CERN-IT-Note-2012-031
Version 2
6 Nov 2012

HEPiX Fall 2012

IHEP, Beijing

15-19 October 2012

Introduction

This was the second HEPiX workshop held in Asia, after the Taiwan meeting in 2008. This one was even more successful in attracting attendees from across Australasia, bringing in people from all over China, as well as from Taiwan, Australia and Korea. We missed a few regular participants, in particular from DoE labs other than FNAL, but the attendance of 67 representing 27 institutes was very respectable and the agenda (60 talks, over 27 scheduled hours) was as full as usual. Once again there was commercial sponsorship (by Western Digital and Huawei) but well controlled and non-obtrusive.

The IHEP team had put in great efforts to prepare the conference and it paid off in a very smooth event. The meeting was held in a well-equipped and comfortable conference room; where projection worked well and which supported video-conferencing successfully, some 4 talks were given remotely. Coffee breaks, lunches and the Conference Banquet were well scheduled and executed. As usual the text below is mine as are any errors but all of the overheads are online and should be consulted for full details: see - https://indico.cern.ch/conferenceOtherViews.py?view=standard&confId=199025

Some highlights include –

- More sites are migrating configuration management to puppet; but Quattor is alive and well and even thriving on some sites.
- Several European sites report severe budget problems, no surprise there.
- The HEPiX IPv6 working group is very active; the Storage group has lots of plans; but the virtualisation group believes its work is done and the issue of sharing virtual images across HEP sites should now be followed up by the WLCG Grid Deployment Board.
- Networking is becoming an increasingly represented topic at HEPiX with a record total of 11 talks. Batch computing was also a popular topic (8 talks). IaaS came up several times.

- Software licences were a major issue, and not only for CERN. Several times, phrases such as "becoming independent of vendor XXXX" were quoted as reasons for new studies or new development.[1]
- Next meetings: CNAF in Bologna from 15 to 19 April 2013, University of Michigan at Ann Arbor in the week starting Oct 28th, 2013.
- On the non-conference side, Beijing continues to expand dramatically. The first time I was there, they had 1 ring road and 2 metro lines; by CHEP 2001 they were up to 3 ring roads and half a dozen metro lines. Today they have 6 ring roads and eight metro lines, which provide a fast, efficient and very cheap way to get round the city. The traffic jams are still prevalent but now as much cars as bicycles. And at rush hours it looks the entire 20 million population are trying to use the busses. Also, judging by the many large and shiny new cars, many well-dressed Chinese and some of the new building seen around the lab, there is an increasing prosperity among the population.

The conference was opened by Prof Yifang Wang who described the work of the IHEP Lab and the Chinese physics community at large.1300 staff plus students and post-docs work on the site. Apart from their own BEPCII electron positron collider, they are members of the Belle, CMS and ATLAS experiments, among others. They participate in astrophysics as part of China's exploration of space and they contribute to a cosmic ray experiment in Tibet. Future plans are still under discussion although they would like to be part of ILC, if it happens, as well as to contribute to the planned Chinese Space Station, targeted for 2020. Other plans include a neutron spallation source under construction, a planned large area cosmic ray detector and many others.

Site Reports

IHEP: registered on their local cluster there are some 1000 users but only 200 active; it has 6500 job slots and the file system is based on Lustre which has been giving some problems recently and they are applying new rules to forbid over-use of small files (see later talk). IHEP is also a Tier 2 site for CMS and for ATLAS. It has 1000+ job slots, 320TB of dCache and 320TB of DPM for data storage. This service has good reliability. An important upgrade of their network is underway to improve their lack of 10G ports. Another current upgrade involves taking their power capacity from 800KW to 1800KW and boosting also their cooling capacity by adding 28KW water-cooled racks.

Australia: Lucien Boland and Sean Crosby from the University of Melbourne gave a report on HEP computing in Australia. They work at the Centre of Excellence for Particle Physics at the Terascale, based in Melbourne, where they run a Tier 2 site for Atlas. This year they have undergone many changes, in staff, equipment and procedures, and these were described in some detail. Changes included moving the grid middleware from glite 3.2 to UMD-1, NFS to CVMFS, local disc to Dell shared storage, KVM to Citrix, a selection of Linux flavours to SL6, cfengine to puppet, and so on. All this with a very small (3 staff) computing team. There is a federally-funded cloud project for which new staff will be hired to support the Tier 2 and Australian Tier 3 sites.

CERN: Helge Meinhard then gave his report. He started with plots of webcast statistics for the 4th July Higgs announcement and he noted how well the infrastructure had coped with the load. The current LHC data rate to tape is some 1PB per week and data being sent out from CERN ranges between 1.5 and 2GB/s. The total data expected in 2012 is some 30PB, twice the design value. Plans for the remote centre in Budapest are on schedule with the network circuits ordered and first servers due to be delivered early next year. The local computer centre, Building 513, upgrade is also on track despite some issues with piping. Vidyo rollout continues, with clients available now for all supported desktops and mobile devices; ATLAS migration is well advanced, CMS is just starting

[1] We heard this some years ago in relation to Redhat wanting to sell HEP sites licences for Linux and now virtually no site apart from CERN and SLAC (if they still do) pay anything to Redhat. Oracle looks like it may be moving in the same direction - out.

and EVO quick launch will be removed from conference room PCs. The TETRA project (upgrade of the radio for the fire brigade) is (at last) progressing and should be ready for the start of LS1.
Oracle 11g migration is complete and CERN will participate in 12c beta tests. The Oracle licence needs re-negotiation. A move of TSM from a complicated RAID configuration to simple discs gave a 38% performance boost. MAC support is now given by the same team as Windows which has provided some common solutions for password management, anti-virus tools and so on. Another licence negotiation is due with Redhat. CERN is testing federated SSO with BNL and INFN and they are looking at classes of SSO. Batch is now running on 35,000 physical cores which has resulted in some hiccups; here the batch tool (LSF) licence has been renegotiated with significantly higher ceilings. There are discussions about the future of Catia and CAD systems in general. BOINC use for Sixtrack has jumped dramatically after July 4th. JIRA is now being used for CERN-wide issue tracking with some 50 projects on the central instance and more on separate instances within the central infrastructure. It will eventually replace PH's Savannah. Change Management is being added to the Service Management portfolio.

GSI: at the Lab level, work on FAIR has started in earnest. In the computer centre, the Infiniband Lustre cluster is fully working and stable; users are happy and more users are migrating to it from other services. Grid Engine is working well on their 10,000 core cluster except for massively parallel jobs where there is a question about network topology support. There are also questions about support for 300,000 cores up and overall for future support for GE, where development appears to have slowed recently. They are building a 1Tb backbone link to various universities in preparation for FAIR. They are moving from a flat Windows domain to a hierarchical view with enhanced security and more use of KVM.

RAL: the e-Science and the Computational Science and Engineering Depts have been merged into the Scientific Computing Dept which itself has 4 Divisions. Noted as being at risk in the last report, the switch gear to the power feeds is being replaced and the risk to power as mentioned then has lessened. In-row cooling is now working in the computing centre and there are now 3 enclosed aisles including one for the Tier 1 nodes. Since the April report, the storage service has been rather stable. Boundary routers have been replaced by two pairs of Extreme x670V units to provide resilient 40Gb/s connectivity for the site.

NDGF: the Nordic e-Infrastructure Collaboration has undergone a painful year of reorganisation, including the loss of some good staff. The Finland Tier 2 has upgraded to dCache FHS compliant 2.2.0. UMEA has added new dCache disc capacity. The Danish site has been in production for over a year now with several dCache pools and 27 HP servers. In Slovenia there are 2 production clusters for ATLAS and more sites are joining the Slovenian grid, although not for WLCG.

University of Michigan: this is a large Tier 2 site for ATLAS, spread across 2 locations, running some 4200 single core job slots and 10 8-core job slots. Most site services are run on virtual servers. A lot of work is going on site resilience, again based on service virtualisation. They have installed a lot of new disc capacity but are concerned about having a large quantity of disc space behind a single head node. They are keen to test Dynamic Disc Pools and they will test total I/O capacity when all the hardware is installed very soon. They will be displaying two demos at SuperComputing next month, along with Caltech and Victoria on 100Gb LHC data transfers and with BNL on high performance 40Gb data performance and these will be the first major test of the new equipment.

Fermilab: a record hot summer forced some load shedding incidents amounting to 30% or 426 hours but they managed to avoid a total shutdown. Their metropolitan network is transitioning to 100Gb/s and a 288 fibre ring has been completed, linking all major buildings on the site. Web site migration to Sharepoint has begun and ISO 20000 certification is underway and due to complete by December. More site services such as printing, logistics and deskside support are being outsourced to Dell consequent to reductions in Fermilab staff[2]. In the tape store, migration from LTO-3 tapes to T10KC tapes is almost complete and LTO-4 migration is underway. Small file

[2] Cause or effect?

aggregation is now supported on ENSTORE and dCache. Work on Fermigrid has uncovered some IPv6 incompatibilities (e.g. versions of Squid) and these are being worked on. Despite the noted staff reductions mentioned above, they are trying to recruit 13 new posts, from Director down.

LAL and GRIF: after a major interruption caused by a chiller incident in 2011, IN2P3 in Orsay has been persuaded to build a shared facility with a target date of October 2013. Budget cuts have curtailed any significant hardware upgrade other than hardware renewal, in particular for storage. The Suncluster was finally abandoned last summer (2011), due in large part to poor support from Oracle and the TruCluster 64 decommissioning is planned for this December. In Windows, the LAL domain has been migrated to the IN2P3 national domain. The EU-funded Stratuslab project to produce an open-source cloud distribution ended in May and has morphed into an open-source collaboration with most of the development team still involved. EDGL is another EU-funded project, this one to integrate desktop grids and clouds into EGI. This one ended in August and discussions on a follow-up are continuing, including the major developer. The GRIF (Paris Grid) has had very limited growth due to budget cuts but the LHCONE network is in full production. There is an initiative called P2IO to unite HEP, nuclear physics and astrophysics in Orsay in order to foster synergies. A first goal is VirtualData, an attempt to build computing expertise around a shared computing platform. This is a new horizon for LAL and it should build on GRIF experiences.

DESY: major work is happening inside the main Hamburg computer room to prepare for the computing which will be required for XFEL, in particular to establish protection ahead of the installation of water-cooled racks. DESY is another site migrating to puppet for configuration management; one of the major considerations was the widespread dissemination of the latter tool and DESY propose to share their use of it. DESY established a National Analysis Facility in 2007 but changes in Terascale management, users' requirements and funding have meant little progress so a remodelling is needed and discussions are in progress. In the production activities, Zeuthen runs Univa grid engine on a 1500 core farm integrated with 8 GPU servers and Hamburg runs S(on) of Grid Engine on a general purpose 800 core farm with some attached GPU servers. Zeuthen is also a heavy Lustre user with 925TB of Lustre storage while Hamburg uses IBM SONAS as user workspace for NFS. They experimented with a Netapp server and were impressed enough to keep it and even expand it. DESY looking at getting out of Microsoft Exchange and has started a study of three alternatives. We were promised a report at the next meeting of the reasons for the change and the result of the study. The speaker ended with some scary photos of two recent incidents – one where a power cut caused overheating and eventual explosion of UPS batteries which caused acid damage; and an incident where careless drilling in a machine room caused destruction via rusting of an expensive network card.

ASGC: since 2008 they have a new, additional, goal of offering support campus-wide so they are building a distributed cloud as an on-demand computing platform for e-Science by federating resources of various scales. AMS will be the first user group to evaluate this service. The software system for this cloud will be based on the creation of virtual machines on demand. Compared to the 6300 cores for the WLCG grid, there are some 3100 cores for the cloud. In the Computing Centre, they are moving towards fanless racks by using conduction cooling using rack walls as evaporators with refrigerant flowing through them: the first units have been installed. A local PanDa system is being setup as the core workload manager for e-Science applications and they are evaluating puppet for configuration management. As well as NFS, Lustre is used for high performance clusters. DPM is the primary grid data management tool and Castor is used only for tape services. Rucio will be integrated into PanDa as a generic lightweight workload manager. The networking is good to US and Europe and they achieved 5Gb/sec for CMS recently.

IT Infrastructure

OpenStack at IHEP: the speaker described the architecture of OpenStack and explained why it had been chosen at IHEP – open licence, open design, open development and 100% python. They use it as a virtual machine management platform. It is integrated with Torque/PBS. It is also used to manage resources at remote sites across China. They have been impressed with the response of the OpenStack developers when they have had problems to solve.

Agile at CERN: presented by Steve Traylen. The current configuration consists of --

- Puppet for configuration
- Foreman for generating kickstart files
- Hiera as a puppet data store
- Mcollective as a client for messaging of commands to remote machines
- and still some use of CDB (configuration database) for old information.

The first implementation of puppet (known as Punch) was set up more or less by hand and with little security but eventually its use expanded and it became hard to manage. This had been expected but it had provided lots of useful experience and led to a new implementation (Judy) being installed this August where more rigour was applied and where scalability was a goal. There are now some 1200 puppet agents with some 100 more added per day at the moment. It started with 2 puppetmasters and 2 foreman backends but adding more of each is easy. Git is used for storing puppet manifests which themselves are easy to prepare (perhaps too easy). Foreman is used to group similar configurations with subgroups as needed. The next steps are to deploy puppetdb for performance improvements and Mcollective but the latter is known to be tricky. The OpenStack deployment is based on the Essex code base and is integrated with CERN's Active Directory via LDAP. The target for OpenStack is production mode in time for use to install the Budapest computer centre. But Agile is not only puppet and OpenStack and it has been used to create an ACL-enabled git service and a Koji service for creating RPMs and publishing these to yum. It can also be used to run JIRA.

Integrating Lemon and Alarm Monitoring into Agile: presented by Ivan Fedorko. He described the evolution of Lemon at CERN for monitoring and how it could be transitioned into the Agile Infrastructure. Splunk has been added for data mining and visualisation and it can also be used for event management, including interaction with Service Now. There is no one single solution to replace Lemon monitoring but a shared Agile Infrastructure can be the basis to cover all monitoring domains. There is a transition plan and progress is steady.

Quattor Update: given by Ian Collier of RAL: most components have been moved to Github; it uses the Maven build system; components themselves are being updated. The compiler is able to output profiles as JSON as well as XML files which opens opportunities for data warehousing[3] although the client needs updating to accept JSON. This work is in progress, as is an option to use yum for package management. As far as configuration databases are concerned, CERN is almost the only (perhaps the only) site still using the CVS-based CDB. All but two sites use SVN-based SCDB, the second generation Quattor. Aquilon is a third generation Quattor configuration database, based on SQL + git. Limitations of SCDB include --

- the inability to test changes on an individual node in an easy fashion;
- users need to wait for other hosts' templates to compile before accessing their own;
- structured files are kept in flat files;
- and, using SVN, everyone is a developer.

[3] Indeed, with the aid of some summer students, RAL has created such a warehouse which allows a profile search by attribute and which should be ready to share with other sites shortly.

Aquilon, developed by Morgan Stanley and now working at RAL also, is supposed to correct all of these. It uses a relational database and not flat files; it defines domains of systems of similar characteristics; it defines sandboxes which can be used for testing. There is an Aquilon broker which permits easy repeat operations and automatic operations and which allows to define services between related servers and clients. RAL have worked with Morgan Stanley to implement Aquilon but they are still learning about the differences between it and SCDB.

Lync: phone, voice mailbox, instant messaging. Presentation given by Pawel Grzywaczewski. It is a pilot project to deploy softphone and unify several common communications devices. Lync is Microsoft and MAC compliant. It has mobile features for smartphones. It is mature, scalable and affordable. It has an address book which has been linked to the CERN address book; it checks for a person's availability; it has instant messaging; and a wifi option permits phoning via your laptop but without incurring expensive roaming charges. It is currently licensed for 8500 users and 400 simultaneous phone conversations. There are clients for Windows and Mac PCs and for Android, Windows and IOS (Apple) smartphones plus an interface to Linux (Pidgen) which supports everything except phone calls. There is an interface to Voice Mail on Exchange and it manages message forwarding. There are 50 users of the pilot service from different CERN Departments and he showed usage statistics. Feedback is positive so far and a production service should start in early 2013. After that, they will try to migrate voice mail from the current Alcatel system to Exchange and create federations with collaborating institutes to offer free calls between sites. A federation with Skype is possible which would allow Skype users to access, with permission, the address book of Lync users but the privacy issues remain to be investigated (and which triggered questions from Sebastien Lopienski of course).

Scientific Linux Infrastructure Changes: presented by Pat Riehecky, one of only two representatives from DoE labs. A new load-balancing plan is being introduced to offload the very overloaded central distribution servers. This will also protect against single server outages and interruptions caused by the addition or subtraction of individual servers. Caching is also under consideration where both Squid and fscached have pros and cons. Other infrastructure changes include

- the evaluation of a content management system for the web server to replace plone and permit automation of content generation
- load testing using FNAL users instead of trying to simulate the average load of 9000 requests per hour
- listserv changes to automatically unsubscribe users whose mail repeated bounces
- move to Django for errata publication which is much nicer to use and gives a more standard format.

DYNES, Building a Distributed Networking Instrument: DYNES is a US-wide cyber-instrument spanning 40 US universities to provide dynamic network circuit provisioning and scheduling. It is linked to 11 major Internet2 providers. Obviously LHC Tier 2 sites are major users as are other large scientific experiments. The hardware components (domain servers, switches. etc) are defined and software components have been defined so that sites can be self-supporting after initial deployment. These software components are thus built into RPMs, distributed by YUM and bootstrapped on to the nodes by Kickstart. Apart from the base components needed to run the various servers, the package also includes monitoring tools such as nagios, and perfSonar is coming. The result has been rather positive: sites are installed in a consistent and functional way.Nagios gives a good view of performance across the network and it is easy to spot malfunctioning sites.

Selecting a Business Process Management System in conjunction with an Identity and Access Management System: this is a joint project of DESY Admin and the HEP Dept to document various admin processes and create electronic workflows in order to handle them faster and more efficiently. The Zachman framework was used to structure the description of the processes – what, how, where, who, when and why. This results in a process map linking the service levels to the management levels. An example of a process analysis was given. To implement this you need a Business Process Management System and an Identity and Access Management System. For the first there are standards and a choice of tool. Solutions for the second are less common.

Like all labs and institutes, identity management is about people and different people at DESY can be put into classes, each with its level of access and trust and then BPM can be used for identity lifecycle management. There are particular issues in Germany because of data protection laws which restrict what information about persons can be gathered and stored. Further, like many HEP sites, DESY is a fairly open campus. Persons are assigned roles and these have to be defined and stored such that changes can be made to the definition without affecting the information stored about individuals and once again BPM can help. At this time DESY are in the process of procuring the tools they will need but it appears a slow process – studies started 18 months ago, calsl for tender were issued at the end of 2011, and benchmarking of the best candidates began recently; they are currently trying to find a consensus because there is no single best solution because, although there are several good and flexible BPM systems, few IAMs really provide identity management but rather account provisioning. So a total solution is not off-the-shelf. Nevertheless, DESY expect to make choices soon and implement the first business processes during the next year.

Scientific Linux Update: given by Pat Riehecky, a newish member of the SL team at Fermilab. Download rates have remained pretty constant over the past 6 months. SL 6.3 was released in August. V4 went end of life in February with 4.9; there is at least one known security problem in it but there is and will be no fix for this, neither from Redhat nor from Fermilab. Security patches for V5 and V6 will of course continue as normal. There have been recent problems with package dependencies which are not easy to repair. One of the problems appears to be a difference in opinion between the speaker and Connie Sieh! There is also an issue with AFS cache in 6.3 with OpenAFS; it seems only to occur with kernel ≥ 2.6.32-279.e16 and not in every case (see slide for the exact circumstance which appears to trigger this along, with their current theories and plans for a fix). It is expected that the lifetime of V5 and V6 should follow the apparent Redhat policy of 10 years. They also plan to release the Redhat Developer Toolset, complete with new compilers which can be installed in parallel with existing compilers.

ITIL at CC-IN2P3: ITIL at CC-IN2P3 started in 2010. Initially they worked in event management, incident management and internal and external communications. These were merged into a "Control Room". Training was given on ITIL V3 and some 19 persons have passed the Foundation course. A Quality Manager has been designated. In 2011, the Control Room was enhanced to cover purchase processing and work started on replacing their old fashioned and rather basic (e.g. weak search, no escalation) ticketing system, which at that time was xHelp. Desired features include assigning tickets to a team, attaching internal notes to tickets, X.509 certification, interface to GGUS, etc. And of course the tool had to be ITIL-compliant, and open source. They tested RT (Request Tracker), Mantis (mostly a bug tracker) and OTRS (complete help desk system): they were each compared to xHelp, weighted against 65 criteria. He showed a graphical view of the results which highlighted OTRS to be the most interesting although some of its scores were affected by the newness of the product for the team. Nevertheless, OTRS was chosen and installed because its pros (rich features, customisable, training available) outweigh its cons (user cannot configure notifications, French translation poor, training not free). The project started in Nov 2011 and the tool has been in production since a week at the current time. Initial ticket response time is set to 2 hours and updates are mandatory after 3 days. Even after just one week, a few bugs have been discovered which did not appear in the tests, so already they realise they must consider taking a paid support agreement (at €4000 per year for 20 service requests per year). They plan to offer user training and explanations and will look at some (not free) add-ons. The next ITIL product will be a Change Management DataBase and specifications are being written. Again tests are planned on open source tools, such as CMDBuild, OTRS and iTop, but they feel they need training. Work has also begun on --

- a Service Catalogue for users[4]
- Business Continuity, which is time-critical since they will have a 2 day power cut in December
- Identity management for both existing and new services; a project is starting, reviewing a prototype from 2009 with Sun/Oracle.

[4] It was suggested (by me) that they may wish to visit CERN and perhaps learn from the CERN experience.

JASMIN/CEMS and Emerald at RAL: Martin Bly described the various services offered at RAL. SCARF is a 2700 core system providing support to various distributed STFC facilities, part of the UK National Grid Service. Next there is a chemistry cluster based on a large SGI system. The Tier 1 site has been described elsewhere. EMERALD is a new 370 node GPU cluster hosted on 27 HP servers running SL6, Platform LSF and a CUDA toolkit, providing resources to the chemists and other sciences. The other new facility is JASMIN/CEMS for non-HEP data-intensive computing, in particular for climate modelling and earth sciences. It offers 4.5PB of global file space as well as a 4.5PB Panasas data store. He showed some performance figures for Panasis, see slides. The STFC support team also provides infrastructure services such as backups, monitoring, virtualisation (using VMware), etc, in many cases but not always (e.g. different virtualisation scheme) the same as for the Tier 1 centre. The e-infrastructure group takes a leading role in national and international e-infrastructure initiatives, for example EGI.

Batch Computing

Condor at Fermilab: given by Steve Timm. Prior to 2002, FNAL ran a variety of batch systems, all locally written, but it was doubted these would scale as the Tevatron experiments built up their data sample. For these who don't know it, Steve described briefly what Condor is, where it came from and how it has developed and spread. At FNAL, Condor first appeared on the CDF Central Analysis Facility. Now it runs on all Fermigrid and OSG batch farms in the form of grid-enabled Condor clients, Condor-G. It also now supports virtual machine submission to the more popular commercial (e.g. Amazon EC2) and open source (e.g. OpenNebula) cloud services. Condor is used exclusively by CDF, CMS and general users but D0 remains faithful to PBS. As heavy Condor users, a number of features have been added by Wisconsin at FNAL's request such as X.509 authentication, partitionable slots, integral support of glexec, extensions to cloud support and so on. As clusters grew dramatically in size, there were scalability issues in the scheduler and again Wisconsin worked with FNAL to resolve these. Currently they can handle up to 30K simultaneous jobs but the target is to get to 150K initially and then double that for cloud support. Current developments cover improved memory usage and packaging and the whole product will shortly be relabelled as HTCondor for High Throughput Condor. In the question period after this talk, a member of the audience posed the most wide-ranging question of the week, namely why were so many different batch systems being used across the sites. But on being asked by the chairman (me as it happens) which one HEP should standardise on, he refused to choose one so I guess we will have to continue with the current variety.

CERN Batch, Monitoring and Accounting: given by Jerome Belleman. Currently using Platform LSF 7.0.6 where all nodes are contained within a single cluster but with different shares for different customers. Today there are >4000 physical nodes, >60,000 cores, > 55,000 job slots and >400,000 jobs per day. The IT future is Agile and plans are being made to create a batch environment within the Agile Infrastructure project, taking the opportunity to resolve problems seen today such as slowness caused by massive job submission and query loads; ensuring fair shares; reducing a complex LSF setup; improving dynamic response to meet changing load patterns; and guaranteeing scalability for the foreseeable future such as going to 12,000 nodes, 300,000 cores. Options for future batch systems include LSF 8, Condor, Grid Engine, Torque and SLURM. For this last a test bed has been established and tests have started, especially on scalability. Turning to batch monitoring, they use Oracle, Python and Django to collect the stats, Cassandra for Fairshare monitoring, OpenTSDB for live monitoring and Splunk for checking historical usage. For batch accounting, they have been working with the APEL team in RAL to make accounting portable to other batch systems, publish local job information with the correct normalisation, simplify where possible and work with the new APEL software.

Batch at GridKa: GridKa provides resources to some 9 VOs with jobs being submitted to a single cluster of 1000 nodes, 14,000 job slots. In 2001**,** they started with OpenPBS but found many issues so moved in 2003 to PBSpro where the situation was (slightly) better but deteriorated slowly as the cluster grew so in 2010 the cluster was split into two to improve stability. Nevertheless there were severe problems several times in 2011 which took some

time to resolve with the PBS development team. Inevitably, a study began to look for alternatives and tests were initiated with Torque/Maui, SGE and Univa GE. Torque/Maui showed up too many issues and GE looked better. The purchase of Sun by Oracle raised doubts on GE[5] so they are taking a licence from Univa and an 80 node test cluster is running under this. If there are no problems, the two clusters will be merged late this year back into a single cluster running under Univa GE. Although long-term statistics show correct use of FairShares, many users complain about long response and a number of possible solutions are being researched. Despite the fact that Univa GE requires a longish learning curve, GridKa are impressed at the rapid responses coming from Univa. Operations so far have been stable and it has very flexible fairshare policies. Manfred Alef closed his talk with the benchmark tests on the latest AMD chips where he has seen some strange results, possibly he thinks because he is still using SL5. Other sites, namely DESY and FNAL, reported similar problems.

Batch BOF: a round table discussion of around 20 people took place the previous evening with the main topic being "is there a role for HEPiX in providing batch solutions?" The idea of a single HEP-wide solution was quickly rejected – differences in budget, scalability, topology. How about a single monitoring framework? How about a wiki to collect best practices? Can HEPiX combine to exert vendor pressure? Can HEPiX be a tool to move towards a common future? No single consensus emerged beyond setting up a mailing list and/or wiki to collect experiences. But it needs a volunteer coordinator to pull this together and to help schedule future presentations on batch and no names came forth[6].

Oracle Grid Engine at CC-IN2P3: this talk comes after one year of experience. The migration from BQS to OGE was completed in December 2011. They now run a single instance of OGE 6.2 with a single master running under Solaris[7] and around 750 nodes in the batch farm. There are 16,000 job slots and some 100,000 jobs per day. There is no failover server but there is an automatic restart procedure and a backup server ready to start. It has taken around 9 months to arrive at a stable server and many patches and much tuning. Parallel and multi-core jobs must run on separate nodes (without understanding why this is necessary but it works) but scheduling time is acceptable at 30 seconds on average although they have seen peaks of 200 seconds. Whereas BQS required 3 FTE to support and develop it, OGE appears to need only 1 FTE for support, plus admin and operations staff. The speaker listed the early problems found during the first months.
The lab appreciates the need for fewer support staff; being able to have a single farm; admin and configuration are simpler and more flexible; resource quota sets; parallel job integration is easier; and the documentation is good. On the other hand, getting job information is difficult and in fact information is lost when a job completes; no easy way to monitor jobs or perform post-mortems; spawning jobs is not smooth; stability could be better; there is an ongoing memory leak which demands regular pre-emptive restarts; no native support to interface to a cloud. Several times it had been necessary to call Oracle support and that was not good until they could establish a contact to a dedicated person in the development team. There is no published road map, which makes IN2P3 nervous. The most serious bugs are fixed but many lower priority ones have been outstanding for sometime. Enhancement requests fare even worse. Despite all this, the lab expects to upgrade to the next upgrade of 6.2. They will suppress multi-core dedicated nodes and integrate OpenStack. But they will also stay in contact with Univa about their version of GE.

Setting up CSP on GE at DESY Zeuthen: DESY has 4 flavours of GE: a general batch system in Hamburg running SoGE, the NAF also in Hamburg running OGS (Open Grid Scheduler, the "official" open source version of Grid Engine[8]) and 2 Univa GE batch farms in Zeuthen. There is no user authentication in any of these flavours in the default setup. This is clearly unacceptable but there are a variety of workarounds – gateway nodes, limit access to hosts without general user access, provide protected client programs or modify clients to limit access. DESY have

[5] Justified as it happens, see the IN2P3 report on Oracle GE
[6] This subject was discussed by the HEPiX Board and all board members were asked to consider it, if not volunteering themselves, then proposing names. Wolfgang Friebel volunteered himself for at least some of the coordination tasks.
[7] Asked why they still run Solaris as the master host, the speaker said because the assigned sys admin knew Solaris best.
[8] See http://gridscheduler.sourceforge.net/

adopted a “certificate security protocol” based on X.509 certificates. Users and GE daemons perform mutual authentication. It is implemented as a “security mode” in GE but note that since GE permits only one such security mode to be active, an additional workaround is needed if AFS is in use and which the speaker described later in the talk. The speaker explained the protocol in some detail and examples of how it is used, see slides for details.

SLURM for WLCG at NDGF: being adopted at four Nordic sites in replacement of IBM Loadlevelerin some places, Torque in others. In Lund they had to repair lots of bugs in the job submission procedures but they now feel they have better control over running jobs. At UMEA they feel it is scalable and efficient and fairly stable, certainly far more stable that Torque/Maui. Another site which switched from Torque/Maui commented that they found it nicer although some running job information is missing or hard to extract and the default settings are not good.

Testing SLURM: presented remotely by Giacinto Donvito from Bari. They feel they need a new batch system (new compared to Torque or LSF) to provide better support for a massive number of cores where, for example, Torque starts to creak. This is the case at Bari where expansion to 4000 cores has shown limitations in standard Maui. Maui can be repaired for this but at a cost, for example processing time in Maui itself. The current Torque daemon is also suffering from a memory leak. LSF is not an option for replacement because of licensing costs. Looking at the requirements of a batch system (scalability, reliability, fault tolerance, specific scheduling functionality, low TCO, grid-enabled. SLURM appears to fulfil all of these and it is used by many of the Top500 Supercomputing sites. He listed the functions tested such as QoS, hierarchical fair-share, priorities, pre-emption and so on. The scheduling functionality is indeed powerful and can be further enriched by using the MOAB or LSF scheduler. Security management is easy. On the other hand, there is no RPM for the install although compilation is easy; no way to transfer output files from WN to submitting host; complex scheduling policy, hard to learn. Performance however was tested under different stressful conditions and no problems were found. Finally it was successfully integrated with Cream CE. Testing will continue, including expanding it into a virtualisation framework but already it looks quite promising for medium to large farms that do not want to use proprietary batch systems. There is a need for improving testing, documentation, best practices and how-to notes.

Storage

Lustre at IHEP: raw and simulation data is stored in a 5PB CASTOR 1.7 data store; other data is stored mostly in Lustre with some authentication data in AFS. There are some 3PB of Lustre data, running version 1.8. Installation is done by Quattor and monitored by Ganglia, nagios and logzilla and also home-made scripts to convert Lustre logs into a machine readable logs. Lustre having been developed with kernel level code, there is less performance overhead, good I/O performance and scalability and it is fully POSIX compliant. All of which explains its widespread use in the Supercomputer Top 500 sites. On the other hand, only offering single replicas increase risks, failure of an OST is a major bottleneck (although this was challenged by Walter Schoen of GSI); it is not system administrator-friendly; and quotas can be reported inaccurately. It is not well suited to store small files and they are looking at other solutions for these and asked the audience for ideas. One reply from the audience was to use NFS; Walter noted that such problems were common in any distributed global file system and encouraged education of users.

Lustre at GSI: still very popular with GSI users. The original instance is being phased out, giving way to one based on newer architecture using a Minicube, with 1.4PB, 50 OSS nodes, 200 OSTs and all linked by Infiniband. They are running also version 1.8 but have plans for version 2.3. Both old and new instances are running stably with only one outstanding serious bug. Migration from the old to the new cluster is left to users but hardware migration proved to be a very tedious task. Beyond the Minicube, they are looking at the Teralink Project which connects institutes in the Rhein-Mainz region.

RAL Tier 1 Disk-only Storage: presented by Ian Collier. Castor is working well but it has some limitations such as scheduling overheads, dependence on (expensive) Oracle and a single point of failure (the name server). With CERN moving towards EOS, they also feel rather exposed. To start a study of alternatives, they made a long list of mandatory features and those which are merely desirable (see slides). There is a long list of alternatives and, in a first instance, they built a table of features for each. This reduced the list of alternatives small enough to build some test beds, namely for dCache, CEPH, HDFS, orangeFS and Lustre. In each case, plus for some of those rejected, Ian listed major pros or cons. Tests are ongoing and he presented some of the early results. So far CASTOR is the most performant in some tests and not far off for others. Nevertheless, tests continue and dCache and Lustre have advantages and disadvantages. They hope to take a decision by December and having a production service by next summer, but there are many dependencies.

CERN Cloud Storage Evaluation: given by Dirk Duellmann remotely from CERN by video-conferencing. The results are still partial but already interesting. There are two test plans for S3 implementations – OpenStack/Swift and the openlab collaboration with Huawei. Storage in this context is defined as clustered storage with remote access via a cloud protocol and a modified storage semantic. Cloud computing and storage are becoming ever more popular but how relevant are they for CERN? Are the price/performances comparable to current costs of CERN services? Tests focus on S3 as a simple storage protocol compared to Hadoop which comes with a distributed computation model exploiting data locality. But S3 does not provide scalability so something must be added. CERN's interest is in the scalability which can be obtained and the TCO of the system. On a broader scale, WLCG's interest is that the S3 protocol could be a standard allowing large sites to run their own private cloud storage and smaller sites to rent it as needed. Dirk listed some common work items, some specifically for the Huawei appliance and others for OpenStack/Swift. First results for single client access showed similar results between CERN's EOS, OpenStack and the Huawei system although some tests may need to be re-done. Server-side tests on the Huawei Appliance showed very good performance, including linear scaling, once some bottlenecks were exposed and repaired. Cloud storage performance of both local S3-based storage solutions looks comparable with current production solutions.

Huawei Massive Storage: given by Jim Hughes, a Huawei engineer. He used the phrase "massive storage" rather than cloud storage and $ per GB is the primary goal, with scalable performance and high reliability. Use cases include massive storage for data analysis, backup, archival and disaster recovery. The trade-offs to enable cloud scales include latency versus throughput, absolute consistency versus eventual consistency and having a fault tolerant data centre against having fault tolerant servers. He considers POSIX as a "last century paradigm" and notoriously hard to scale, Similarly, he claims the same for Oracle and Fibrechannel. S3 is not POSIX but rather focused on simple file storage and transfer. It has no seek/write which makes it easier but you need to write the whole file if you change a byte. There are no partitions, no linked files and directories are simulated. The fundamental storage API concentrates on the physical attributes of the disc which are defined as keys (sector, track for example) and these are stored in a distributed hash table. Each node will know about its 4 neighbouring nodes so the system scales linearly. There are only three simple operators, put, get and delete, and having atomic operations permit stateless clients.
Disc drives will not get faster but rather probably slower so data will need to spread out and we need to map S3 to a distributed hash table of disc keys. Metadata and small data are stored in 3 copies (for reliability and quick recovery) with the file name as the key. Larger data is chunked into 1MB chunks protected by an error code and every block, chunked at 1MB, is stored at a pseudo-random address (but the same every time). How to make disc storage inexpensive? Disk performance has not changed in recent years. How about using simple cell phone processors with one disc per processor? This is the basis of Huawei's nano-scale servers. The system scales linearly and he showed some impressive figures for a 60PB system. The goal is to require no backup, for example writing three copies such that the chances of failed recovery of the file are 11 9s of probability. Currently a 384 node system is installed in CERN's openlab since January and a smaller system is installed in IHEP.

Mucura, Cloud Storage at the Desktop: this is an exploratory study by an IN2P3 staffer working at IHEP. They implemented a prototype of a multi-user remote file repository backed by unstructured data stores, usable both interactively and by grid jobs. The targets are users and service providers. The vision is to provide personal space on a cloud with which you interact in the way you do today to your personal file base. But now file sharing would be much easier. He claims his proposal would be more attractive than commercial solutions such as Google Drive. It would be operated by HEP computing centres. The use case excludes directly serving I/O intensive applications. Users would start with a few hundred GB. He expects files to range from 1KB to 5GB but typically would be, he guesses, around O(100)MB. He has defined a restricted set of basic operations and the service should be operator-friendly, for example no operator intervention on file recovery. File confidentiality or not is an open question but I/O performance should not be important.
He has produced a client/server with Amazon S3 as the API. The server side is in fact 3 nodes – a front-end to expose the Amazon API, a meta-data store and the file content store. The design is modular so the back-ends can be interchanged as required. A client exists; Amazon S3 type credential can be created for authorisation; both commercial and open source options exist for the various servers but they are currently using redis for the meta-data store (with the restriction that all the meta-data for a file must fit in memory). The S3 servers have been implemented over Tornado and most of the operations are ready. Current work is to support ACLs, provide operational tools and, perhaps, file encryption. They plan to demonstrate it against ROOT, test different back ends and do performance tests. They need to explore better clients which are better integrated with desktops.

HEPiX Working Group Status: Andrei Maslennikov, live from Rome. The WG received a new block of worker nodes in July at KIT. They are now installed and being used for a new series of CPU-bound tests. A new questionnaire was distributed in late September. The first results are that there appear to be three main data stores/access technologies – dCache, Xrootd, Lustre. New systems since the last survey include home-written EOS at CERN, SONAS (IBM) at DESY and ZFS via NFS at JLab. Other, graphical, results were shown, see slides. Andrei then described the so-called Storage Laboratory at KIT. The load farm has 70 8-core nodes so can support up to 540 jobs in parallel. The use cases are a CMS Hammercloud-based job, Nova from FNAL and an ATLAS job is in preparation. He showed how the tests are performed and what can be tuned and the first results. From the CMS job, NFS V4 on SL6 looks up to 40% more performant than Xrootd. In the Nova case, NFS4 saturates significantly later than the others. These tests will complete shortly and the results will be published on the WG web site. After that, tests will concentrate on Gluster and repeating the test with the ATLAS code.

Security and Networking

Network Traffic Analysis using HADOOP: nProbe was used to compute the network flow, nfcapd to store netflow data to disc and nfdump to transform the data to readable text. All this clearly results in large data files and Hadoop was chosen to manage and distribute the data for analysis by tools such as Map/Reduce to process the large data sets and the data is visualised by drawing tools such as RRDtool and Highstock.

ZNeTS, scrutinizing your Network: in France there is a legal requirement for network providers to record data to identify users for the previous year. CNRS/IN2P3 thus developed a network traffic supervisor, ZNeTS, to perform traceability of network flows, analysis tools, detection of anomalies and collecting statistics. It is built around HTML, and is easy to build, install and configure. It is compatible with NetFlow from Cisco and it is IPv6 compatible. It measures bi-directional flows which reduces the amount of data to be stored and analysed. It is built to flag many types of alerts and these are configurable. The GUI is via an integrated HTTP web server. Inside IN2P3 there are 21 instances and at least 50 outside in France. It is free for public institutes worldwide.

IPv6 at IHEP: there is government pressure to deploy latest generation networking in China and the first IPv6 links appeared in China in 2008 and have grown steadily since. The first stage in the transition from IPv4 was router to

router tunnels between IPv6 “islands”. As IPv6 built up, it became faster to carry IPv4 traffic over IPv6 networks, especially for LHC data transfers. Methods were created to create IPv6 addresses automatically. The deployment at IHEP is dual stack, IPv4 and IPv6, with similar management and security policies. Host addresses, both v4 and v6, are issued by the respective v4 and v6 DHCP servers and based on the host’s Mac address as stored in a central database. Thus the user should be unaware of the differences. The current status is that deployment and monitoring is in production along with security and the current effort is concentrating on management issues. Longer term they will work on creating a virtual environment with OpenStack and then enable IPv6 data transfers for BEPC experiments.

IPv6 at CERN: given by David Gutierrez. CERN’s Communications Group considers that dual stack is the only viable transition option but the IPv6 services must be available at the same level as those of v4. An addressing plan has been defined with two sets of prefixes, one for the Geneva site and a second for the remote centre in Budapest. The domain field will be split by project or experiment and so on down a hierarchy, leaving 64 bits for the host address. LANDB is IPv6 since March; all information is dual stack and IPv6 is now the main navigation source. Current work is on provisioning tools, user interface, network services. User training is planned and they hope for a full IPv6 production service by 2Q2013. Devices will need to be registered to use the network infrastructure with the Mac address as the key. DHCPv4 will provide a special pool for unrecognised devices.

IPv6 in FZU, Prague: this was an update from the last meeting and concentrated on issues in the FZU computing centre. Every server should be assigned an IPv6 address. They too use DHCPv6 in preference to stateless address auto-configuration (SLAAC) to assign v6 addresses although they see problems when a NIC is changed (Mac address changes) or where a device has multiple NICs. Nagios is used for monitorng and particular checks on the IPv6 testbed include DNS name resolution for IPv6-only nodes and GridFTP uploads and downloads. They deploy separate IPv4 and IPv6 instances of nagios and wonder if they should install an IPv4/IPv6 version. Smokeping is installed for latency monitoring to traffic flows between FZU and the HEPiX IPv6 testbed. Results show similar performance to that of v4 and sometimes even better. No support in the current hardware for PXE over IPv6 and he gave a long list of yes/no support in various hardware devices, see slides. He also listed some software products which work (e.g. GridFTP, EMI UI, etc) and others which do not yet (Torque has problems, YAIM).

HEPiX IPv6 Working Group: Dave Kelsey presented the current status. He started by listing the predicted v4 address space exhaustion dates of the different regions. Asia and Europe have passed this stage; the US has a year to go. Dave is currently conducting a new site survey of IPv6 status by site. Just over one-third of the 42 replies so far reported being IPv6-enabled, others are moving towards it. Two sites reported extensive use of dual stack which is encouraging in an LCG context. Of the remaining sites, 10 have defined plans within the next year. Only two sites (FZU, South Africa) report having too few available v4 addresses although some predict problems with increasing use of virtualisation. Concerns were expressed cover IP address management, security, applications not being ready.
The (small) HEPiX IPv6 testbed has been used to test a number of middleware components and tools, mainly in data management. The UK (RAL) should join shortly and hopefully also FNAL and maybe IHEP. EGI has its own testbed for other EGI components. The tests themselves check that dual stack usage works; does the application try to use IPv6? if it fails does it fallback to v4? and so on. GridFTP and globus_url_copy work and FTS can be made to work. An IPv6 DPM has been installed. The long list of non-working products is headed by OpenAFS and dCache although the developers are working on the latter. All batch systems tested give problems at this time. The WG will build a table to show component readiness although the definition of “working” is not really a yes/no question. The WG is working closely with the EGI IPv6 testing group which is led by Mario Reale of GARR in Italy. The EMI 4 site testbed is concentrating on EGI components. Similarly, EMI has a testbed for their components. Despite a US Government mandate for front-facing services to support IPv6 by end Sept, there is no enforcement and DoE national labs are out of scope anyway. Nevertheless, the HEP-related DoE labs are making good progress. From a recent WG meeting at CERN, a new test plan has been defined and tasks allocated to various players.

Data Centre Network Changes in CERN: given by David Gutierrez. He described the process to change to Brocade routers which were required not only to cope with the expected growth plan in the number of central nodes but also to introduce 100Gb switches. Next came the firewall changes which resulted in 2 active firewall gateways rather than an active+passive combination. Finally he described the changes required for remote operation of the Budapest centre.

Federated Identity Management for HEP: presented by Dave Kelsey. This was an update from his talk at the previous meeting in Prague. He started with an introduction to FIM and why it is desirable. The federation must include the user, the service provider and the identity provider and the picture can be further confused by the addition of an authentication and authorisation agency. Example federations include Grid X.509 certificates, the eduroam (and similar) federation of academic institutes, the TERENA certification service and so on. A collaboration was started in June 2011 called Federated IdM for Research (FIM4R) which includes not only particle physics but also other sciences. There have been 4 workshops to date and they have documented common requirements and some recommendations. Since Prague, there has been more work in publicising the issues and the requirements, which themselves have been prioritised (the list was shown and commented on). They believe that pilot projects would be a desirable next step to make use of existing federations, one of which is a WLCG pilot led by Romain Wartel to build a service enabling access to WLCG resources using home institute-issued federated credentials. It should not be yet another web portal. There are some existing building blocks but there are also a lot of technical questions to be answered. Nevertheless, it is hoped to quickly establish a pilot to demonstrate proof of concept and an architectural design.

Cyber Security Update: given by Sebastian Lopienski. He used a recent hack attack to demonstrate the risk of interconnected accounts, being as weak as the weakest link or password. A survey showed that more people consider email account security is more important than your bank account security than the other way round. He then went through the traditional scary list of hacks and vulnerabilities which have arisen in the last 6 months. He recommends to disable Java in browsers, or perhaps use a different browser than the default for Java sites. Statistics show that Windows, Linux and MacOS are all now more or less equally affected by malware. He described Stuxnet which is thought to have delayed the Irainan nuclear programme by 2 years as up to 30,000 Iranian computers were damaged. It is estimated that it cost 10 man-years development. Since Stuxnet in 2010, at least 2 more recent hacks appear to be based on similar techniques. Sebastian mentioned the 35M node botnet uncovered in Europe and Steve Timms, noting that this was orders of magnitude more than WLCG resources, suggested that perhaps we should consider creating botnets rather than grids or clouds.

Service Provisioning and Security at CSTNET: CSTNET is the China Science and Technology network and is one of the earliest Chinese national networks, dating from 1994. It covers 30 provinces, 200+ institutes and services 1,000,000 users and links to many dedicated grids for different sciences. The backbone is mostly 2.5Gb/s. and they offer connections to domestic ISPs at 22G and to international ISPs at 14.6G. There is now a 2.5Gb link to GEANT as part of the ORIENTplus project. They provide a network management cloud service which includes security management but there is also a distinct network security cloud network in which they perform intrusion detection and provide an early warning and emergency response centre. There is a software suite called Duckling to allow users to build their own collaborative environment, including conference planning and operation and video-conferencing. The speaker showed some applications running on CSTNET, such as light-path provisioning, tracking for the Chinese lunar mission and services for ITER.

Networking Tools for Sysadmins at DESY: this was a list of the various network tools used by DESY.

- For monitoring router performance, they used to run home-made scripts but have now moved to netdisco, developed in the US, which was a good match to their scripts. The migration happened at Zeuthen in early 2011 and in Hamburg this past Spring. A local enhancement allows them to locate IP phones on the

Campus. Netdisco is useful to respond to simple queries such as computers in a given room, port status check, unused IP addresses, IP misconfiguration. It is IPv6-enabled.

- The second tool is netflow, coming from Cisco. It collects traffic data but only inbound traffic so all router ports must be measured to include also DESY outbound traffic. It easily handles the Zeuthen traffic of over 1000 flows per second and is reputed to scale up much higher if needed. The package includes many scripts for different offline analyses of the collected data. V9 is IPv6-enabled
- ntop is used to display snapshots of current network traffic. It is IPv6-enabled.
- nfdump reads net flows and stores the data on disc. It is IPv6-enabled.
- nfsen is a web GUI to display data stored by nfdump. It has plug-ins to enhance functionality.

Grids, Clouds and Virtualisation

High Availability Fermicloud as a Service Facility: the Fermicloud has been described in previous meetings. It is been a 4 phase project where phase 2, deploying management services and extending the initial infrastructure (which was Phase 1), is more or less complete and phase 3, establishing production services, is underway. The specifications for Phase 4, extending services to more user communities, are being prepared. They have added an X.509-based authentication plug-in to OpenNebula. Based on bitter experience with Fermigrid, they are actively working on creating a distributed fault tolerant infrastructure for Fermicloud production services, for example the cloud is split across two buildings, the network is designed to be fault tolerant, they use SAN discs, and so on. They intend to monitor carefully virtual machine activity, including detection of idle VMs. They also apply accounting. One of the initial goals of the project was interoperability, for example being a hybrid cloud and sharing resources with Fermigrid when there is a need; or to join a public cloud such as Amazon EC2 or private clouds. These options are known as so-called grid busting and cloud bursting respectively and are achieved using VMs. For the latter, cloud bursting, a joint project with Kisti resulted in a tool called vCluster.

Scientific Data Cloud Infrastructure in the Chinese Academy of Science: CAS has 12 branch offices around the country and links to 117 institutes and 100 national labs. Within CAS there is a Computer Network Information Centre to provide informatics support and within that a Scientific Data Centre which provides storage, data and high speed networking services. Faced with the explosion of scientific data, the SDC looked at cloud computing and established the CAS Scientific Data Cloud (CASSDC). There is a distributed storage system of some 22PB today, 50PB in the future, and 5000 CPU cores, soon to rise to 10,000 cores. These are spread among 12 centres spanning the entire country. The cloud hosts some 37 large databases for different sciences. As well as linking the CAS sites, there is cooperation with various international institutes. Their biggest challenges are on-demand data services, handing “big data” and integrating all the various cloud services.

EGI Federated Clouds Task Force: Ian Collier presented the results of the first year’s work. The objectives were engagement with resource providers and user communities, integration of cloud resources, and making recommendations on issues as they arrive. The deliverables were to be a blueprint document for users and resource providers on how to engage with the federated virtualised environment, and to provide a test bed. The test bed as it stands now consists of 4 services, 2 management interfaces, 9 cloud infrastructures, operated by 7 resource provider and other providers are soon to be added. Services include a marketplace which is a repository where resource providers and EGI publish metadata about, and links to, VMs which can be installed. There is an LDAP-based information service; nagios is used to provide monitoring services; and there is an EGI-based accounting service. This was all demonstrated at a recent meeting in Prague in September. Turning to the blueprint document, Ian described the current content. As of May 1st this year, the Task Force became a Task within EGI InSPIRE which means that resources can be specifically allocated to it and there is now an outreach team. He then listed some user communities. They feel they have achieved the objectives in the adoption of standards for VM and

data management, a federated model consistent with the current EGI infrastructure and, most importantly, inter-operability between multiple cloud management platforms.

Virtualisation WG Report: given by Tony Cass, WG chair. A major achievement is an agreed policy for image endorsement by the people who are trusted by the sites at which particular endorsed images run. A framework for image endorsers to publish and distribute images has been developed. CERNVM images are compatible with the HEPiX WG policies. When the WG was created, there was a need to give the user control over the execution environment but now CVMFS provides this control, even for physical machines. So the WG consider that their work is done and there exists no reason to prolong its life. Rather, in the LCG environment at least, the baton should be passed to WLCG, to the GDB for example, to use the output of the WG to fully exploit the installed resources.

Global Accounting for Grids and Clouds: presented by John Gordon, concentrating on APEL accounting. He started with a history lesson from the early days of LHC and the distribution of tasks, when accounting "landed" on RAL, and how APEL was born, nurtured and matured and has since modified to absorb data from other accounting systems. APEL now takes data for over 270 sites via APEL client software and 90 others and it is a worldwide reference point for accounting data for LHC VOs among others. Peaks reach 73M jobs per month. John showed some views from the APEL visualisation portal. He then described some of the APEL internals and how a new version should be available next year along with a new regionalised structure. APEL is now in a good place to add new accounting record formats such as the CAR revision proposed by EMI, the proposed storage record (STAR) and, inevitably, accounting records from clouds.

Cloud Computing at RAL: given by Ian Collier. Service virtualisation started some years ago based on Hyper-V hypervisors and the System Centre VM manager. They are now up to some 200 VMs, mostly for services, and growing fast. But it is based on Windows with which the team is not so familiar so a likely move to an open platform is foreseen. The RAL Tier 1 has introduced Infrastructure as a Service, based on StratusLab, to integrate grid and cloud resources, but at the moment this is still very much a prototype. However, it is good for initial testing of VMs and a number of use cases have been defined, including internal development, providing resources for other STFC departments, a possible virtualisation layer in their batch farm and for participation in the EGI Federated Cloud. It is hoped that a new version of StratusLab will help lead to a more serious service. Contrail is a new, 3 year, EU project looking at federated clouds, where STFC's contribution is to study identity management, QoS and security. Finally there is Jasmin/CEMS which is a super-data cluster targeted at the climate and earth system modelling community and which is covered in a later talk. Given the wide range of activities described, it is hoped to find some overlap and shared benefits as cloud technologies evolve.

CERN and HelixNebula, the Science Cloud: given by Fernando Barreiro live from CERN. He showed the collaboration structure of CERN, EMBL and ESA, along with resource providers. ATLAS was chosen as one of the first pilots, can ATLAS jobs run on cloud resources? The configuration had an interface to each cloud provider, which was the first problem: each offering is different, each has a different concept of IaaS. Each one requires a different CernVM format and the interfaces are proprietary. No chance of a conceptual interface, all had to be prepared by hand. The other labs have still to complete their proofs of concept so all results so far are from the ATLAS job. Running (of some 40,000 CPU days) was generally smooth with most errors being network related. The results are shown graphically in the slides. Confidentiality agreements mean that the cloud providers cannot be identified. In summary, the result was positive – it is feasible to run ATLAS jobs on the cloud, although MC jobs are best suited for this setup. But cloud technology is in its infancy and there is lots of scope for standardisation. There are no cost estimates at this time. The next steps are for the cloud providers to propose some form of federation – broker, common API, image and data marketplace, etc; improve connectivity by connecting cloud providers to GEANT; and collaborate with the EGI Federated Cloud Task Force.

Miscellaneous

Mobile web development, and CERN mobile web site: given by Sebastian Lopienski. There are around 10,000 mobile devices present on the site, mostly Apple and Android. Each has a native development and operating environment which makes support costly. Many apps are mostly simply displaying info so concentrate on web apps and for (mobile) web apps the suggestion is to develop on the server side and then use HTML, CSS and Javascript on the client side, all of which are standard and well known. You can then select your preferred development environment on the server and distribution to the mobile device is easy. He showed a table of common apps which could be converted into web apps. From this, he has created a demo CERN mobile web site and he showed some screen shots, noting that other apps are coming, including Indico. The technologies he has used are based on jQuery, a Javascript library. He is considering using PhoneGap to create native apps with web technologies, which should allow the creation of a hybrid solution able to profit from powerful native features and mobile web flexibility.

Alan Silverman
6 Nov 2012