled the developments at Intel and elsewhere. He presented some charts of how they rated the installed capacity of their computer rooms measured against required air flows. [Tony Cass noted that the room size in these charts is only one-tenth the size of the IT Centre although the speaker told us later that his data scales.] He illustrated the talk with examples of how not to cool cabinets and then how to improve the situation. Cable management in the rear of the racks is very important. They have worked with cabinet manufacturers to manage better the air flow space: air comes from the front of the cabinet; there is a rear door and a flue for hot air exhaust into the return air space in the ceiling. With these cabinets they are able to sustain much higher power ratings per floor space. The overall air flow efficiency has increased by some 75% since this upgrade started and he has ideas for further improvements. He then presented a case study of refurbishing a computer room for a customer to install high performance blade servers. He presented layout charts, heat maps, power ratings by floor area, animated air flows and so on. One spin-off of the work was that they could suffice with a much smaller raised floor (18 inches against the original 36) which itself saved $500K. During a lively question session he said the cabinets he had described cost $16K from the vendor. .

Storage Day

Tape Technology (Don Petravick) – at Fermi, the tape growth is around 1PB per year but 16PB of data is moved on and off tape per year and they read 3-4 bytes per byte written – unusual in the outside tape market. He says that in HEP, tape's roles are for HSM[35] and for archive. The largest risk of loss of data is at tape writing and this can be aggravated because not all faults may be signalled at this time but only discovered on read-back later. But during its storage lifetime, risk of loss is low (which Fermilab checks by sampling random tapes for read). There is an end-of-life if it is kept too long but via their checks, Fermilab has yet to see end of life effects, the media is declared obsolete before that happens and is replaced by a more modern device or media. The most worrying aspects of tape are that it is highly mechanical and it is a specialised technology. Also performance is highly variable, depending on which tape the files required by a job may reside and where on the tape they reside and therefore we need to cope with these effects. At Fermi, tape capacity doubles every 18-24 months, LTO-3 drives currently store 400GB but there is no inherent tape density limit as there is for disc technology. In summary he claims tape offers high quality retention technology and simple, reliable units of expansion but it does complicate

[33] International Committee for Future Accelerators
[34] Voice Over IP networking
[35] Hierarchical Storage Management

data handling and it requires specialised skills to manager and operate. And the future roadmap appears to face no fundamental engineering limitations.

Disc Technology (Martin Gasthuber) – He presented various disc configurations such as FC[36] SAN[37], SCSI FC and others. Important components are not only the discs themselves but also the interconnects and the disc and network controllers. Expected performance is 40MB/sec throughput per TB of storage and it is important to be able to isolate fault domains (individual discs for example) to avoid breaks in service. He listed issues to consider when acquiring discs – disc type, controller, network, etc. And having a knowledgeable local vendor or integrator is very important. In large sites, high-speed networking, 10GigE or Infiniband, may be needed to access disc storage. Discs are getting just too slow and price per GB is flattening out. He offered with some predictions – no further increase in FC use but rather Serial Attached SCSI (SAS) which will come with smaller form factors; SATA[38] will be around for a while but there will be no real improvement in performance. He ended by describing Object Storage Devices (OSD) which he believes will come in the coming years – storage in a box and offering multiple protocols.

Hardware Potpourri (Andrei Maslennikiv) – Andrei described what he called a “fat” disc server contender. He compared what CERN requires for CASTOR performance with what his configuration can achieve and he believes it could satisfy the needs for CASTOR for a cheaper price.

GPFS and StoRM (Luca dell’Agnello) – GPFS[39] has been in production use for over 2 years. It had been extensively tested with both generic codes and physics (LHCb) applications on a dedicated testbed. Similar tests were made with Lustre. Lustre out-performs GPFS but the latter was chosen because of problems encountered with Lustre which was considered (then) as too intrusive, requiring kernel changes for example. They have “quattorised” their GPFS installation. GPFS is stable and easy to install; it is free for academic sites but support from IBM is considered poor and even trying to buy support has not proved easy! StoRM is a disc-based Storage Resource Manger developed at INFN Bologna which is optimised for GPFS. They may also interface it to Lustre in case they look again at the latest releases of Lustre. It is not in production yet and they are not actively working with the SRM developers although they do read the material being produced by the SRM Working Group.

Local File Systems (Peter Kelemen) – The presentation was concerned with the choice of a journal file system to be used on CERN’s commodity Linux servers. He described briefly ext3, a block-structured file system, probably the simplest implementation of a journal file system and the only one supported in Redhat Linux version 4 (RH4). The next contender is XFS but it can only journal metadata and is probably the most complex file system under consideration. It is disabled in RH4 but it is feature-rich and powerful. JFS has similar properties to XFS; it is still actively being developed by IBM but is also disabled by RH4 and not widely used. ReiserFS is another contender but also only journals metadata and is designed for small files and considered fragile; also disabled in RH4. CERN commodity disc servers need support for large files and large filesystems and mostly streaming IO (RFIO). They would also like delayed allocation, space pre-allocation and online defragmentation. These all lead to the choice of XFS and this therefore has to be added to the SLC[40] distributions. Various operational problems have been seen in the SLC3 distribution but with care these can be avoided and are mainly hardware-related. CERN has 650TB of XFS space. In SLC4, XFS is included in the RH4 release but not enabled. However ext3 is catching up fast in performance and the future is unclear.

AFS/OSD Project (Ludovico Giammarino) – this is being developed in CASPUR in conjunction with CERN and FZK. The principle goal is to improve AFS performance and scalability by extending AFS to

[36] Fibre Channel
[37] Storage Area Network
[38] Serial ATA
[39] General Parallel File System, from IBM
[40] Scientific Linux CERN, version 3

support object-based file management. The work is based on the SCSI T10 OSD standard. There three basic components – an AFS client extended to support the T10 standard; a similarly-extended AFS server and an AFS DB server capable of supporting OBSD. He explained, rather too quickly, the operation of the components and their interactions. One issue is a lack of support for volume replication and this is under active investigation. The target is deployment in production later this year.

WAN Access to a Distributed File System (Hartmut Reuter) – He described several ways to achieve this, most of them previously described, their advantages and disadvantages or risks. RZG is part o the DEISA European Supercomputer collaboration and they have chosen GPFS and AFS (although the latter not at all sites). For their (RZG) needs, GPFS performs much better than AFS and they are quite happy with it although they believe that current work with AFS could improve that alternative.

Disk to Tape Migration Introduction (Michael Ernst) – Michael started by reminding us of long-ago predictions that tapes would "go away" but apparently they are still with us. Data Management issues are not only from disc to tape. The challenges are a lack of standards, portability issues, location finding and of course performance. He then listed some criteria to be considered, and in particular some security aspects.

CASTOR 2 (Sebastien Ponce) – He presented a quick overview of CASTOR 2 and how it has changed from version 1. A major difference is that the clients are much lighter and the service is more DB-centric with stateless deamons. Tape handling has been optimised. He reviewed CASTOR 2 according to the list of criteria from Michael; for a full list of the comparison, see the overheads. Much of the functionality listed by Michael is indeed now available. Policies and protocols are now pluggable and prioritisation of requests is possible but only with LSF for the moment. There are interfaces to SRM[41] V1 and V2. On the security side, there is a new authorisation scheme, resiliency against hardware failures and regular DB backup but not specifically disaster recovery. Strong authentication is under development. 46M files have been saved to date, 4.6PB of data with an average file size of 100MB. There is 715TB of disc with 1M files staged. Scalability tests with 50 disc servers and 30 tape servers achieved 2.2GB/sec incoming, 1GB/s read back and 1.2GB/s to tape.

dCache (Patrick Fuhrmann) – again, most of the talk was to compare the features of dCache to the criteria list of Michael and again the full details can be consulted in the overheads on the agenda web page of the meeting. He noted with some satisfaction the growing list of dCache developers, more and more of them outside the original development team from DESY and Fermilab. He illustrated the data and control flow of dCache. Compared to the criteria list, there is no way to prioritise individual requests, only protocols or VOs[42]; there is support for transactions; SRM V1 is supported but support is not yet complete for V2; several protocols are supported and more can easily be added; there is support for pluggable VOMS[43] integration.

HPSS (Andrei Moskalenko) – he described the operation of HPSS[44] and in particular the architecture and operation of version 5.1. HPSS is configured via storage classes and classes of services and hierarchies. At CC-IN2P3, HPSS is at the centre of their storage infrastructure and is used by dCache, Xrootd, SRB[45] and other clients. He then compared its features to the criteria list and found most requirements satisfied in whole or in part. CC-IN2P3 are rather happy with its features and performance and they have faith that it will continue to do satisfy their needs although they do have some wish-list items.

[41] Storage Resource Management
[42] Virtual Organisatrons
[43] Virtual Organisation Management System
[44] High Performance Storage System, originally from IBM
[45] Storage Resource Broker

TSM (Jos van Wezel) – Jos described the Tivoli Storage Manager which FZK use as a backend for dCache. They chose that as they already had good experience with it elsewhere in the centre and they appreciate being able to offload some support to a commercial supplier. He described the working environment and explained briefly how TSM operates. He showed the data flow and how scalability is handled.

Discussion – some concern was expressed about the level of local support required for some of the options described in this session although at least some (most?) Tier 2 sites will have no tape and no HSM requirement. And those Tier 2 sites who do want an HSM can choose the gLite Disc Pool Manager (DPM) which is designed for that.

Bulk Data Movement
(Convener Jamie Shiers)

Hardware Solutions (Martin Gasthuber) – Martin had sent out a questionnaire and from the input he had received back, most sites have moved from dedicated hardware configurations to standard commodity systems. Also, most sites replying appear to have created a bypass to get the bulk transfers past (rather than through) their local firewall. All servers are Linux-based, mostly storage-in-a-box; most sites have GigE networking. Security is a big issue everywhere and in particular how to handle the scale of the transfers needed. The BNL solution of 2 DNS[46] domains is interesting.

Software Solutions (Graeme Stewart) – The lowest level of the software is of course TCP streams; on top of this GridFTP can be used to move the data around as needed. But it has drawbacks – it requires dynamic ports, for example, which are problematic for firewalls. GridFTP version 2 fixes some of the problems. An alternative may be the use of HTTP and work at Uni Manchester by Andrew McNab has made this an interesting option because it does not require any dynamic ports. Another method of bulk transfer is to build a transfer service; gLite FTS (File Transfer Service) and Globus RFT (Reliable File Transfer) are implementations of this and some testing has taken place at Uni Glasgow. Both work in similar ways. Although FTS has a slightly more complicated architecture it has some advantages over RFT. FTS works using “agents” to interact with the VO and with channels which permits some flexibility and is more scalable. RFT also has the disadvantages of only having a Java client and of being slower and less robust. Going up yet another level, you need to interact with the file catalogue via tools such FPS (File Placement Service) and DRS (Data Replication Service). Stress testing has proved that much experience and practice is needed to arrive at a reliable and stable service.

Summary (Jamie Shiers) – Jamie agreed with the previous speaker in that there remains a lot of work to be done to recover from “normal” operational problem. And communication, human as much as computer, is essential. Jamie then showed how LHC startup will mean full capacity will be expected by the experiments from the beginning. He displayed the blog which has been created to encourage more sites to participate and contribute to the current Service Challenge 4 exercise.

Backup
(Convener Harry Renshall)

Site Summary (David Asbury) – David started with a summary of replies to a questionnaire sent out by Harry: a variety of products are used but the most often-quoted is TSM (all of them on AIX boxes

[46] Domain Name System

although CERN is planning a Linux installation). A lot of INFN sites have developed their own tool for backup. The frequency of backup varies, both for “full” and incremental (although TSM does not have the concept of full backup). Backups are kept for typically a year but again there are large variations. Although restore information was not asked in the questionnaire, CERN restores ~100 files (~1GB) per day and the rate seems to be increasing. FZK is by far the winner in quantity of data stored in backup and archive store.

TSM at CERN (David Asbury) – Tivoli Storage Manager has been in use at CERN for a very long time (15 years). There is some residual Legato but it is being phased out; AFS home directories are backed up to CASTOR at this time. Experimental data is not backed up as such although experiments may make copies, especially via CASTOR. .David described some of the reasons why TSM is used and he showed the configuration of the servers. In particular he explained how data duplication is used to protect against loss of data. For monitoring, they use Servergraph/TSM although he admitted it is quite expensive. CERN tries to maintain a disc buffer large enough to store at least 24 hours of backup and then most restores are from disc and not from tape. Future plans are to investigate Linux-based servers and LANless backups (data directly from client to TSM tape drive via FibreChannel): TCP/IP overheads are reduced but certainly it will have down-sides.

AMANDA at TRIUMF (Steven McDonald) – AMANDA is an open source backup and archive product. It is used only for backup of ~150 mostly Linux desktop systems and a handful of servers, not as an LCG Tier 1 solution. Has been around for over 10 years and is now rather mature. It runs on.many operating systems, including Windows DFS via Samba but TRIUMF does not offer this. It runs on a smallish node attached to a large disc pool (for the backup) and a 7.5TB tape store (for archive). He described the backup cycle in operation. The clients compress the data before sending it for backup with a typical compression rate of 40-50% and they typically backup 180GB of compressed data per night. Data is archived every 2 months and stored “for ever”. He ended with some advantages (low cost, efficient, configurable, easy restore, reliable, etc) and disadvantages (no web GUI, no long term data trending and difficult to archive).

HEPiX Storage Task Force Report

(Convener Roger Jones)

This interim report was already presented at CHEP06 and to the LCG GDB. Its mandate was to review the storage market place, especially the disc market. in view of imminent purchases by, especially, LCG Tier 1 sites. Many of the talks in the previous session covered much of the material in the report. The main outstanding issue when the report was first published was archival storage. The task force believes that the field should be regularly reviewed and the report updated. The report covers areas such as disc technology, archival media, procurement procedures and how the experiment computing models affect storage. In each of the areas, he summarised a few highlights – for example the steady price decrease in discs may be slowing down; the importance of archiving; and if (or when) disc prices may make tape archival unnecessary. A number of recommendations are made (see report) and it is felt important to test these. It is suggested to share hardware experiences via a web site, such as HEPiX.org, taking account of course of commercial confidentiality. Further, a similar task force should be set up to review CPU technologies.

Operating Systems

(Convener Alan Silverman

Evolution of Managing Windows at CERN (Ivan Deloose) – Ivan described how the “Windows for Controls” project spawned the Computer Management Framework (CMF) project to manage the 5100+

Windows XP desktops at CERN. He described the installation and patch methods in use, mostly using Microsoft tools such as SMS for patching and Active Directory for group policies. He then explained their drawbacks such as the level of knowledge needed to work with SMS which would have made it hard for the Controls community to work with. Also, due to the nature of control system operations, the users wished to have more control on when their systems were upgraded or patched and what patches are applied. CMF offers these features both for the Controls users and also for the general desktop users. At installation time, the initial boot is via the network (PXE[47]), no longer via floppy or CD except for old PCs, and the user is guided through the installation either for a Controls user or general desktop user. NICE (CERN's Windows environment) services are split into named sets of services (NSS) and CMF can pick the ones of interest. It also supports named sets of computers (NSC) and these can be assigned to sets of services. Within these sets, delegation is supported at the level of NSS, NSC and package within a set allowing individual groups or users to select non-default packages to be included or not. Packages can be published, pre-installed (but removable by the user), installed (not removable) or denied (not available to that user/computer). Installation of a package can be postponed, forced at a later time (security patch for example) or free choice. To implement this, the native Windows Add/Remove applet has been replaced by one adapted by CERN. Secondary functionality includes hardware and software inventory of the computer and loaded applications. It is currently under test in IT Department and in the Controls community and rollout across the site will begin soon. Ivan ended with a short demo of CMF.

Virtual Servers for Windows (Alberto Pace) – Alberto started with a demo of creating a couple of virtual systems on his desktop (one Windows, one Linux using SLC) and while they were being created, he started the presentation with a history of how virtual computers have long been a dream of computer scientists. As the Intel X86 architecture is becoming by far the most commonly-found system in our environments, running virtual X86 systems on real X86 systems is more attractive than previous implementations of virtual computers. In CERN there is an ever-increasing number of requests for dedicated servers running individual applications or services. But limitations of space, management overhead and the often-underused CPU load on many of these servers makes virtualisation an interesting option. The CERN team has built a number of different configurations of Windows 2003-based servers and Linux (both SLC3 and SLC4) virtual systems which can be called up on demand. The scheme uses the Microsoft Virtual Hosting Server. The user can configure the hardware down to the size of memory, the presence of a floppy or CD/DVD, the number of discs, etc. He or she can request use of the server for a finite time or long-term and more options will be offered in the future.

Scientific Linux Status and Plans (Troy Dawson) – Current usage of SL is at least 16,000 installations (total of SL3 and SL4) with SL4.2 alone reaching 5000 installations since its release in December 2005. After the US and the UK, Taiwan is the country with the most SL sites. Version 4.3 is approaching general release and they are working on 3.0.7. The e-mail list is growing and there is more use of the Contributions Area. They have announced a date for the end of life for SL3. Fermilab itself is standardising on SLF 4.2 and trying to phase out all the unsupported distributions (those before SL3). Working on a hardware and software inventory tool. They are gearing up for SL5, although they are bound by Redhat's release date for RHEL[48]5 and they realise it will not arrive in time to be packaged and deployed before LHC startup. He ended with his usual plea for more helpers to coordinate scientific and other applications. He also asked if there is a need long-term for Itanium releases or any other architecture; the answer, at least from this audience, was no. He would like a better definition of "compatibility", for example between SL, SLF and SLC but in discussion doubts were expressed if this was worth the significant effort it would need. Instead it was suggested to publish a list of tips and hints of things to avoid which might break compatibility.

[47] Pre-boot eXecution Environment
[48] RedHat Enterprise Linux

Scientific Linux at CERN (Jarek Polok) – 2100 individual SLC3 installations, 3559 centrally-managed installations and 2400 SLC3 installations outside CERN. SLC 4.3 is just coming into use after its official release at the beginning of April. As explained above, the projected release date of RHEL 5 (only next year) means that SLC4 will be the officially-supported release for LHC startup. It is planned to start migrating to it on the central clusters in September this year. SLC4 uses yum as the default updater now and is moving slowly towards a fully Kerberos 5 infrastructure. He ended with an overhead explaining the level of compatibility between SL, SLC and SLF.

Security Update

(Bob Cowles)

Bob covered a range of topics, starting with the dangers and risks of Skype, especially of becoming a Supernode when connected to a powerful network; apparently this does not happen to systems behind NAT[49] boxes. Skype is banned at CERN and monitored at SLAC. Turning to topical m atters, service providers should be concerned about the risks of a "bird flu" epidemic – if people start seriously to get infected and have to stay home, how to run the operation; what happens if they use infected home PCs to login? As usual he displayed the list of some 30 passwords he had sniffed during the week from among the HEPiX attendees. For instant messaging, he recommended OTR (Off The Record) which encrypts the data passed and supports authentication. He listed 10 tips to improve security – see overheads.

Alan Silverman
18th April 2006

[49] Network Address Translation