HEPiX Spring 2011, GSI Darmstadt IT/NOTE/2011-007

May 2nd to 6th, 2011

Alan Silverman (CERN/IT)

Release 1.1 (9th May 2011)

Summary

The spring 2011 HEPiX meeting was held in GSI outside Darmstadt. It attracted 85 attendees from 17 countries across Europe, North America, Russia and Asia (Korea, Beijing and Taipei). As usual (almost) all the overheads shown are online in Indico and readers interested in a particular topic are referred there. Also, I repeat that these notes are my personal impressions. I take responsibility for errors and omissions and I will accept requests for factual corrections.

The conference was well organised, including transport from town to the site which is located in the countryside. The room was comfortable and well-equipped. The staff was helpful and friendly. The banquet in a local hotel was excellent. Even the weather was clement.

Some Highlights

- Lustre: starting at the board dinner preceding the meeting, the support of Lustre was a hot topic.GSI are founder members of an open source movement to support Lustre independently of Oracle and this was highlighted in the first site report and in a dedicated talk.
- Very interesting presentation by one of the GSI/FAIR Cube designers, especially covering cooling aspects.
- First report from HEPiX's newest working group, on IPv6 networking

- A dedicated Oracle session including
    - an overview of Oracle Linux (should CERN not be using this open source Linux for the Oracle service instead of the paid licences for RHEL?)[1]
    - a review of some of the open source products which Oracle inherited from Sun – good news for MySQL
    - a discussion session with Monica Marinucci and 2 other Oracle managers on Oracle's attitude to open source and its use in HEPiX sites
- Update reports from two of HEPiX's three working groups – Virtualisation led by Tony, with some practical examples of implementations, and from the newly-formed IPv6 group led by Dave Kelsey
- In general, good talks from CERN attendees, especially the younger generation; clearly presented, interesting overheads; good publicity for our products and services
- Next Meeting will be organized by TRIUMF in downtown Vancouver from October 24th to 28th. It will mark the 20th anniversary of HEPiX, formed in 1991, and some form of celebration is expected. Discussions have opened for 2012 sites. Prague looks likely for late April 2012 and Asia for autumn, IHEP Beijing is favourite, Kisti in South Korea as second choice.

Introduction

The meeting was introduced by the GSI director of research, Karlheinz Langanke. He described some of the research being conducted at the Lab, concentrating the latter part of his talk on the newly-approved FAIR project. There are 16 international partners. An interesting snippet is that the project is costed at 1B "2005 Euros" so they must clearly be suffering already from the decrease in value of that currency.

Site Reports

GSI: GSI is a founding member of the European Open File System (SCE) project to support and develop Lustre independently of Oracle. Other members include many large and small centres and universities. More detail on this initiative will be given in a later talk. Usual increase in CPU and file servers since the last meeting. The 2000 cores on order will fill the last empty rack spaces so the cube[2] is eagerly awaited; however in the meantime, they are forced to create "mini-cubes" of 50 racks on two levels. A small test cube dedicated to Lustre is now in operation with a new cooling system rated at a PUE[3] of 1.1. More Lustre acquisitions are planned and they predict that their Lustre configuration will be "full" by the next HEPiX with a total capacity of some 3.3PB.

Fermilab: With the Tevatron on its final run, computing is considered vital to the scientific output of the experiments and will remain so for many years with a declared minimum support duration of 10 years. This implies increased interest in data conservation. For the post-Tevatron era, several other smaller new projects are under

[1] Apparently this has been discussed but dismissed because of fears of having to maintain a third Linux release at CERN; However there are large potential major investment savings in new licences for new DB servers. At least it needs a wider discussion

[2] See previous HEPiX site reports for details on GSI's Cube project, e.g. Autumn 2009.

[3] Power Usage Effectiveness or Efficiency

discussion and/or development such as MINOS, NOvA, accelerator-based research with Fermilab's other existing accelerators (those currently used as Tevatron feeders) and also cosmic and astro-physics. The MINOS experiment was recently hit by a fire at the mine where the detectors are based. Damage was not as extensive as first feared and repairs are underway but it is unclear when data taking will resume. Turning to computing, he showed impressive performance statistics for the availability of Fermigrid, used by all the current major experiments at the lab. They got ARRA funds[4] to upgrade the cooling in the Feynman Centre. On the negative side, an overheating incident in November 2010 caused a 16 hour shutdown and in February, network spanning problems twice took out the network for some (short) time; eventually they had to take it down for 8 hours to replace a switch. The Computer Division is expanding Apple support to cover iPhones, iPods and iPads. They will also support Android devices. Remedy is being replaced by Service Now with a target date of 1 Aug 2011. The Computing Division is being renamed as the Computing Sector with Vicky White, the Fermilab CIO, in charge. There will be two divisions in the sector.

GridKa: usual expansion with some 106K HEPSpec06 now available and more planned. They are performing some tests of many-core servers. They complain of many many hardware failures, a lot of them recurrent. HP support was particularly singled-out whereas Supermicro, supported by the Swiss firm DALCO[5], suffered rather few failures. While GridKa concentrates on HEP, other parts of the Karlsruhe centre supports other sciences. To my question of who supports their HP systems, the speaker said they call HP who outsource the repair to a local repair shop.

NIKHEF: Microsoft support is moving from WSUS and non-MS software to Desktop Central from ManageEngine, initially only for centrally-managed packages. The network is being redesigned to provide higher bandwidth and more redundancy. Single sign-on based on LDAP has been introduced. Topdesk Enterprise has been chosen as their trouble ticket tool; this allows other groups to build their own schemas.

CERN: Helge presented the report for CERN. He reported on the success of LCG in handling peaks of more than 6GB/s to tape. He referred briefly to ITIL and the new service desk and new ticketing service but refrained from details since there is a later talk. He reported on the Monitoring Review Workshop and the subsequent creation of a working group on monitoring. He referred to the new community support scenario being implemented and how it applies to smart phones, tablets, browsers and mail clients. He noted that Building 42 is equipped from the start for IP phones. The Oracle service is moving away from SAN/FC discs for physics to NAS, which was previously only used for admin databases. The new 5TB cartridges tape drives are being commissioned and he presented the dramatic write improvements gained by buffering tape marks (see later talk on CASTOR). Most of the other points in Helge's talk will be covered by dedicated talks.

DESY: their grid has grown to 5000 nodes and 4PB of disc space. They have found a local firm offering support for the open version of Grid Engine (previously Sun Grid Engine). DESY is upgrading its LQCD cluster and the conventional parallel computing facility. There is a GPU testbed in Zeuthen and a pilot one in Hamburg for photon science. Zeuthen is the European Tier 1 for the IceCube astrophysics experiment. The photon science experiment predicts identity management issues with its thousands of new users split into very many small VOs all across Europe. Similar to the report from Helge, DESY is also experiencing ongoing AFS problems (short freezes of several minutes) and they are investigating fixes suggested by the OpenAFS developers because, unlike at CERN, the latest

[4] American Recovery and Reinvestment Act funds

[5] Helge informs me that this Zurich firm had previously expressed interest in working with CERN but has never replied to a market survey.

release did not make the problem go away. The current version of Lustre crashes SL6 but fixes (from Whamcloud[6]) are rumoured to be on the near horizon. DESY is acquiring IPv6 hardware and tests are underway. It is the leading organization in Germany for LHCone.

RAL/STFC: luckily not too badly affected by the recent UK government spending review. However, the rest of the report was rather negative. As reported last time, there is still a lot of resistance among the staff to a change in email addressing but it is gradually progressing, mostly through staff retirements or replacement – so it's going to take a long time. UPS in the new computer building is still giving some problems but finally a solution is thought to have been found. Sporadic packet loss which started in December got steadily worse and was eventually traced to traffic shaping rules and fixed by removing these rules and using a hardware bandwidth limiter. Since the last meeting there has been the usual round of hardware acquisitions but some recent deliveries have had trouble passing acceptance tests and some older batches of servers are now giving problems. Inspired by CERN's successes (a quote from the speaker) they are experimenting with Microsoft's Hyper V for virtualization but they are struggling with it, partly based on their Linux backgrounds and partly from poor support from RAL's Windows team. RAL is undertaking a major effort on CernVM-FS, see later talk.

SLAC: Amber Boehnlein was hired from Fermilab as Head of Scientific Computing, starting this week to manage the support for Linux, storage and scalable databases. For the moment at least, the speaker, Randy Melen, is Deputy Computing Division Director and Amber reports to him. A committee of the 5 scientific directorates has been created to build and grow a SLAC scientific computing programme. The data centre capacity has been increased and with the addition of 2 new substations, the average power load has been doubled from 1,7MW to 3.6. There is also more cooling and a new UPS and a new generator and exploration has started for further expansion elsewhere on the Stanford campus.

CC-IN2P3: usual hardware upgrades including 36 Dell systems due next month with 30K HEPSpec06.Storage is still a mixture of HPSS, dCache, TSM and GPFS. Their new machine room is now in service, to be covered in a separate talk. The BQS to Grid Engine migration is covered in two later talks. They are looking to replace their trouble ticket system and introduce ITIL best practices.

IHEP, Beijing: possibly the first site report from this site, certainly the first since many years. Originally built in the 80's for BES, the computer centre was rebuilt in 2005 for BES III (50%), Tier 2 centres for ATLAS and CMS (20% each) and cosmic ray experiments. There are 6600 cores running a mixture of SL4 and SL5. They have HSM storage based on Castor and a shared file system based on Lustre and NFS. There are two IBM 3584 tape libraries with 5800 slots and 26 LTO-4 drives. To (an old version of) Castor they have added a file reservation component to prevent reserved files migrating to tape when disk usage is above a certain limit. Asked by Michel Jouvin why they did not move to a later version of CASTOR which addresses the same problem, the speaker expressed reluctance to change something which works and which, today, costs them little local support load. They are linked via international lines to Europe (via ORIENT/TEIN3) and the US (via Gloriad).

NDGF: they are looking at creating a new legal entity instead of being hosted by Nordunet but lots of red tape is needed, including acts of parliament. Meanwhile NDGF is shrinking with most of operations merging into EGI/NGI and most development moving to EMI. Various sites have undergone or are undergoing various upgrades, both in

[6] Whamcloud is one of the spin-offs of Oracle's take-over of SUN. Of the 72 Lustre developers at the moment of takeover, 48 have left and formed Whamcloud to support an open version of Lustre and 23 of the rest formed another firm, also supporting Lustre, leaving only 1 of the original team at Oracle. Thankfully both these firms and the newly-formed Lustre user groups (see later talk) are in communication and working together.

infrastructure and in available compute power. The NSC site is planning a new building for computing with 4 rooms totaling 2800 sq.m; only two rooms will be initially commissioned.

St.Petersburg (PNPI): PNPI covers a range of science including HEP. The Computing Division offers various services such as mail, a computing cluster, networking, etc. They have a 6 node so-called micro-sized compute cluster which runs under SGE with AFS clients (no server) and a number of XEN virtual nodes and supports some 140 users.

Diamond Light Source: first site report from here. DLS is sited on the RAL site in the UK and operates a small synchrotron with an increasing number of beamlines for its users. Their (small) computing team looks after some 250 servers and 250 workstations. They are largely independent of the main on-site RAL computing department although they do share JANET links, mail and account management. They use Redhat Enterprise Linux having decided against the free CentOS or SL. The beamlines are rather independent, own sub-network, own local storage, but connected to the central core networks and central services, including a central compute cluster of over 100 nodes running under SGE.

BNL: in late 2010 they reached the UPS-backed 920KW limit in the old facility so expansion is happening in the CDCE building where only 30% of the installed 1MW is in use. A long=-planned power and temperature monitoring system upgrade is underway. The ATLAS network had a major upgrade in January and IPv6 is under serious consideration. An ATLAS procurement is in progress for 142 Dell nodes which are much more power-efficient than installed nodes, and a RHIC procurement is due in the summer. After major problem on the Dell nodes, Dell had to replace 90% of the heat sinks. BNL also traced disk drop-out problems to a faulty fan which Dell were also obliged to replace. ATLAS have transitioned to using group quotas on Condor 7.4.2 and PHENIX will similarly move soon. BlueArc is still in use with SAS drives now replacing 4 year old FC-connected drives and BNL plans to stress-test the new release of BlueArc with ATLAS.

ASGC: added 150 six core nodes for e-science. The upgrade to Castor 2.1.10 was completed last month and the ATLAS D1T0 storage class is being merged into DPM to reduce unnecessary data transmission between ASGC and TW-FTT, which are actually on the same site. 2.3PB of newly-delivered storage is being used as a buffer for this migration. They are merging sets of four 1GbE 40TB disc servers into a single 10GbE 160+TB server to save rack space and power requirements. They are experimenting with a 10GbE cluster on 150 IBM blades and an Extreme x650 edge switch. Performance is less than with Infiniband but acceptable. On the other hand, IBM blade hardware is being blamed for cluster instability. They are testing the ceph open source distributed file system but find it buggy. They are also surveying pNFS but seek an open source solution such as EXOFS or PVFS2 and the investigation is ongoing with plans for a consultant coming from the DESY dCache team and possible collaboration opportunities with CERN's DPM team. Their various cloud/virtualisation studies include work on Open Nebula, Openstack and SL6 with a KVM hypervisor. They have developed a portal to allow easy access to the grid infrastructure for users.

PSI: the first comment was the similarities between PSI's role in Switzerland with GSI's role in Germany. New planned detectors will produce much more data than the current generation, with obvious implications for computing – major network upgrade, more compute power, more storage. They now have an Infiniband network, HPC-like storage for the beamlines and SATA 1TB discs. They have used GPFS since 2006, despite a new licensing model imposed by IBM last year. They have built a virtualisation service based on VMware on NetApp Metrocluster storage. They have 4 clusters with 12 servers each for a total of 130 VMs currently. They have a local flavour of SL but took a hit when they lost their main local developer. They run a small CMS Tier 3 cluster; only 20 SUN XEON blades, 160 cores, run under SGE with no CE. There is an SE and dCache is used.

Networking and Security

Secure Messages (Owen Singe, DESY)

Although it is a basic security technology with many benefits, message signing is not much used in HEP despite the growing inter-application use of messaging between asynchronous services. It also fits well with message queuing. He listed the benefits and drawbacks of signing and encryption and he believes the latter has predominated in HEP at the expense of the former. Signing only, or encryption only, of messages both have severe drawbacks, they really need to be used in tandem. He illustrated this with an example using S/MIME[7]. With message signing and encryption covered, he then turned to message queues and explained their advantages, in particular compared to using ssh. He described various message delivery models. He recommended AMQP[8] which is supported by many programme languages and he showed some examples using Python.

Security Update (Romain Wartel)

Romain started by emphasizing the importance of a site's reputation and how it can be, and has in some cases been, damaged by poor security. He described recent attacks on RSA, the Sony Playstation network and various newspapers. The HEPiX/academic community has had 4 serious attacks since the last meeting and he described them briefly. Three involved ssh and 1 was via a web application. He brought the forthcoming Identity Federation Workshop at CERN in June to the attention of the audience. He explained Linux rootkits and rootkit checkers. Since even the best rootkit checkers cannot detect all rootkits, a site needs a strategy to manage security risks; for example a good patching policy, a user access control policy, in depth monitoring, etc. He ended by discussing the security implications of virtualization. Two types of virtualization are that of a virtualized infrastructure with the use of trusted images which are beginning to be trusted by sites; and virtualized payloads which are generally not (yet) trusted.

Host-Based Intrusion Detection (Bastien Neuburger, GSI)

He described the principles behind intrusion detection and the three types of tools – network monitors, host-based tools and hybrid tools. Each has their drawbacks and he listed them. He then described in very great detail the open source OSSEC tool used in GSI, see overheads for details.

HEPiX IPv6 Working Group (Dave Kelsey, STFC)

As the IPv4 address allocation rate mushrooms towards exhaustion, the rise in IPv6 address allocations is an encouraging sign. 18 sites responded to Dave's original questionnaire of which 12 sites so far have offered representatives. A first meeting of the group occured last week. DESY and CERN are actively preparing for IPv6, DESY especially so although a testbed should be available at CERN this summer. However, Dave is unaware of anyone doing any work on application testing under IPv6. The first plans of the WG are to share experience, ensure no duplication of other activities and expand the group to include applications, security, middleware and monitoring experts. During 2011 they hope to deploy a distributed testbed on which the various areas can be investigated in detail. These plans should be finalized this year, leaving 2012 for testing and perhaps being ready to use the LHC shutdown in 2013 for some deployment. Further video/phone meetings of the WG will now continue on a weekly basis with occasional face-to-face meetings, the first after the World IPv6 day on June 8th when sites round the world are invited to interconnect via IPv6.

[77] Secure MIME

[8] Advanced Message Queuing Protocol

Computing Systems

Batch Monitoring and Testing at CERN (Jerome Belleman)

The first motivation for this was the huge number of nodes (>3700) and cores (>30,000) making up the batch service at CERN. Other triggers were the large number of customers and jobs per day. Current monitoring is based on Lemon but it does not cover everything, so logs still have to be checked in case of problems for example. The review started by asking the kind of questions posed on a regular basis, why is a job not running, are fair shares being properly implemented, and so on, and the shortcomings of existing tools to answer these questions. Then they defined the views required by managers, by sys admins, by users. As a target, they wished to create a pluggable framework to allow people to create their own view of their own data. Existing tools were surveyed with criteria of being independent of the batch scheme in use. The intention was to create an IT-wide tool built in a Lego-like fashion. He showed a layered diagram of the process from raw data collection to display and the tools currently being surveyed for each sub-task. A first prototype is targeted for 6 months time and he invited people in the audience to share their experience in this field.

Selecting a new Batch system for CC-IN2P3 (Bernard Chambon)

As reported at the previous meeting, IN2P3 have decided to replace BQS and this talk described how they made the replacement choice. BQS was a local development from the early 90s which was only used in Lyon and which was requiring increasing manpower to maintain and add missing features. A decision was taken in July 2009 to replace it and a market study was begun, covering all batch systems in common use across known HEP sites. One of the studies included analysis of 15 questionnaires returned by these other sites. A list of criteria was established and the different options evaluated against these – see slides for the criteria and the score achieved for each tool. Sun Grid Engine (SGE) got the highest score, closely followed by LSF. Both were installed for a short trial and SGE appeared more suitable for their needs, independent of procurement costs, and this was duly agreed in February 2010.

Setting up SGE (Bernard Chambon)

The previous speaker then continued by describing the initial testing performed, how the final installation was configured and how they solved the various problems met. Since SGE is of limited interest to CERN readers, I will not cover the details here but refer the reader to the overheads on Indico. The service entered production 3 weeks ago on 128 worker nodes and more WNs will be added for the official opening in mid-May. When John Gordon asked how they would improve the current-poor support for SGE features needed for grid jobs, someone in the audience from the ARC community suggested that ARC supports these, although he did not go as far as to say they would support the features for non-ARC sites.

OpenMP (Jie Tao, GridKa)

OpenMP is a programming model for multiprocessor systems with shared memory, developed by the OpenMP Forum. The speaker established a test of this on a virtual setup involving XEN as hypervisor on a multi-core Debian-based configuration with 1 to 8 cores. She then showed performance figures versus the number of cores with a simple test routine, mostly linear with a few exceptions. But in measuring execution times with her target NAS application, she saw these go up where she expected them to go down. She used the profiling tool ompP to understand why and discovered that some of the openMP overheads outweighed the benefit expected from multiple cores and found an anomaly in the number of loops in the execution sequence.

CMS 64 bit and multi-core plans (Giulio Eulisse, FNAL)
CMS have noticed up to a 30% improvement in some applications when converting to 64 bit execution. There is a definitive memory cost, although it is not at all a factor of 2 as sometimes suggested, more like 25-30% if measured properly. Hence the porting of code started in 2003-4 and is now considered mission accomplished such that 32 bit versions are no longer built. Turning to multi-cores, applying single core scheduling is not the answer, memory needs explode, I/O rates increase and the number of independent jobs mushrooms. Rather exploit the fact that most of the common data and code can be loaded early in the application, so create a fork at that point and leave it to the kernel which is smart enough to share these pages, the so-called C-O-W principle, copy-on-write. He showed graphs of memory usage with this scheme, a reduction by a factor of 2.5 over a simple single-core scheduling mode. This leads to whole node scheduling where the application takes over the management of the whole node to properly share the node's resources and this triggered the creation of a whole-node job submission task force. Multi-threads could be also interesting but on a practical basis it does not appear to be worth doing in the near future. Keith Chadwick noted that, as long as the same CMS application is being run, this forking and C-O-W would be automatically handled by some, at least, hypervisors, saving CMS the overhead of changing their codes. Further, if the CMS code needs to know about every different hardware configuration on which it has to work, it will be a much larger and unwieldy object. From several questioners, it was suggested that CMS should consider more use of virtualization and be less afraid of its overhead.

Performance of multi/many core systems (Manfred Alef, GridKa)
Manfred pointed out that in general core speeds are not going up (much)[9] but rather CPU performance increases are coming from multi-core systems. He therefore performed some tests on these systems and presented the results, including HEPSpec06 (HS06) results, job throughput and Ganglia performance plots. See the overheads for the very detailed tables and charts.

IT Infrastructure

Drupal at CERN (Juraj Sucik)
Now used by some 1.5% of the world's web sites, including some major entities. Since its initial installation as a pilot service in CERN in September 2010 and subsequent upgrade to version 7 in January, there are now 60 web sites at CERN using it, with positive feedback. Support comes via the community model mentioned by Helge in his site report – for example, the FAQ is maintained by the user community and the Service Desk will handle up to second level using this FAQ forum but there is no third level support. Juraj listed the various server components installed and the software configuration. Drupal is installed with security rules to ensure site separation and access to the site files is through secure Webdav protocol. The (pilot) service is integrated into CERN's SSO and Drupal roles are mapped to e-groups. There is a CDS module to import CDS records and similarly an Indico module is planned to link to events. The service is monitored via Lemon but more detailed site monitoring is planned. They hope to open a production service later this spring but there are still some stability issues with various Drupal 7 modules.

Indico (José Benito González López)
First used at CHEP 2004 in Interlaken, Indico is now used at more than 90 institutes worldwide. Current statistics include more than 130K events, 590K presentations, 770K files and around 10K visitors per day. Jose described the

[9] Performance of cores seems to be frozen at around 10 HS06

resources allocated to the service before turning to its evolution, with emphasis on the recent re-design to improve its usability and performance and to introduce new features such as room booking, integration with video services, paper reviewing and chat room integration. The final target is to cover the full lifecycle of conferencing management and execution. Despite the wide usage, they only hope to release what they still call Indico 1.0 next year(!).

Invenio (Jerome Caffaro)
Described as an integrated digital library or repository, Invenio has become the platform of choice for managing documents in HEP. Jerome showed the development line from the first days of the Library preprint service through CDS to the current state. Today there are more than 1M records, >700 collections, 18K search queries per day, more than 200 new documents every day and over 100 submission workflows. Some 4500 users have established 6250 baskets or bookmark collections. He showed the modular relationship between the CDS kernel and associated Invenio components and covered some future development plans such as new generation submission workflows and integration into other CERN services such as Drupal, Inspire, etc[10] .

FAIR 3D Tier 0 Green-IT Cube (Volker Lindenstruth)
This was a talk in the GSI seminar series and many GSI staff joined the HEPiX meeting for it; it was introduced by the Lab Director. The speaker started by showing early estimates of FAIR computing needs -300K processing cores plus 40PB of mass storage for a year's data. The current centre does not have the capacity for this. He showed a calculation by which a simple laptop needs 1litre/sec of air cooling and a 1MW computer centre needs 33,600 l/s of air, but only 50 l/s of water cooling. Assuming a PUE of 2, a centre of 1MW of computer equipment needs 2MW of power. The speaker has designed a special heat exchanger as the rear panel of a rack which uses water cooling to convert hot air from the rear of the rack and then pumps it out cold air for entry into the rack behind.

This shows the prototype cube on its steel girder.

[10] But not to the SVN service, they use GIT

This shows the rear cooling panel of a cube rack. Note the thickness of the door.

A first example is the Loewe-CSC cluster at the Goethe University in Frankfurt where 450KW were needed to cool a system rather than 3 times this power which would have been needed in a conventional configuration. The water moves at 5 cubic metres per hour; it heats up by about 8 degrees in the cooling panel and is then re-cooled in an external evaporator heat exchanger. The Loewe site is rated at 600KW of which the primary pump needs 5KW, the secondary pump needs 27KW and the PUE is therefore 1.062, or 1.082 at 450KW which is the current usage.

As a result, apart from the very front rack, the working temperature (for example in the aisles between the racks) in such a configuration remains cool. From there came the idea of cube with layers of racks mounted on steel bars. Their cube is planned to be 20 metres tall to house 890 racks. The project is almost approved (approval came through the following day) with a plan to start ground-breaking in 2012. Conceptually, they have considered a tier structure like LCG but they have many nearby user groups so they will install short-distance high performance links direct to nearby universities and a grid to more remote sites.

The new computer room at CC-IN2P3 (Pascal Trouvé)
Pascal is the new Facilities Manager at Lyon. In the existing building they have installed new transformers, a diesel generator, a second low voltage distribution panel and some other minor changes to the power circuits. Other, final, changes in the old room include a third cooling unit (600KW), bringing cooling power to a legal maximum for that space. Fire hazard precautions have been strengthened. Since these installations, power and cooling have been more stable than before. The new computing room will be commissioned in stages, Phase 1 is for worker nodes for the LHC experiments using 600KW for 50 racks and needing a total building power of 3MW. These numbers should jump in 2015 with just over a doubling of the installed power and in 2019 to 216 racks, 3.2MW of power for these and 9MW in total. Water, power and other services are delivered from the ceiling, no false floors. A single firm was chartered to design and build the three level centre, to be completed within 18 months; in the event it was only 6 weeks late. Technical commissioning is in progress with production and the official commissioning of Phase I scheduled for end May. For more technical details on the building infrastructure and some of the installed equipment, see the overheads. Planning for Phase II will start in October as well as studies into a power supply redundancy project and the provision of additional office space, which was included in the original plan but dropped for lack of funding. This last item should be merged into a pole for outreach and interaction with the public at large.

CERN's Computer Centre Plans (Wayne Salter)
Wayne showed the current, almost-saturated situation of the CC and the particular issues this raises. Experience of the local (Geneva) hosting is mostly positive but the need for several onsite interventions by CERN's sys admins and even more by some service managers is worrying in the context of a future remote hosting option. Apart from that, the main lesson is that setup, from tender to installation, took longer than expected. In parallel, we are implementing some of the suggestions from Intel from their report at the end of 2010 on improving the efficiency and performance of the existing CC. These include running at a higher temperature, running without the chillers when possible (expected to be for most of the year), using variable speed fans. Next he described the CC upgrade goals and plans. With civil engineering work starting this month and due to be complete by November 2011, the increased physical power should be available in August 2012 and additional critical power by November 2012. Finally to the topic of remote Tier 0 hosting, starting with the original plans for a building on the Prevessin site and the first Norwegian poposal. Leading on from this, CERN issued a call for what could be made available for 4M CHF/year and we have already received 23 expression of interest and these are currently being analysed, mostly by a series of visits. A go/no go decision is expected this summer and if go then a tender should be prepared and dispatched with adjudication targeted for early 2012. Installation could then start in 2013 and built up as needed.

Service Management Processes and Service Now (Zhechka Toteva)
This was a very tightly packed talk in which Zhechka described, in some detail, the studies which resulted in the choice of Service Now (SNOW) to support the new combined IT and GS first-line Service Desk. The basic configuration data needed by SNOW (people, organizational data and configuration database objects such as locations) had to be extended in a CERN context and needs to be synchronized on a regular basis. Merging SNOW authentication with CERN's SSO needed a workaround and similarly for SNOW authorization. A 3D Service Catalogue was built with both user and function perspectives (the third dimension is weight of a service element on a functional element). This Service Catalogue is used by SNOW. Email incident creation is allowed for a temporary period but it creates too many exceptions in the mapping of mailfeeds to functional elements to be handled long-

term. In the longer term, users should use the powerful user portal. SNOW incident management uses a state-driven workflow and this was extended with respect to the Service Catalogue and how the tool should be used by the Service Desk. The LCG trouble ticket system GGUS was interfaced to SNOW via a SOAP interface. Zhechka described the user portal and the Knowledge Base. Change Management is the next ITIL process to be implemented and work has started as well as integration to EDH. Eventually we will need to migrate all Remedy workflows into SNOW.

Scientific Linux (Troy Dawson, FNAL)
Troy started with the usual download statistics, noting the huge jump created by the downloads of 64 bit versions and the record number achieved since the SL6.0 release in March (2.3 million downloads of the 32 and 64 bit versions combined). This release now includes OpenAFS as a result of HEPiX requests. The expected final SL4 release, 4.9, was released in April. Security updates for SL4 will continue until February 2012 at which time they will decommission SL4. SL 5.6 is due this month with a beta now available, 5.7 should come in winter 2011 and 6.1 in fall 2011. They are looking at automatic checking for security patches for added packages Some really good news – Fermilab have added 2 new developers to the SL team. Meanwhile, Troy's manager has publicized a web survey inquiring about usage and future desires and everyone is encouraged to reply.

Version Control Services at CERN (Alvaro Gonzalez Alvarez)
There are currently 3 version control services offered by IT, 2 based on CVS and one on SVN. Shutdown of the non-AFS CVS service which was requested by some users is in progress, only 10 projects remain, and shutdown of the main AFS-based version is planned for 2013. Meanwhile the newer and recommended SVN service continues to grow and now hosts twice as many projects as the main CVS service. The main outstanding issue on the SVN service is load balancing where there are a number of small but frustrating problems. Solutions for these are being worked on as well as service consolidation and improvements to the documentation. Alvaro closed with a very short description of the Twiki service and some of its statistics.

CernVM-FS (Ian Collier, STFC-RAL)
In a phrase, the topic is this talk was virtual software installation by means of an HTTP file system and plans to use it for the future installation of WLCG software. This would have the advantage of installing software updates once at the zero'th level (so-called stratus zero installation) and files are updated on remote sites in a much shorter delay (via an hourly cron job currently) than today's manual update procedures which may take days; the same versions (and the same bugs) would be introduced everywhere which means problems need to be fixed once (although the downsize of this is that they would occur on all sites at once); and problems with scaling NFS or AFS go away. Performance was tested by PIC last month in comparison to NFS and the results were impressive, after a slow start if the cache is empty. Ian described the release and update processes in more detail as well as how to install CernVM-FS. It is in production for LHCB, ATLAS and others and will soon be added to the SL release. Question: if BaBar, for example, wished to distribute its software in a similar manner, would CERN put this on their stratus zero server? Answer: why not create a separate stratus zero and point BaBar clients there?

Oracle

Oracle Linux (Lenz Grimmer, Oracle Corp.)
This senior product manager started with the usual disclaimer that we could not commit Oracle to anything he said. He was originally a SuSE developer and one of his first tasks was to interface to MySQL to whom he moved in 2002

before later joining SUN and so eventually to Oracle where he was in the MySQL team until last month. Oracle Linux is the new name of what used to be called Oracle Enterprise Linux (or Unbreakable Linux) which was first announced in 2006, is based on Redhat sources, freely available and downloadable (but see the last 2 sentences in this section). They offer24x7 support both for this and the original RHEL Linux. They claim to be binary compatible to Redhat Linux and no counter-claims have been recorded. It is the default platform used by Oracle for their Linux release of Oracle and no RHEL is installed at Oracle. They have 6000 paying customers for Oracle Linux, including many well-known names. One major advantage of Oracle Linux in combination with Oracle RDBMS is clearly no finger-pointing. On installation, they offer a choice of kernel, the original from Redhat or a strengthened version (the so-called Unbreakable Kernel), better suited to support other Oracle products. They also bundle in some tools and performance improvements aimed at large installations. They are working on data integrity features to detect in-flight data corruption and prevent corrupted data being written. Luckily Troy asked why he could not see any 6.0 updates on their open download site and it emerges that they only make the original releases available under open source. Updates are only available to support customers!

Open Source at Oracle (Gilles Gravier, Oracle Corp.)
Another ex-Sun employee, working in Geneva, presumably in marketing judging from his presentation. He started by emphasising Oracle's commitment to open source, although not a strategy but rather part of their general software development strategy. On Grid Engine; they have withdrawn from the open source version of this and are working on Oracle Grid Engine which will be a paid project. They have uploaded the most recent code to sourceforge with no support and from now on open GE and Oracle GE are separate products. Better news for MySQL - it will continue to be released in a community edition with source and binary releases and a GPL licence. Java will also continue as open source, similarly Glassfish and Java development tools such as Netbeans and JavaFX. He closed his talk without mentioning Lustre. He referred us to oss.oracle.com for link to all Oracle open source products.

Discussion
At this point Monica Marinucci (Oracle Corp.) joined for the open discussion. Some of the questions were -

- Tony: the open source philosophy includes contributions from anyone, will Oracle accept contributions from the community? Answer: Oracle will certainly consider outside contributions if the addition is "going in the direction Oracle wants to go for its customers". "Overlapping interest" is the relevant phrase.
- OpenOffice: work has stopped but they are talking to other bodies to take it over and keep it open source
- Grid Engine: he confirmed explicitly that Oracle would no longer contribute to the open source version; all their work would be on the closed, commercial Oracle Grid Engine product.
- Lustre: not mentioned on the slides and not on the recommended web site. No one from Oracle claimed knew its status or future. I was invited to mail a question to Mr.Gravier. But in the next session (see last session in the Storage track below), it transpired that Oracle have dropped Lustre into the user community as they have done for OpenOffice.
- Monica: Oracle wants to work with the research community on topics of common interest for mutual benefit and their presence at this meeting is to listen to HEPiX as customers.
- Wolfgang Friebel: we are not powerful enough to make Oracle listen. Answer: not true, the Educational group exists to interact with the research and academic community.
- Future: risk that Oracle may stop an open source product down the road. Answer: Oracle would never just stop and close it, they would at a minimum put it into the open source community for others to continue. OpenOffice is an example.

- PSI had to buy a new licence to upgrade their properly-licensed Solaris software on their Sun hardware. Monica was aware of this and a fix was being discussed.

Storage

Evaluation of Gluster at IHEP (Chen Yaodong)

Faced with the well-known Lustre drawbacks, they looked for alternatives based on a checklist including a minimum level of performance, stability, scalability, etc. Gluster [11] was chosen for evaluation. He listed some features and the evaluation tests undertaken. Its performance and scalability were rated as satisfactory but its reliability less so. For example, striped volumes did not survive a disc or server crash although thankfully replicated volumes did. It may support replicated volumes but not replicated files. Further some actions on the global file tree (such as ls) are performed on the clients which can overload them. Nevertheless, IHEP intends to stay with Glsuter.

GridKa's mass storage (Dorin Lobontu, GridKa)

Their storage system is based on dCache with TSM as the tape library manager. Between these two entities they have developed a tape staging server (TSS) to distribute the data evenly among their 4 tape libraries. TSS must look after maintaining the integrity of the data coming from an individual VO no matter on which library/ies the data is stored. Between TSM and the physical drives is the ERRM (enterprise removable media manager) which effects dynamic library and drive sharing, queues mount requests and gathers reports and statistics, some of which were then presented.

CASTOR Status (Eric Cano)

Eric presented the latest performance statistics coming from the LCG with some key numbers. He then described some of the recent improvements, in particular the tape mark buffering and the dramatic performance improvements which resulted. The repack process will be used as a test of this new scheme and is expected to handle the current volume of data (50PB) within a reasonable time (2 years). There is a scheme in place to pro-actively scan media and data in the background to ensure its availability. In the 17K tapes scanned so far, only 4 tapes had to be recovered by the vendor and only 10 broken files could not be recovered, all discovered before the user was aware. Work done with "inefficient" tape users and general user education has greatly reduced the number of mounts and associated overhead. Eric then moved to the testing done on the Oracle/StorageTek T10000C tape drives and their 5TB cartridges. See overheads for details. Finally he presented some software improvements recently introduced or planned, in particular the new transfer manager to replace LSF in its client disc scheduler role and the new tape scheduler.

The DESY Grid Lab (D.Ozerov)

This is a small grid testbed re-using old CPUs. It has 32 nodes, 256 cores and 80TB of disc space. The speaker described some tests on various protocols including NFS 4.1, dcap, dcache and xtootd. The results are displayed graphically in his overheads. dcache and xroot gave similar results and the standard NFS 4.1 was also doing a good job. Other activities include studies on CernVM-FS, tests on DESY hardware boxes and new tests on dcache.

Lustre at GSI (Thomas Roth)

Lustre at GSI is long considered as a success story and the configuration is continually being expanded; currently

[11] See www.gluster.org for its description.

1.2PB on 3000+ nodes with 514 clients. GSI use a version ported to Debian by a local firm of “Debian kernel hackers” who now maintain this port “officially”. Since installing version 1.8.4 in September 2010 they have found no serious problems and all subsequent crashes were traced to hardware problems, largely blamed on Adaptec 5401 RAID controllers. The old Supermicro configuration for the Lustre MDS[12] was showing signs of age and was to be replaced by a new more powerful model. Unfortunately the update is non-trivial and gave some problems and eventually, after 12 days of frustratingly-poor Lustre performance and frantic debugging, they had to return to the previous configuration! They claim to have learned some lessons although they still do not understand what went wrong. After further tests, they eventually re-installed and configured the MDS on the new hardware at the third attempt and they have plans to increase further the capacity.

Distributed File System Evaluation (Jiri Horky, Prague)
Given the difficulty of configuring a system to run real user jobs while measuring only the disc performance, and the doubt on the relevance of a synthetic benchmark, the speaker prefers to use a trace and replay mechanism. He used strace to record all storage operations in his chosen benchmarks (IOzone and a real ATLAS job) and he went through the steps to replay his benchmarks on different configurations.

European Open File System SCE (Walter Schoen)
Founded Dec 2010 in Munich, non-profit, web site hosted at GSI - eofs.org (not yet ready, waiting for Walter to fill some content). Founder members include many German universities but other countries are represented also (e.g. CEA, ETHZ) plus a few firms (Bull, HP) and Whamcloud which employs 48 members of the original Lustre team. The group’s mission is not directly to be a Lustre support group but only Lustre fulfils their current needs. Xyratex Technology (UK) joined in March; founded by another senior Lustre developer and staffed by many (23) of the other members of the original Lustre team. In the US, the HPCFS and OpenSFS groups also claim to be part of the Lustre Open Source Community and they plan to merge. There is a close relationship between EOFS and OpenSFS. All meet under the banner of the Lustre User Group, most recently (April) in Orlando. Lustre 2.1 will be released by Whamcloud in summer 2011. EOFS members agreed on a Lustre roadmap, some part of which should be contributed by EOFS and OpenSFS/HPCFS members. There will be a 2 day Lustre workshop in Paris in December.

Why a European user group? The US groups are targeted at Exabyte-scale installations. Oracle: Lustre has been passed to the community, no work in Oracle now. Randy Melen: could Oracle release the name Lustre? Answer: unlikely for legal reasons.

Clouds, Grids and Virtualisation

Secure VM image transfer (Owen Singe)
Owen is a member of the HEPiX Virtualisation WG chaired by Tony Cass and he presented some of the output of that WG. The objectives are to propose methods for secure image transfer, administrators must be allowed to authorize privileged images and sites must be able to revoke images. The proposal is that an endorser signs an image list and a site VMIC[13] subscribes to the appropriate endorsement lists. Image security is guaranteed by binding the image to its X509- signed meta-data, for which the WG has proposed a first definition. Owen then stepped through an endorsement procedure including the fabrication of the meta-data, for which CERN has an

[12] MetaData Server
[13] Virtual Machine Image Catalogue

automatic method. CERN and three or four other sites already have this scheme in operation with their own VMICs and the HEPiX Virtualisation WG is promoting general adoption of this scheme.

Status report on Virtualisation at CERN (Ulrich Schwickerath)
The CERN Virtualisation Infrastructure (CVI) is based on Microsoft's System Center Virtual Machine Manager (SCVMM). Since the last meeting, CVI 2.0 has been deployed, the hypervisors now run on Windows 2008 R2 SP1, SL6 templates are available and the growth in VM use has doubled in 6 months among the 8 principle customer groups to a total of 1250 VMs on 250 hypervisors. More blade servers will be added soon to the service allowing for a further 500-800 VMs. Ulrich then described LXcloud status and plans. Since the Cornell meeting, a VMIC has been added in accordance with the previous talk along with some image distribution scripts. There are now 96 virtual batch nodes in production and a prototype of a public cloud interface using Open Nebula (ONE). Work has started looking at SLC6 which will offer some needed features, better support for ONE and is needed for an OpenStack evaluation, but there is still only a beta system available. Ulrich showed some interesting benchmarking results, see overheads. Of the 12 hypervisors controlling the 96 virtual batch nodes, half are controlled by Platform ISF and half by ONE. It is too early for a decision and it depends on the chosen future directions but ISF has created more problems than ONE although some of these are reputed to be fixed in a newly-received release. Further evaluations of both and performance benchmarks will be made as well as launching an investigation into OpenStack, a new provisioning system in which interest is growing rapidly.

Sharing virtual appliances with Stratus Marketplace (Cal Loomis, CNRS/LAL)
StratusLab is a small 2 year project with 6 partners to create an IaaS[14] service a la Amazon in the public domain. On the basis that machine image creation is a barrier to the adoption of clouds, Stratus Marketplace is supposed to facilitate the sharing of images by supplying a registry of meta-data, storage of images in the cloud and supports trust between creators, users and administrators. It has simple interfaces, a test endpoint and tools to create the meta-data. Cal then showed how users, image creators and site admins would use the Marketplace in practice.

Operating a distributed IaaS Cloud (Ian Gable, University of Victoria, BC)
The goal was to apply this to a number of projects including a HEP Legacy Data project to store BaBar data and a some astronomical survey projects whose jobs are just as embarrassingly-parallel as those of HEP and which read large data samples but output only small amounts. Why choose IaaS – primarily because of the simplicity of creating many instances of a VM image to run simple jobs[15]; but how to manage the VMs and how to schedule the jobs? From the possible solutions they chose Condor plus a cloud scheduler. Users select a basic VM created by the service and add the experiment software to create an image and then create jobs to run on these VMs. Ian then stepped through the scheduling process. There is primarily support for Nimbus and Amazon EC2. The astronomers are making steady use of the service and are quite happy. For BaBar they got funding for resources at ANL, Victoria and Ottawa, as well as funding to use Amazon and a private cloud at Victoria. Lessons learned include -

- Monitoring cloud resources is difficult
- Debugging VM problems is hard for users
- No two EC2 API implementations are the same
- Users are nicely insulated from cloud failures

Fermigrid improvements (Keith Chadwick)
Fermigrid has been described in several previous meetings/reports. They user Condor Classads to allocate user jobs to the appropriate sub-cluster and this mechanism was also used for a gradual transition from SLF4 to SLF5. They use a kind of static cloud virtualisation to offer high availability service to users. They have a locally-developed central banning service, SAZ, and a new version of this, with a new protocol which pushes the DN, VO, etc parsing

[14] Information as a Service
[15] It's interesting to compare this statement with that of Steve Thorn in the last session of the conference where image creation was the single largest problem for his users.

to the client, has been installed to improve its performance. After recent power problems, they decided to split the grid across multiple buildings.

Adopting Infrastructure as Code (Misha Zynovyev, GSI)
The mission of this PhD project is to simplify deployment and operation of scientific computing applications and their infrastructure by describing everything in code and to provide a mechanism which will allow to use external resources provided by IaaS clouds in a transparent way. Infrastructure can be deployed manually using tools, by ad-hoc scripts or by coding, and there is a similar choice of methods for scheduling and running applications. The concept of infrastructure as code came from a paper by Adam Jacob, creator of chef[16], and it should enable the construction of the environment from nothing but a source code repository, an application data backup, and bare metal resources. It requires a language to describe the infrastructure and applications, an API for every infrastructure element and a control mechanism to deploy the result. They have applied this to deploying a virtual computing cluster executing a HEP application (FLUKA) for ALICE. See talk for details. They have also tried this on Amazon EC2, an OpenNebula installation at GSI and a community cloud in Frankfurt. The results include an Alien grid for ALICE at GSI, the use of FLUKA for FAIR simulations and a testbed for a virtual environment at GSI. They hope to further expand their horizons both within GSI and outside, including targeting an entry in the Top500 Supercomputer list with an IaaS cloud facility in Frankfurt.

The UK National Grid Service – a Cloud pilot (Steve Thorn, Uni Edinburgh)
Question: should the national grid become the national cloud? A 2 year, 1.25 FTE study group was established to answer this question. Is there enough interest; are IaaS clouds suitable; and what is the ease of use? The team, from the universities of Edinburgh and Oxford created a pilot cloud, attracted some real users and performed some case studies. They based their cloud on Eucalyptus and would probably recommend this, for example over Nimbus, at least in the short-term future, simply because it works and they have practical and positive experience with it (but see later). They got some 200 user applications so the interest is there. 23 were chosen as case studies and Steve quoted three of them – ATLAS, a distributed scientific computing MSc practical course and a geographic data library. They tried to identify key IaaS features – "unlimited" resources, scalability, offloading full local cloud workload to an IaaS cloud (popularly known as cloud bursting), control over the VM environment on which the application will run, etc. He showed a matrix of use case against the key IaaS features being used showing that cloud bursting was most used, followed by scalability and control. The answer to the ease of use question was less encouraging – users had problems creating virtual images and then deploying these. It is clear that proper training and support are vital, in particular for image creation. Steve ended with some technical issues raised, for example with Eucalyptus – bottlenecks and stability in particular – although they could perhaps have installed a newer or even a commercial version of this which may have been better but they preferred to offer a stable (?) service with no upgrades. Despite the poor ease of use, they will maintain the service until at least September to allow users to complete their work and then decide on the future – maintain the current infrastructure or create a federated cloud with more institutes. His faith in Eucalyptus was questioned by members of the audience and he agreed it was debatable.

Virtualisation Working Group Report (Tony Cass)
The mailing list has 66 members of whom 10 are regular attendees at the meetings. Tony reported on the five areas chosen to work on by the WG. For trusted image generation policy, a set of guidelines was established – see overheads for the list. A scheme was designed for image creation and exchange, including expiry and revocation. Following on from this, Owen Singe and Cal Loomis have already presented at this meeting their practical implementation of some of the WG proposals and the University of Victoria is distributing VM images to CERN on a test basis using the WG recommended method. Despite these three examples, the distribution of a catalogue of endorsed images remains loosely coupled and needs to be made more coherent. Hence future work will concentrate on delivery of a distributed catalogue of endorsed images. Although CernVM-FS seems a neat solution

[16] Chef is an open source systems integration framework

for VO software distribution, VM exchange remains an interesting option and Tony asked for volunteers to help investigate this.