HEPiX Fall 2010 CERN-IT-Note-2010-008

Cornell University, Ithaca, New York

1st to 5th November, 2010

Introduction

This was the first HEPiX in Cornell, brought here by Chuck Boeheim, formerly of SLAC. It attracted 59 attendees from 24 sites from 12 countries in Europe, North America and Taiwan, a healthy turnout for a North American meeting. It was held in the Statler Hotel, part of the Hotel Administration school at the University. Networking was available but while excellent in the guest rooms, it was overloaded in the meeting room and very slow on the first morning and then only improved slightly. The agenda, which had gaping holes only a few weeks before was very full and sessions had to be started earlier than planned each day to accommodate all of them. Overheads of all of the talks are available on the meeting's Timetable page. Interaction, both in questions at end of talks and in informal discussions during the breaks between sessions, was as lively as always.

Highlights

- A large number of HEP sites rely on Sun technology, including CPU or disc hardware, tape drives and robots, Solaris, the SGE batch tool and Lustre. All are concerned with the implications of the Oracle take over. DESY reported that Oracle have removed academic discounts for the Grid Engine product forcing them to move to the free – and thus unsupported – version. For Lustre some form of user consortium is in the works, in fact two of them. A special Birds of a Feather session was added to the programme.

- Once again the workshop was organised by theme and once again this allowed the Virtualisation and Storage working groups to demonstrate their value. A suggestion was made to create a new group on IPv6.
- Helge made an interesting comment about the different atmosphere between the recent CHEP (parallel stream) talks and those at HEPiX. On consideration I think one difference is that CHEP presentations are often made to get publication citations while HEPiX reports are often part of an ongoing debate. Also CHEP talks almost always start by explaining the environment, at HEPiX it is almost always assumed the audience knows the environment. This leaves much more team to get into the meat of the topic.
- Once again the agenda was pretty empty up to 2-3 weeks before but once again we filled all available slots and had to extend the length of the morning sessions.

Site Reports

LEPP – the Cornell Lab for Elementary Particle Physics: a long history of physics at the Cornell Electron Storage Ring (CESR), now de-commissioned for physics but reinvented as a test accelerator. Notable experiments were the CLEO detector and the CHESS synchrotron source. Currently they participate in CMS (they are a Tier 3 site) and CDF. They also perform accelerator research for a variety of sciences. They have a very large local IT installation, including some systems running VMS (yes, it's still alive although they have plans to replace it by Scientific Linux) used for several central infrastructure services such as mail, print DHCP, DNS and NTP plus the CESR control system. They also run some Solaris for an SGE service and for NIS and NFS services; and Digital's Tru64 for various monitoring tasks, for Apache, MySQL and other services. In total they have over 1000 computers, 2000+ network-attached devices all managed by 7 FTEs. Invenio and Indico are installed.

Fermilab – making use of ARRA[1] funds, their main computing facility, the Feynman Computing Centre, now has a third, high availability, site designed for low power (5-7 kW) racks. It uses overhead cooling only, no raised floors or dropped ceiling. The Grid Computing Centre, also broken into 3 rooms, has not changed much recently apart from the usual addition of new CPU and storage capacity. They claim 100% CMS Tier 1 availability over the past year because the speaker discounts time when no power was available! Migration has started to LT05 tape drives. A recent DoE review recommended removing UPS for scientific computing sites, stating that sites should rely on "reliable" utility power. Not yet accepted by Fermilab. Work moves ahead on ITIL with many of the processes in place (incident mgmt, problem mgmt, change mgmt and release mgmt).

INFN Tier 1 – a tender is ongoing for a 25% increase in capacity in spring 2011. More emphasis on their virtual machine service (Worker Node on Demand Service WNoDeS, described at the last HEPiX as well as at CHEP). WNoDeS is now deployed on a second INFN site. CNAF have re-christened their mass storage system as GEMSS (Grid Enabled Mass Storage System), still based on GPFS, TSM and GridFTP.

IN2P3 – a number of staff changes among the management, including the creation of a new post of Quality Manager. Latest farm acquisition was based on Dell. Migration in progress from BQS to SGE. Although this product does not seem to be an Oracle priority, a new incarnation, Oracle Grid Engine (OGE) appears to be interesting; more news soon. First real users should be live this month and a production service open next spring. A new farm, LAF or Lyon Analysis Farm, used initially by ALICE and ATLAS, is starting.

[1] American Recovery and Reinvestment Act

CERN – Helge started by describing the new Service Management structure being put in place, why we are doing this and how and where we are now. He then reported on some of the recent activities of the different IT groups. Much of the material is covered in later dedicated talks.

ASGC – the speaker reported on the performance over the past year of this Tier 1 site as well as percentage usage of the installed CPU and storage capacity. He spent some time describing the power supply system at ASGC. With a wide variety of customers for their services, ASGC are investigating the use of cloud technology to share their resources, based on OpenNebula.

NDGF – not much new since last meeting. New HPC2N machine room at Umea should handle >20kW per rack so more power being installed, upgrading from 400kW to 700 with no UPS backup. This will saturate this location's power and cooling capacity.

DESY – in the middle of redesigning itself as a centre for accelerator research, particle physics and photon physics. At Zeuthen there is an intense astrophysics programme. At Hamburg, photon science has created a set of new collaborations and new investment, and lots of new user requirements, including a request for MacOS support. PETRA III and the XFEL will have multi-PB data requirements when fully commissioned. DESY sees cluster computing making a return to the scene, including Infiniband networking; NUMA architecture is needed for photon science and an SGI is being installed this week. One difference from the LHC environment is the presence of hundreds of small groups or VOs. They expect to use virtualisation and cloud technology to ease their support load.

RAL – UK Government spending review: the good news, no further cuts; the bad news, no compensation for inflation. Implications will not be fully understood for several months, for example for the next phase of GridPP. Significant user resistance to change of e-mail addressing (from rl.ac.uk to stfc.ac.uk) and exceptions will be allowed. In their new computer room UPS problems continue but are now better understood although the preferred fix is not possible so a secondary UPS is likely to be needed. Also problems with dust such that for some time face masks were needed. Link to SuperJanet doubled to 20Gb. New disc servers will be 10GbE which may "challenge" their LAN capacity. Usual round of acquisitions, some SuperMicro, some Dell with whom they have a framework agreement. They suffered a repeat of a 2008/2009 storage acquisition commissioning failure with some of the 2009/2010 purchase but a swap of an interface card may have solved the problem. Quattor is being used more widely and shown to save significant sys admin time.

SLAC – first year of LCLS (photon science) was spectacular, good performance, happy users. Use of scientific computing is now charged back to users, which of course was not popular and one cluster was shutdown because the users could not afford it. A new charging model has been designed and appears more acceptable but time will tell. After a high level review of the department, they are looking at ITIL for the first time but are at a very early stage with the first SLAs only just in place. Another side-effect of the review is that a diesel generator, long demanded by the computing group, is now being installed. Almost immediately after DoE-enforced network traffic monitoring was implemented, staff was reminded that surfing porn sites is against the computing rules and a serious exposure of personal data was uncovered. AFS backup has been switched to TIBS from Teradactyl, only to disc for now while other Unix backups remain with TSM for the moment. The Astrophysicists (KIPAC) are introducing GPUs which are now also spreading to other groups. SLAC's latest acquisition, for LCLS, was based on how much they could acquire for a fixed sum rather than the cost of a given capacity. Finally they are looking at how to replace Sun/Solaris which is heavily used for storage servers.

Jefferson – most significant new feature is a serious investment in VMware services for production services, including some business services. They have applied power management for window systems, for example scheduling when they go into standby mode, and this has shown significant power savings. For single sign-on, they are participating in an ESnet collaboration and the Shibboleth pilot SSO server has been installed. Web services are being moved to Drupal. They have used $5M of ARRA funds to acquire high performance clusters for lattice QCD theorists. Beta tested LT05 drives successfully but soon after purchasing 4 of them, several problems occurred (tape damage, spooling off the reels) and drive replacements are now being installed.

GSI – Regarding FAIR, see the CHEP report; in summary, international contracts are signed, construction has started, for example the magnet testing facility where the computing group hope to build a demo model of their Cube 3D computing facility, creation of new departments for FAIR computing. Work continues on designing the Cube. A construction plan exists and now begins the search for funding. The first proof of concept model achieved the target PUE[2] rate of 1.1, a very challenging value. GSI are still fans of Lustre although the speaker proposed that stability issues should be discussed in coffee breaks and not in open session! Despite this, they have over 1100 hours of heavy production running on 3000 discs, 1.6PB capacity. With Oracle's take over of Sun, owners of Lustre, GSI has joined a European organisation to promote some open source tool that looks very like Lustre (see later).

LAL/GRIF – new DN 6620 storage servers at LAL, 60 discs in 4U format, 350TB usable space, good performance. Two computer rooms have been merged and there are plans for a new GRIF computing room. As did the previous speaker, Michel noted an implication of the Oracle take-over of Sun, in his case the "future" of Solaris which they use for storage servers (as well as 2 Tru64 systems). On Windows, MacAfee has been replaced by the Microsoft Security Essential anti-virus tool. GRIF continues to expand.

IRFU (Saclay) – member of GRIF although IN2P3 remains their primary computing resource. Tests are happening on the GRIF Vlan using jumbo-frames.

Prague Tier 2 site – history of acquisitions over the years and the building-up of their centre. And a tale of blown fuses.

KISTI (South Korea) – the lab is situated in the centre of "Science Alley" (6 universities, 20 government labs, 33 private labs, 824 hi-tech firms). The lab has a history with supercomputers, currently 15th in the Top 500. On the physics side, they are involved with ALICE, CDF, Belle, Star and Ligo. Their physics computing team is 14 FTE strong but there is only a single system admin (the speaker, on loan from IN2P3) and urgently seek another. For the ALICE Tier 2 they have some 219 CPU nodes and 30TB of storage. There is also a small analysis farm.

BNL – Tony Wong[3] reported increasing use of the new computing centre, doubling since the Lisbon meeting; currently at 230kW while the capacity is 2.5MW; most of the new systems are for ATLAS and RHIC. Various problems in cooling (water leaks) and power systems (2 UPS flywheel failures). New research directions include astrophysics and neutrino physics. The new RHIC systems, quad-core Nehalems from Dell, have suffered from over-heating CPU issues and failing hard discs. The CPU issued is blamed on poor assembly of the systems, in particular the absence of thermal grease which connects the CPU to the heat sink. BNL have also concerns about what to do with their Sun storage servers. On the network side, they have migrated from Cisco to Force 10 routers. Performed the first Condor upgrade in 3 years and seen improved performance with OSG for ATLAS and STAR. The DoE has

[2] Power usage effectiveness

[3] Formerly Tony Chan but he has changed his name recently

mandated a new security protocol at RACF (which comprises computing support for ATLAS and RHIC) which demands centralisation of Unix management, down to the level of desktops and printers.

PDSF at NERSC – NERSC's current projects include building a new Cray XT supercomputer with 150,000 nodes and testing the Magellan Cloud. PDSF is hidden behind these massive projects in the computer room. Apart from not being very visible, PDSF suffers by being the "little brother" and often has to fight for resources, in particular power but they finally got their new computer centre with the injection of ARRA funds. A major clean-up was done during the move, removing old equipment. They have just added 6 core Westmere systems. They have SGE for batch so they are another site affected by the Oracle take-over of Sun.

Virtualisation

Working Group – Tony Cass, by EVO link from CERN, reported on the progress of the Virtualisation Working Group established at the Fall 2009 HEPiX. After a slow start, the group has built a mailing list of 66 members from a large number of sites and has regular meetings. They had identified 5 working groups, each with a coordinator -

- Image generation policy, in particular trusted image generation
- Image cataloguing and exchange
- Image expiry and revocation
- Image contextualisation
- Multiple hypervisor support where an image could be generated which would run on multiple hypervisors. notably kvm and Xen.

Tony reported on the current status of each of these. The next stage is how to get the agreements of the working group endorsed and/or accepted by site management and the VOs and he noted a couple of initiatives which may advance this.

CernVM File System (cvmfs) to deliver experiment software to WLCG sites – Ian Collier described first the problem which he thought cvmfs could solve, namely the scalability of the distributing new releases of experiment software, load failures, etc. Cvmfs is not really a virtualisation project as such but was developed in that context to deliver VO software to CernVM. The interesting features include download on demand only, support for read-only images, local caches, reduced traffic flows. Tests at PIC were encouraging and further tests at RAL confirmed that it "just works" and they are starting to accept ATLAS analysis jobs.

Trusted VM Images – the speaker, from LBNL, described a different scheme to define and distribute trusted VM images. They propose to define a recipe to create a VM image rather than define the image itself, leaving the option for VOs to customise the images while sites can run VO-provided images without compromising local security. There are 2 sets of recipes, actually configuration scripts, one for the VO with VM specifics and one for the site with security and binding mechanisms to the local batch scheme for example. Reasons for this approach include the size and number of files which need to be distributed and the much easier task of vulnerability checking between recipes (scripts) and images. The tool used is Puppet, a configuration management language. He then went through the various steps in image definition, creation, distribution, etc.

Overview of Virtualisation at CERN and service consolidation – presented by Helge. He presented some use cases. Why do we need different platforms for virtualisation? Very different requirements – some service hosting tasks, others for very large scale batch systems with flexible provisioning. The former is based on Hyper-V, the second is undergoing an evaluation between OpenNebula and ISF (from Platform) over kvm (or Xen). Helge noted that in some longer term it is to be hoped that we can find one scheme to satisfy both needs. For the service consolidation task, he listed the short term goals and the particular issues around providing and running service hosts. After tests inside the group and IT, the voatlas cluster was chosen for the test since its CPU and I/O load on individual nodes was so low. He explained the hardware and software configuration. The result was that users are unable to distinguish between a real and a virtual machine but the lifecycle management is much simpler. The service currently hosts some 130 VM servers, growing at about 4 per month. Next, the plans are to validate this scheme for small disc servers.

CERN's image distribution for the internal cloud – Romain Wartel presented some work done in the virtualisation working group. The concept of an "endorser" was presented, each of whom publishes a VM image catalogue (VMIC). Sites then decide which images to "approve" which defines what can run at a site. A solution was devised for LXCLOUD, focusing on image transfer (a solution based on Bittorrent was chosen) and management of the images. He then showed a number of screen shots of the process. They intend to continue the tests to gain more experience and to investigate better local user management of the VMIC.

CERN's Virtual Infrastructure – status report given by Tim Bell. CVI provides the underlying support for quickly custom-building on demand long-lived VMs for users. Use has doubled already this year from 300 to over 680, two-thirds Windows although Linux VMs show a faster growth. Many hypervisors, 170, grouped into 6 "hostgroups" which allows delegation of admin privileges. Largest single user is Beams Dept; also some 130 VMs for physics. CVI is based on Microsoft's System Centre VM Manager (SCVMM) whose rich feature set resembles Vsphere. To this CERN has added browser- and OS-agnostic web and SOAP interfaces for VM creation, deletion, etc. The simple web interface was shown. The VMs are based on CERN standard templates. Working on supporting user-supplied images. Virtual machines are optionally backed up as virtual disc images, not for file recovery but for disaster recovery of the full image if necessary; only used twice this year. For Linux VMs, a new feature called Integration Components (IC) has radically improved disc I/O speeds; these drivers are not yet in the Redhat release but they are pulled in from a staging area.

Benchmarking

Evaluation of Intel Westmere and AMD Magny-Cours processers – the Intel move from the previous generation was from 45nm to 32nm and the AMD move was from 6 cores to 8 or 12. There is little to say here, anyone interested in the details is recommended to check the Indico web site to access the presentation.

Storage

Storage at IN2P3 – home directories on AFS; semi-permanent storage on GPFS; backup on TSM; tape backend on HPSS. RFIO still used as the access protocol between HPSS and dCache and Xrootd. Soon plan to move to HPSS version 7.3. A single dCache instance for 3 LHC experiments but, to solve various problems, they are considering moving to one instance per experiment. Still missing an HSM interface to GPFS and problems with too many small files. Lustre has been rejected for now because of its inconsistent data placement policies and the lack of a reliable and transparent online data migration service.

Scientific Mass Storage at Fermilab – a separate dept to that which provides home file storage, backup, etc. Their main customers are CMS and the large Fermilab experiments. Although the main Computing Dept is aiming for ISO 20000 for central services, the mass storage dept is not involved. They are implementing ITIL V2. dCache is the main mass storage system for CDF, CMS and the public service but Lustre is making an appearance, initially for the QCD theorists, later for CMS. D0 use their own system, SAM. Migrating old generations of tapes to LT05. The decisions around tape storage are complicated by the uncertainty of Tevatron run extensions. A year by year extension scheme would really make the choices difficult. As is becoming a recurring theme, they are concerned with the Sun take-over because of their reliance on 8500 tape robots and the use of Lustre.

Storage at BNL – Star uses 1PB of Xrootd space. ATLAS and Phenix use dCache, 7.8PB and 1.1PB respectively. BNL offers NFS access via BlueArc, initially with expensive Titan servers, now moving to the much-cheaper Mercury architecture. Both have proved very reliable since installation.

The NFS 4.1 initiative, first results – presented by Patrick Fuhrmann. The important new feature of interest to HEP is that NFS 4.1 (or pNFS[4]) supports distributed NFS data. A number of major suppliers requested an independent consortium to validate this new model. EMI and dCache are participating in this initiative. EMI will provide DPM, dCache and probably StoRM interfaces. Patrick then described the specific tests performed at DESY on stability, I/O throughput, interface to Root and use in Hammercloud and showed the results. He considers that pNFS is stable and able to sustain production services, performance is as good as comparable storage systems but more work is needed.

HEPIX Storage Working Group – during 2Q10, there were more performance tests and Fermilab joined the study. Andrei reported the results to the WLCG data storage jamboree in Amsterdam quoting new use cases. Since then an evaluation of a new openAFS release was conducted. This uncovered a major problem which led to another new release to fix it. New use cases are being added to the test suite, some provided by LHC experiments, and tests are being performed on them at the test site in KIT. Andrei's talk (and thus the overheads) contained a lot of detail on the test setup configuration and the tests themselves and anyone interested is recommended to read the overheads. In the questions, Patrick Fuhrmann noted that "standard" use cases from the experiments may depend on whom you get then from. He recommended making the tests also on native file schemes, in particular Xrootd which appears to perform not so well in the tests. Also in answer to a question, Jeff Altman, the openAFS guru who is becoming a regular HEPiX attendee, confirmed that some of the issues brought up by the working group were in the plans for future openAFS.

CASTOR development – Lukasz Janyst presented the first of his two talks. CASTOR currently stores around 35PB of data; CERN has 6 instances of the stager, with 1600 disc servers and 16PB of data on disc. Lukasz went through some recent improvements and then plans for the future. He noted the success of the recent (short) heavy ion run tests for ALICE and CMS. He concluded that CASTOR 2.1.9 performs well if used for what it is designed for.

High Performance Storage Pools for LHC – Lukasz then described EOS, a tool to provide a simple scheme for data analysis users. He presented the requirements. Development started in May and first tests on a prototype started in August. It is a set of Xrootd plug-ins speaking Xrootd protocol. It treats discs as JOBD (just a bunch of discs, no RAID) but it implements high reliability software within groups of nodes. He presented some of the internal detail, for example the two namespace views, by directory or by storage location. High Availability has been included from the start. The team is pleased with the results of the prototypes, is working in a new version and looks forward to a

[4] Parallel NFS

rollout for production users. Patrick Fuhrmann asked why CERN had invested such effort into a new service without seriously considering dCache, neither talking to the DESY team nor to major users of dCache. I promised to take this message back to CERN.[5]

Lustre BOF – this extra lunch-time session was called to discuss Lustre futures in light of Oracle's attitude to the product. If there is a community, or more than one, looking into this, should HEPiX get involved? At least our labs should speak with one voice if possible. Walter Schoen of GSI explained that Oracle have stated they will only support Lustre on Oracle (Sun) hardware. Julich Supercomputer Centre organised a meeting, inviting some US labs and US vendors, including Oracle. Discussions centred on creating a non-profit organisation to take Lustre back into the open source arena. US labs and vendors have plans to launch an OpenAFS-like organisation and have since done so and this will be announced during Supercomputing next week. The entrance fee is high so that they can hire a development team. The European labs decided to do something similar (!) but with lower joining fees. The launch of the European organisation should be launched before Christmas. Everyone promised to keep both organisations in step – not to fork Lustre. Oracle agreed to be the repository of the code and they would include the developments of the respective development teams but it is unclear how this will be achieved or what this will mean. What can HEPiX do? Encourage individuals to join? Several people suggested that as many individuals as possible who are interested in Lustre should join their respective groups. Certainly we should get regular reports from both new organisations at future meetings. Can we trust Oracle? Andrei may ask Monica Marinucci (who used to work for him) if Oracle will send someone to talk at a future HEPiX meeting.

Networking and Security

IN2P3 new network architecture – Despite an upgrade in mid-2009, e.g. doubling the backbone capacity, they are starting to see bottlenecks. A full network analysis was performed with the aim not only of improving bandwidth but also to build-in future scalability. He showed the new architecture based on a new core switch, a Cisco Nexus 7018 which has an 80G backplane per slot. Because of its central role, it has 4 power supplies backed by 2 UPS's. Initially only 6 of the 16 slots will be used. He then described the various steps in the preparation for and eventual switch to the new setup. The preparatory efforts paid off in a smooth transition. Core capacity has increased by 250%, paths have been shortened and bottlenecks removed, they have moved from a flat to a starred network which suits their needs better and they have a clear upgrade route when needed. Next come tests with 100G, e.g. on the Renater link to CERN and Ciena.

Single Kerberos Service at CERN – On behalf of John Hefferman, Lukasz Janyst explained how we plan to merge our two parallel Kerberos services, one on Windows and the second on Linux and MacOS. The aim is to migrate to the Windows-based service. He described the various steps and some of the particular problems to be solved. The timeline calls for the Heimdal KDC to be switched off in spring 2011.

Security Update – presented by Romain Wartel. He explained in some detail the most "popular" attack vectors, how botnets are built for example, the "business" of buying and using malware. He then moved on to Linux rootkits and then discussed the impact of grid computing on security. Thus far, no security incidents investigated by the team have been caused by the grid itself; conversely, the grid community has helped discover and contain the incidents which have occurred. He described the Pakiti tool used for monitoring security of grid nodes and work

[5] In offline e-mail discussion with the team back in CERN it transpired that dCache had been considered but rejected. And various team members had discussed several times with Patrick.

done to detect and isolate compromised user accounts, probably the single largest attack vector. He ended with a scary demonstration of session stealing with the Firesheep browser add-on, noting how many application Ids he had uncovered during the week of HEPiX so far (some 45 accounts in various apps such as Google, Facebook, etc).

IPv6 at INFN – this talk consisted of many questions.

- Why? One reason is to regain transparency between transfer layers – remove the need for NATs. Another is to introduce some of the inherent security features in V6.
- When? Possible now but almost no one is doing it. Will there appear a killer V6-only application?
- How? Dual stack approach – doubles network support load. IPv6-only nodes? Need IPv4 on IPv6 tunnels. What about IPv6 on IPv4 tunnels?
- What exists for address allocation, how to configure the transport layer.
- What monitoring tools exist.

Conclusion – we cannot ignore IPv6; it will not ease the network security problem; the transition from IPv4 will be difficult; we need to create best practises. HEPiX can help but in the meantime the speaker apologised for asking so many questions without firm answers.

HEP and IPv6 – presented by Dave Kelsey. He went through the history of IPv6, noting that the US Government has recently issued an official directive on the transition to IPv6, promoting its adoption with some very aggressive timelines – public services to be IPv6 by Sept 2012 and internal services by Sept 2014. There is an estimate that the global v4 address space will be exhausted on 9th June 2011 and the regional address space by 22nd Jan 2012. Dave then went through the questionnaire which he had recently published to the HEPiX mail list. Of the replies, half of the (13) individual sites have testbeds and all the (six) NRENs support it. On applications which are compatible with v6, gLite is claimed to be 99% compliant (but not tested), OpenAFS should be by Autumn 2011 (at the earliest) and Condor is working on it. Only one site (a university) reported v4 address problems, blamed on a peculiar setup there. Virtualisation was cited as a possible cause of V4 address space exhaustion,. A number of other concerns include immaturity of the software and tools, lack of vendor support, time needed for the transition. What can HEPiX do to help? Dave suggests creating a HEPIX working group on IPv6 with an initial mandate to perform a gap analyse of the problem, create a distributed HEP testbed and eventually propose a timetable for implementation and deployment. No immediate decision, delegates were asked to contact lab management.

Grids and Clouds

Caching in ARC – a presentation of the caching scheme in the NDGF grid middleware with examples taken from ATLAS jobs.

L-Grid – a light portal to access the grid, developed in INFN. Designed to permit newcomers to submit jobs to the grid.

VOMS and VOMRS convergence –a presentation given by phone from Italy. Today there are two tools to manage VOs, VOMS Admin used by small VOs and the more flexible and mature VOMRS used for large groups. The talk discussed changes to the former to add the features of the latter. The final phases of this work are currently under test.

CloudCRV – how to deploy your cluster into a virtual cloud, a presentation from LBNL. Any cloud user can set up their own cluster but what to do with 100 computers and how to scale this on demand? They define a Virtual Cluster Appliance which contain multiple VMs, multiple apps and defines the relationship between these. These VCAs consist of a set of scripts and specifications. Cloud CRV is a tool to work with these scripts to configure the VMs.

Fermicloud – Fermicloud considers itself as a (scientific) Infrastructure as a Service cloud. They use SL5 with both Xen and Kvm. Originally made up of old systems which were fairly unreliable by themselves, some modernised now; currently 23 physical nodes, plans for 13 more next year. Used by grid and dCache developers, the LQCD testbed group and a dark energy experiment. After an evaluation of hypervisors, they chose Kvm as the preferred direction but retain some Xen systems, especially for DB and I/O applications. For cloud frameworks they evaluated Eucalyptus, Nimbus and OpenNebula. Keith showed the results. The first has scheduling issues; Nimbus has issues with privileges; the third has poor generalised document and default security is poor. They currently have a few Eucalyptus and some OpenNebula systems. Fermicloud is being used to analyse some storage solutions such as BlueArc and Lustre accessed from Kvm. A phase 2 has been approved to add more low-CPU-load services. They are investigating the Amazon EC2 interface, in particular authentication issues (GSI certificates).

CERN Internal cloud infrastructure – Ulrich Schwickerath gave a status report. Like the previous example, this is also considered as an IaaS cloud. Initial deployment is to deliver IasS internally to IT service managers and not (yet) to end users. And for now only batch service customers to get started. Supports Xen and Kvm but Xen being dropped because doubts of support in Redhat and SL 6. For a central scalable provisioning system they are looking into a commercial product from Platform, ISF, and the open source OpenNebula (ONE) and Ulrich showed some comparisons. No decision yet. Customers create a so-called Golden node which is a centrally-managed VM which is a clone of a batch node and which is used to clone worker nodes. VMs are limited to 24 hours lifetime (with clean run-down and a scheme to fill empty slots up to 24 hours). Feasibility studies on scalability showed some issues with ISF at 10,000 VMs but this is now understood. ONE went to 16,000 nodes. Some 16 Kvm VMs have been running for a month as public batch systems with real payloads and they hope to scale this up gradually.

StratusLab – a cloud-like resource delivery for production grids – Michel Jouvin described this work going on at LAL. Michel demonstrated some complementarities between the grid and cloud models. The project started in June with 6 international partners, including OpenNebula, with the goal to provide a coherent open source private cloud distribution and an open source API to access clouds. It is planned that there should be limited development, because there are limited resources, and mostly integration of existing tools and methods. The first results are an appliance repository to store stock images and grid service images (appliances) and a testbed at LAL. As future challenges, Michel sees the need to build more trust between the actors at the same time as meeting the expectations of both users and administrators.

Magellan at NERSC – a cloud prototype funded by ARRA funds to investigate use of clouds for a variety of DoE scientific applications, not only HEP; the project has two partners, the other being Argonne. They first started with a user survey, what makes sense to run on a cloud, what features are needed? What are the security implications? Does it make sense to distribute a cloud over multiple sites? What are the real costs? 720 nodes, over 5000 cores, have now been installed since February but few real users so far. They use Eucalyptus which has an API compatible to Amazon but there are many problems. They also offer a Hadoop stack. Users like having root access to the VM, allowing personal tailoring of the images but of course this causes problems for the admins. This has led to some user selection but they invite all DoE grant holders to contact them to join the evaluation.

DataCentres and Monitoring

Batch Infrastructure Resource at DESY (BIRD) – Thomas Finnern described a layer over Oracle Grid Engine to bring together local support and know-how and offer resource sharing for the "poorer" user groups. AFS is used as the shared file system, dCache for mass storage and they have a growing cluster of currently 500 nodes running Scientific Linux. Any DESY user can access BIRD and some 250 from 25 user groups do so. BIRD selects the queue for the users' jobs depending on resource demands. It implements a fair share policy as well as quotas. He noted that Oracle have removed academic discounts for the Grid Engine so DESY are running on the last free release of Grid Engine and considering moving to the Open Grid Engine version. In the discussion it became clear that different branches of Oracle have different price policies for Grid Engine, for example IN2P3 are negotiating a package deal with Oracle France while Oracle Germany will only offer a deal if the site uses Oracle (Sun) hardware!

LSF Scalability Tests – presented by Ulrich Schwickerath. Given the size of CERN's CPU farms and job queues, we are starting to see slowing response times on job queries and virtualising the farm can only exacerbate this. On the other hand, using virtualisation permitted this scalability test on a farm of 16,000 worker nodes using only 480 physical nodes. Ulrich detailed the configuration and test methods. The first lesson learned was that too-fast VM registration saturated DNS updates. The LDAP query rate was very high and we may need to add more LDAP servers in the future. There were issues in the LSF resizing of the farm and with the master batch demon and above 10,000 nodes, LSF had trouble filling all available job slots, some 4-5% remained free. With Platform's collaboration, we were able to demonstrate factors of 20 times the number of jobs and 4 to 5 times the number of nodes compared to today's production load. However these seem to push LSF to, and perhaps beyond, its limits and if we arrive at around 20,000 nodes we may have to re-think our batch farm policy.

CERN IT Facility Planning and Acquisition – Olof Barring described the relations between the various players in the procurement cycle. He noted the difficulties of massive deliveries and how no new system can be installed in the machine rooms unless something is removed. He then explained a little about CF's inventory scheme and how this is used to track not only vendor interventions and hardware failures but also power trends. Finally he noted the problems dealing with low margin suppliers – missed repair delay targets and occasional bankruptcies.

GPUs and Lustre at Jefferson - JLab recently added a large GPU cluster for LQCD via ARRA funds. In total they have over 530 GPU nodes with 200,000 cores. The attraction for the theorists was their FP performance which greatly out-guns that of the host CPUs; in dollars per MFlop terms they see a 10 times factor. Sandy described the configurations of the Fermi GPUs chosen. CUDA is still used for preparing the software model but they are looking at openCL, despite the fact it is programmed at a much lower level. Initial reviews shows that the GPUs are largely (up to 80%) idle, partly because of memory bandwidth limitations, so having a GPU expert programmer can pay dividends. Turning to Lustre, JLab installed 300TB on commodity hardware, again with ARRA funding. After an original stable run-in period, when load was recently applied, many problems began, mostly system hangs, and a new configuration was required, moving from a single RAID to multiple RAID pools and installing a new Lustre release. During this upgrade, a particular file system check failed and 200,000 small user files were lost. The good news is that since the upgrade completed, only a week ago the speaker admitted, the service has been stable. Bottom line: Lustre upgrades take a long time.

Quattor update – presented by Ian Collier of RAL. At RAL, about half of the disc servers and all 700 batch worker nodes are now installed with Quattor and the remaining disc servers will use it soon. They estimate a saving of up to half an FTE. They use the Quattor Working Group format, at least for those machine types supported by QWG. For other types, they are contributing their configurations back to QWG. They hesitate to move their existing Castor service to Quattor but they will use it for a new instance of Castor. He then philosophised about why the recent Quattor workshop in RAL was so successful, why there is such a good community spirit, apparently undaunted by the failure of their FP7 bid.

IN2P3 infrastructure improvements – there have been improvements to power distribution in the existing computing room, the addition of new transformers and a diesel generator and the cooling power was uprated to 1800kW. But since the cooling requirement is 1600kW, there is no redundancy and no further capacity can be accommodated. But there were no power problems this summer. There is a project for a new 650 sq.m computer room, initially dedicated to new worker nodes. A single contractor was chartered to design and build the centre within a period of 18 months. Due to budget constraints it will start in March next year with limited power (600kW) but the modular design will allow staged increases, up to 3.2MW eventually in 2019. No raised floor, services routed from above. There is a plan to have fully independent, fully redundant power supplies but no decision is taken yet.

CERN infrastructure improvements – Wayne Salter presented our plans to upgrade the existing installation in B.513 from 2.9MW to 3.5MW useful power, to increase the available critical power to 600kW and to restore UPS redundancy for both critical systems and physics.

Operating Systems and Applications

Windows 7 at CERN – Tim Bell explained the roll-out plans. NICE Windows 7 is available on 11 supported laptops and 9 desktop models. He explained the plans to migrate the large XP and smaller Vista populations but some 20% of the XP systems are under-configured for Windows 7. Tim then presented Remote Desktop Gateway to permit access to one’s Windows Desktop from “abroad”. It’s been under test since last year. It uses https protocol with Active Directory access control. 450 people have registered to use it.

Exchange 2010 at CERN – Pawel Grzywaczewski explained how it is proposed to upgrade from Exchange 2003. Exchange 2010 has a new improved webmail interface. From the point of view of the support team, a leading new feature is the high availability feature DAG, Data Availability Group, each mailbox database copy is hosted on 4 separate disks/servers. In case of failure or maintenance operation users' connections are automatically redirected to another available database copy - this operation is transparent for users and they can still access mailboxes without interruption. Deployment of a pilot service has started. There are known problems with Alpine and Thunderbird IMAP and Macmail clients, all of which have been reported to Microsoft. I expressed concern about making mail archiving too easy if there are no quotas on this and someone from Cornell claimed there is a specific Exchange archive quota which can be switched on.

Anti-spam at CERN – Pawel then turned to the subject of spam fighting. CERN gets 1 million messages, up to 95% of them spam. Many CERN addresses are forwarding addresses and CERN cannot afford to forward spam for fear of being blacklisted. If spam is recognized, it is rejected, never silently thrown away. The scheme was based on the built-in Exchange tool but has now moved to Forefront Protection 2010 from Microsoft. He explained the filter mechanism, based on “fingerprints” of known spam messages. 94% of the rejected messages are rejected based on the message source (blacklisted senders), 4% by the protocol filter and only 2% due to inappropriate content. The tool also has 5 different anti-virus engines which scan messages in parallel. The introduction of Exchange 2010 will permit sender whitelists as well as blacklists.

A Distributed C/C++ Compilation Service (lxdistcc) – Peter Kelemen, who won the best-dressed speaker of the week prize, presented this new tool, developed to get round very long compilation times of the large codes being produced at CERN these days. It came from an open source, almost a *de facto* standard, product, distcc, already heavily used in CERN. In the software cycle, distcc can help with the compile and assembly stages. He explained how it is implemented. The clients (the user's machine) and servers (lxdistcc) communicate via TCP, shared file services are not needed. Distcc only helps if a compile/assembly task can be parallelized and the user must specify the parallelization. A CERN technical student has added Kerberos authentication, white- and blacklists and log timestamps and these have all been posted back to the source tree.

Tools used by the SL team – Troy Dawson presented some of the tools he and Connie Sieh use or are considering using to package Scientific Linux. Many of these come from the Fedora work flow.

- Koji is used to build the RPMs; they are experimenting with certificates to grant external users the right to access the SL build system
- Bodhi moves RPMs between tags and keeps track of releases and updates
- Mash is used to sign RPMs, gather all RPMs from a tag to create a repository and update a repository as RPM updates are built
- Revisor builds distributions from one or several repositories; it has a GUI and a text-driven command mode. Troy encourages sites building their own sites to consider it. (Pungi does something similar but Troy prefers Revisor.)
- Spacewalk is the open source version of rhn or Redhat's source repository

Scientific Linux – Troy's regular update. The team (both of them) has been moved to a different department and there is a possibility of reinforcements. SL 5.5 was released in May; it included XFS Utilities and a new version of OpenOffice. SL6 is in preparation and SL 6.0 could possibly be available in 4 months but this assumes a Redhat 6.0 release next month. SL 4.9 will come if and only if Redhat release a 4.9. SL 5.6 may come in spring or summer 2011, again depending on Redhat. In the discussion, he proposed adding openAFS into SL6 and possibly CernVM FS to the EPEL repository[6]. Romain suggested some fast tracking of fixes for zero day critical security exploits and he and Troy will discuss how this could work because the SL team is reluctant to patch the kernel and risk getting out of step with the Redhat releases. Troy said he would consider adding XFS to SL 5.6 but he thought it might actually be included in Redhat 6; to be confirmed.

Miscellaneous

Digital Library – Tim Smith provided an update on Indico and Inspire. In answer to a survey, 50% of the scientists who replied quoted Spires as their main information repository. Inspire is a development merging the meta-data features of Spires at SLAC with the technology of Invenio from CERN, fed by various information sources. The official launch was 2 weeks ago. It offers a variety of search interfaces such as the SPIRES original or a more modern Google format and Tim showed a number of uses of Inspire. After the first release, work is going on to add personal

[6] Extra Packages for Enterprise Linux, a volunteer-based community effort from the Fedora project to create a repository of useful add-on packages

libraries, personal alerts, inclusion of non-text material and more. Indico grew out of CDS Agenda and went live at the CHEP conference in Interlaken in 2004. Recent changes included making it timezone-aware (contributed by Fermilab), adding e-payment (sponsored and paid for by EPFL) and room booking; it can be used to request recording of lectures and webcasting and chat rooms have been added. Drag and drop are eagerly awaited. At CERN it hosts some 115K events (including 1000 planned already for 2011 and nearly 100 each in 2012 and 2013) and there are 100 known instances across the world. And finally the official 1.0 release is only due next year!

The CERN Search Engine – status report by Tim Bell on CERN's enterprise search tool which comprises not simply document retrieval but also the concept of document privacy, user authentication and authorisation. The enterprise search market exploded over many years but subsequently contracted as the smaller firms disappeared and major players merged or were taken over. CERN's search requirements cover web data, admin data and central IT data. After an evaluation the FAST search engine was chosen and has been in production since 2007. Tim explained how the search database is constructed from the documents and their access policies and how the search includes your authorised access roles. The database includes over 3M documents comprised of 1.5M web sites, 1M CDS documents, 61K Twiki pages. 470K conferences and events (two-thirds Indico). Ranking of the results has been an issue, largely due to poor use of keywords. What about Google, for example the Black Box solution? For CERN's number of documents, Google may be more powerful but it is also much more expensive and a test at BNL showed that although the ranking was better, the indexing of protected pages did not work. Future work on FAST includes adding more protected content, supporting Drupal and moving to a new version of FAST which has some interesting features.

Next Meetings

- 2nd to 6th May, 2011 in GSI, Darmstadt
- The Fall 2011 meeting, the 20th anniversary of the creation of HEPiX, is likely to be held in Triumf but official confirmation and exact dates are still open. An attempt will be made to invite the originators of HEPiX plus several former long-term members.

Alan Silverman
5 Nov 2010