HEPiX Fall 2011 CERN-IT-Note-2011-009

Triumf, Vancouver

October 24th to 28th, 2011

Introduction

In fact although the meeting was hosted by TRIUMF for the third time, on this occasion it was held downtown in the Simon Fraser University as major work is commencing on the conference facilities at TRIUMF. And it was somewhat special since it was the occasion of the 20th anniversary of the creation of the group and several special sessions were included in the programme. Just under 100 people attended the meeting, including at least half a dozen who used to be regular attendees but have since moved to other activities and in some cases new careers (or retirement). Two of the four "founding fathers" of HEPiX were among the attendees, one of them your author. The meeting was very well organised: the meeting room was the correct size and well equipped, the breaks were long enough to permit the very valuable face-to-face interactions which are so important at HEPiX and the social event was most enjoyable. To partially offset the expense of being offsite, the meeting organisers had attracted a number of commercial sponsors but they did not intrude into the HEPiX programme.

As usual, this report is the responsibility of the author and any errors, misinterpretations or serious omissions are his responsibility also. The overheads are all online at http://indico.cern.ch/conferenceTimeTable.py?confId=138424#20111024.

After the chairman's (Steve MacDonald) introductory remarks the meeting was opened by Mike Vetterli, professor at SFU and an ATLAS project leader, who welcomed HEPiX to SFU. He explained the physical structure of SFU, spread across 3 campuses in and around Vancouver. Since it was established in the 1960s SFU has grown to over 30,000 students. It is involved in 2 large and some smaller projects at TRIUMF and Mike briefly described these 2 projects, one on nuclear astrophysics and one on condensed matter studies. He ended by describing the ATLAS physics activities in Canada.

Highlights

As usual I try to identify some common trends and important points brought up during the meeting. Individual readers may find other subjects more interesting.

- If there was one word used more than others, at least during the site reports, it was "cooling". Apart from BNL, whose new computer centre now has massive under-used cooling capacity, sites from Paris to Chicago to Beijing reported problems with cooling over the summer and more recently and many of the these incidents have created significant service disruption.

- ITIL was mentioned several times also and several sites are now introducing Service Now (PDSF at NERSC and Fermilab).
- Virtualisation is a common thread and everyone, it seems, is evaluating OpenStack.
- More and more sites are moving away, or considering moving away, from SUN/Oracle products, both hardware and software. And a number of Sun Grid Engine users seem to be finding a safe haven in the arms of Univa.
- The departure of Troy Dawson from Fermilab's SL team caused some concern on SL's future but Connie Sieh reassured the meeting that the support team is being maintained, even boosted. [Troy himself was also present wearing his new (Red)hat – but also his traditional Hawaiian shirts.]
- This meeting had significantly more network talks than before, including several on IPv6, and they generated very healthy question and answer sessions.
- Once again a strong CERN contributions, many talks, of good quality and well presented in general.
- John Gordon has a new competitor in the "asking the most questions" stakes, Owen Synge.
- In this, its twentieth year, HEPiX is alive and well. We hold 6-monthly meetings attracting typically around 100 people on both sides of the Atlantic and there are three very active and productive working groups (on storage, IPv6 and Virtualisation) plus one which is currently quiescent but is prepared to be re-awoken when the SPEC organisation releases a new version of CPU benchmarks, expected next year.
- Next meetings – Prague from 23rd to 27th April 2012 and Beijing in autumn 2012.

Site Reports

TRIUMF Tier 1 Centre: the support crew is one group leader and 5 experts in various fields who all share the 24/7 cover. There are also 3 user support experts. The site consists of 1210 cores in 232 blade worker nodes, 2.1PB of disc, 5.5PB of tape space and 97 servers performing the usual range of services. Site availability is rather good currently. Still running some 22 Solaris servers. Recently introduced 4 XEN virtualisation servers with plans for more but in 2012 they expect to move from SL5 XEN to SL6 KVM hosts. Testing CVMFS since this summer and there is a local CVMFS Squid server. A major issue for them is the current need to drain the batch queue when performing CVMFS updates. At a higher level, too-frequent failures of IBM blade nodes have been linked to these nodes running at a higher temperature than the others by some 10 degrees. A new large purchase procedure is about to be launched for 4PB of usable disc space, 3.9PB of tape space and 3500 cores. Turning to the TRIUMF network, the speaker noted that the limit had been reached on traditional telephony capacity and VOIP was being introduced in all new buildings and the core network is being upgraded to 10Gps.

PDSF@NERSC: NERSC runs 2 Crays each with thousands of cores as well as many other similarly-sized Linux-based clusters and, in a corner, PDSF[1] has some 1500 compute cores and 1PB of globally-accessible storage. They are a Tier 1 for STAR at BNL, a tier 2 for ALICE in the US and a tier 3 for ATLAS. They offer fair share scheduling and a home-grown virtualisation service. In the area of procurement, ALICE, ATLAS and other LHC experiments benefit from special prices from Dell. Servers are operated beyond the standard 3 year warranty but usually fail in their fourth year. NERSC were considering a site licence for torque/MOAB but in the end PDSF joined in a site licence and support for Sun Grid Engine from UNIVA. They are adopting Service Now, initially only for trouble tickets. Plans for a new computer centre, noted in a previous site report, are meeting opposition from local inhabitants.

Fermilab: of course physics analysis continues on the Tevatron experiments after shutdown with a commitment of at least 5 years of analysis support and at least 10 years access to data. The 300 FTE Computing Division has become the Computing Sector which has 2 divisions, the Core Computing Division headed by Jon Bakken and the Scientific Computing Division whose head is not yet known. The speaker described the target areas of each division. Fermilab has expended considerable effort in being green and has received awards for its efforts. The speaker reported on several cooling incidents over the summer and steps being taken to reduce the frequency of these. There have also been a number of power incidents and again there is work in progress to avoid these as far as possible. After testing LT05 and T10KC tape technology, they have chosen the latter for future acquisitions. CMS has introduced CERN's EOS storage for user data files with considerable success. Service Now went live last week for ITIL support. Migration to Exchange 2010 is progressing slowly and rollout of Windows 7 is starting in earnest. The speaker ended with recent and planned acquisitions. They are working with KISTI on cloud computing and another collaboration with Open Nebula.

[1] Parallel Distributed Systems Facility

NDGF: it is being reorganised and will in future be hosted by Nordforsk rather then become its own legal entity, the original plan. It will also be renamed NeIC and will have a new director soon. It has lost many staff during this period of uncertainty. The speaker described some upgrades to some of the distributed sites of NDGF, both in capacity and in compute power. They run Ubuntu as well as Scientific Linux.

RAL/STFC: STFC has a new head, John Wormsley, who was previously at Fermilab. Recruitment is underway for 5 new posts at the Tier 1. There has been the usual acquisition cycle and this included the decommissioning of the 2005 and 2006 acquisitions. The speaker reported on issues with various components (Jware and Adaptec controllers and T10KC tape drives) and the fixes applied. CASTOR is the disc and storage manager, comprising some 18M user files; a major upgrade is planned during the 2012 Jan to March technical stop of the LHC. Despite more tests, they still suffer occasional sporadic network packet losses and asymmetric data transfer rates, causes of which have not yet been identified. They are starting to rollout a Hyper-V virtualisation local storage service (see later).

INFN: INFN is requested to change its statute and a new manager was appointed, a former member of L3 (Fernando Ferroni). INFN is working on plans for SuperB, a two ring electron/position accelerator with some 100 times larger luminosity than Belle or Babar. SuperB is an international collaboration spanning the globe from Canada to Russia and China as well as some EU partners. The MoU is under negotiation among the partners and a Technical Design Report is planned for 2012. A possible layout at Tor Vergata, near Frascati, was shown. INFN also participates in AMS. Turning to technical matters, INFN is the implementing agency for the Italian Grid Initiative, part of EGI.

CERN: Helge presented the CERN report. He noted the ongoing smooth operation of WLCG with peaks of 200TB of data being written to tape per day, a large part of which is repack data. A working group has been established to review how we run our services including the role of virtualisation, clouds and fabric management. The Call for Tender for a remote T0 Centre has been issued and should be adjudicated next Spring. Cold air in the main computer room has been raised gradually from 14 to 21 degrees, producing significant energy savings. Future acquisitions will use a new warranty scheme based on the supply of spare parts by the vendors and use of a CERN contractor to effect repairs. EVO is being replaced by Vidyo; SPIRES was switched off as INSPIRE moves to production. CS group is about to rollout a new generation of routers which will support 100GbE. Plans for IPv6 progress (see later). On the database side, preparations are ongoing for Oracle 11g. Limits on mail sending have successfully blocked spam attacks being launched from hacked CERN nodes. PES group has moved CVMFS Stratum 1 into production and they are established JIRA and BOINC services. The effect of IBM's acquisition of Platform, authors of LSF, are not yet understood.

GSI: they are concentrating on preparing for computing for FAIR, including testing new highly efficient cooling techniques, in particular for the so-called 6 level Green Cube. As a prototype they have established a 2000 core test setup and ATLAS and ALICE are running jobs to test this. LSF is being replaced by Grid Engine and CVMFS is being used for software deployment. 10,000 AMD Bulldozer (see later) cores will arrive in Nov to let them setup a first Minicube. Meanwhile they remain very satisfied with their Lustre configuration and performance and further expansion is planned.

ASGC: the speaker described some recent purchases for networks, servers and storage. They are migrating their CASTOR service to 10GbE but it is taking some time. They have eliminated their UPS to save the 30% overhead that this takes and looking at applying space technology to improve heat conduction of the data centre to increase its thermal efficiency. Plans are being made for an ASGC cloud for e-Sciences across Asia; it will be based on Open Nebula, vNod*e and OpenStack; it will use CERNVM and CVMFS is already deployed for ATLAS.

Prague Tier 2: This was a review of the FZU centre supporting ATLAS, ALICE, D0 and STAR, as well as solid state and astrophysics. The speaker demonstrated how the centre had grown over the years to a 2011 capacity of 30K HEP-SPECs with more planned. But they are now approaching the limits of installed cooling capacity and water cooling is being installed for new acquisitions. They run SL and were using PBS for batch but the latter was considered too expensive with poor support and they are moving to torque/maui.

BNL: after a 6 year campaign, finally BNL's cooling capacity is more than sufficient for installed CPU capacity. There will be a week shutdown of HPSS in Dec for a major software upgrade moving to a new version of HPSS and from AIX to Linux hosts. There has been the usual round of new acquisitions, especially in storage. They continue to suffer some quality issues with Dell. They are trying to build an NFS 4.1 service but both BlueArc and Netapp have been unable or unwilling to help. As part of their Unix Centralisation Project mandated by the BNL funding agencies, they have chosen puppet for configuration management. and RACF for security compliance and migration to these, for both servers and desktops, is in progress. Some new smaller experiments have applied for support and often make demands which have already been satisfied in a different way for existing experiments and this creates additional and unwanted support load – and some frustration - on the BNL IT team. They are investigating cloud computing and they are evaluating a Cloud Edge server with AMD Interlagos (see later). As a precaution, they shut the centre in advance of Hurricane Irene and when they re-started, they lost some discs on the restart, mainly older models, but on the whole they escaped major disruption.

IRFU: the Saclay site of the Paris Tier 2 distributed node. 2011 has seen the extension of their liquid cooling capacity and more storage capacity and of course more worker nodes for their grid service to fulfil their pledge for 2012. They have migrated to Exchange 2010 but the service is no longer based at IRFU, now centralised at CEA where three AD domains have been merged.

LAL: compared to IRFU, the cooling situation at LAL, installed 4 years ago, is in a very bad shape and they get little help from the supplier. In fact since yesterday, much of LAL computing is closed. On the hardware side, their Tru64 cluster will be decommissioned by next summer and the SUN cluster should shortly suffer the same fate because it has become rather unreliable. On the other hand, a Netapp service is being put into production with 60TB. They are now installing an HP blade storage server with 200TB and a StratusLab private cloud for service virtualisation. They are about to migrate to a new domain within the IN2P3-wide Windows AD domain. GRIF offers some 30K HEP-SPEC and 2.2PB of disc to the LCG grid and is running smoothly, except for the LAL part. They are connected to the LHCONE prototype in production mode since some days.

DESY: Lots of construction sites – PETRA III, FLASH II extension, a new centre for laser sciences, one for structural systems biology, and last but not least the first physical appearance of buildings for XFEL, for which tunnelling is approaching completion and for which the first calls for tender for computer equipment have been issued. DESY will become a national astrophysics centre inside the Helmholz Association and hosts an IceCube tier 1 site for this. DESY’s dCache installation is now 4.7PB and supports a growing number of user groups. GridLab is publically available for tests, for example on NFS 4.1. There was a recent AFS and Kerberos conference at DESY and the slides are available at indico.desy.de. Their Lustre service remains very stable and they have recently added a new large Netapp 6280 installation of 1.2PB. For bulk file store they are investigating IBM SONAS, possibly with native GPFS. DESY hosts a National Analysis Facility and they purchased some Dell C-blades for this but Lustre does not offer a good access pattern for this application and SONAS may be chosen. For batch, the Hamburg site has been having problems with torque/maui and are looking at both Univa Grid Engine and S(on of)GE. Zeuthen are using UNIVA Grid Engine with a support contract and they are happy. A DESY Image Sharing service (DISH) is starting to share virtual images and there are plans for cloud services as from 2012. Migration to Office 2010 is complete and that to Windows 7 just starting with a target of 80% completion by end 2012.

SLAC: still looking for a new CIO and a search committee is considering names. Several reviews have taken place to identify strategic focus areas and new directions and the speaker showed the top 10 selected projects for computing. They are examining Globus Online to handle large volume data transactions into and out of SLAC and they are working with user groups on prototypes. The SLAC computing centre support group has taken a leading role to support the offline computing environment of the LCLS photon group and they are creating a Babar long term data access service based on Dell servers, NFS home directories and xrootd /HPSS for data access. They are evaluating BlueCoat WebFilter to protect SLAC’s computer network in order to meet Federal and DoE requirements. Windows 7 rollout continues; it is one-third completed. Mac support is implicit rather than declared so no defined support model exists. Stanford Uni has approved plans for a new Stanford Research Computing Facility for which building work should be started next year and completed in 2014. Despite being a long-term Sun customer for hardware, SLAC are moving away from the latest Oracle offerings, towards Supermicro for example. Still using HPSS heavily but looking at Lustre for the future and a test system is planned. An XLDB (extremely large databases) conference was held recently at SLAC, with almost 300 attendees and 50 turned away; 2 new user communities were identified and lots of interesting sessions were held. The speaker ended with a longish list of recent, indeed current, cooling problems which have affected SLAC computing, in some cases closing many systems.

IHEP Beijing: 1000 staff, 2/3 scientists and engineers. Blade servers for CPU capacity and a 1.7PB Lustre installation for file services. 8000 CPU cores in total, about to rise by 20%; running SL 4 and 5. Batch is based on torque/maui and they have tried to merge AFS into this to prolong the lifetime of AFS tokens. They are considering adding CASTOR with the HSM feature to their Lustre service. Two IBM 3584 tape libraries with 26 LT04 drives and 5800 slots. They have 10Gb links to Hong Kong and thence via Taiwan and directly to the US and also 2.5Gb links direct to Europe. The LCG Tier 2 site runs well, mainly for CMS and ATLAS. LCG storage is based on dCache with 320TB installed. A CVMFS client has been installed on all LCG worker nodes. And once again there are cooling problems having reached 75% of maximum capacity which leaves little room for future acquisitions, as well as generating local hot spots. Hence the call for tender for water cooled racks.

JLab: Jlab had a major cyber attack in June and went off the net to investigate and recover from. The attackers had been present but undetected for weeks so contamination was serious by the time they were discovered. The speaker listed the steps taken to recover, a process which was under the control of a sympathetic DoE team and which took several weeks in total and related changes are still ongoing as a result. One direct result was a major redesign of their Windows services, a new certificate authority and the clear separation of public and private web servers. Turning to less dramatic issues, a user survey took place and the main demands were support for PDAs, videoconferencing, calendaring and remote access. In IT Division, VoIP continues, with much success; they are installing Zimbra for calendaring but the recent take-over of the supplier resulted in a price increase which will limit future use. Both the Infiniband and GPU clusters for QCD are being upgraded but the percentage which should be spent on GPUs has not yet been decided. The 100 node experimental physics cluster has been upgraded to CentOS 5.3[2]. They added some 32 core AMD nodes for interactive use but there has been little use so far. They have added a further 6 servers, 68TB, to their 14 node , 24TB Lustre service. During the upgrade, some copy steps reported compare problems and they were recommended by Whamcloud to upgrade to Lustre 1.8 but this had no effect. Worse, the previously well-working Lustre nodes showed lower performance than before. They still do not understand either problem and they cannot afford formal supplier support[3]. In preparation for the major upgrade to Jlab's facilities in 2013 to 12Gev, the IT Division is planning a full external review in 2012 to be ready for data taking in 2015.

IT Infrastructure

CERN Facilities: Wayne presented the status and plans for the computer centre. Local hosting (Safehost) is working as planned with installed capacity increasing and approaching the physical limits of what we contracted for. There has been no need for system admin interventions at the remote site for 6 months. In the CC, using outside air in the cold months saves some 200kW and, adding in raising the operating temperature, the net savings total 4.7GWh per year. Civil engineering work progresses on the CC upgrade to 3.5 MW and we are on target for completion of this phase around the end of 2011. Purchases and contracts to equip the new rooms are in process of being placed. During the work, there were a number of incidents related to water leaks or dust in the main computer room. Regarding the establishment of a remote site, a call for tender has been issued with minimal specifications to permit flexibility by the bidders. Replies are due next month and we hope for adjudication at the March Finance Committee meeting. Studies have begun on creating an independent network hub outside the main centre to ensure business continuity.

DESKA, a tool for central admin of a grid site: this comes from FZU in Prague and allows to centralise many kinds of monitoring data from different sources ranging from warranty information to online performance data. The speaker listed some existing candidate tools but none coped with the range of data he wanted to cover. He showed the database schema he had devised to store his target data.

Scientific Linux: with the departure of Troy Dawson from Fermilab to Redhat, Connie Sieh returned to HEPiX to provide the traditional SL update. Linux Format, a UK publication, in an article on SL next month has the byline on the front page "if it's good enough for CERN, it's good enough for us"[4]. As usual, the usage graphs show continued increase in its usage although, thankfully, the use of V3 and now V4 is slowing. 6.1 was released in July and 5.7 in September. Around the time of Troy's departure, two new members have joined and more effort will go into automation of SL build and release procedures.

SINDES-2: Veronique Lefebure presented the status of SINDES, CERN's Secure Information Delivery System. The tool allows service managers to upload confidential information, for example password files, grid certificates or configuration files, to a central repository such that only authorised client machines may download the information. It has been in use at CERN and outside since 2005 and a new version was required to increase the flexibility of the tool and replace unmaintained code; it was re-implemented from scratch. Veronique described the types of file found in SINDES and the various privileges of the service managers and the client machines. She explained the

[2] There was the usual question, why not SL; the speaker replied there were philosophical reasons behind the choice

[3] Discussions with other Lustre users during the week gave the speaker some clues to investigate on her return to JLab. HEPiX to the rescue we hope.

[4] See http://www.linuxformat.com/files/lxf_covers/150-big.jpg although, as John Gordon whispered, it was not good enough for CERN since we made a local version.

differences between SINDES-1 and -2 and gave the timetable to complete the migration to the new version. Once it is considered stable it will be available, with documentation, on Sourceforge under the Apache2 licence. She noted finally that although it will be published on the Quattor web site, there are no Quattor dependencies.

Use of OCS for Software and Hardware Inventory: Matthias Schroeder presented this subject. There is no real knowledge about how many Linux nodes are being supported, on what hardware, what versions, etc. While registration has been needed to get initial access, there is little information provided and no check of continued use or not. OCS is an open source multi-platform inventory tool. To this CERN has added some specific data fields such as last boot, machine architecture and so on. On the other hand, some collector fields were dropped or are ignored. For desktops, it was added as a config tool dependency so it eventually arrives on all Linux nodes in CERN's public network. It was also installed on all the central Linux nodes. Matthias then displayed some screen snapshots. Although targeted at Linux it can be installed on Windows and a few people have done this. Only snapshots are kept but by feeding this information into another database, history can be maintained. There are currently some 10,000 clients since its initial use in Spring 2011. A Mac client is being prepared but here user cooperation will be required for installation. In summary, OCS has been easy to adopt, extend and install and CERN now has much more knowledge about its SLC population.

Configuration Management at GSI: GSI is considering moving away from cfengine; although they are quite happy with it, there are concerns about its scalability in face of the expected rapid growth of GSI computing ahead of FAIR. There is a newer version of cfengine but the speaker thought no other HEP site used it; he was informed that in fact Fermilab uses it. For GSI, the tool known as chef looked more interesting and the speaker described some of the features of chef. There is a chef server, an api, and rule books known as cookbooks. A pilot chef server was installed in a newly-installed batch cluster. Already they recognise some deficiencies in chef rather similar to those of cfengine but they nevertheless intend to push ahead with chef. Asked why they did not go with puppet, the speaker felt puppet was too similar to cfengine.

Computing and Benchmarking

AMD's new processor: an invited (and entirely technical) talk by John Cownie of AMD on their new 16 core Opteron processor[5] which will be formally announced shortly. He was not permitted to give too many details but he had measured its performance against HEP-SPEC06. The processor is based on 32nm technology which allows them to go to 16 cores. He presented how AMD had moved from single core to the latest chip and how the HEP-SPEC score had moved with the advances, both the single-core score and the overall processor score. For the 2010 Magny-Cours processor, scores are 9.3 per core, 400 for a 4 socket configuration at 2.2 GHz. He described some of the new features, for example to assist virtualisation. Although Intel and AMD cooperate and try to stay compatible with each other, some of the new features in this chip won't appear on Intel chips until later and software writers, or perhaps only compiler writers, must be aware of this and he showed a chart of which compilers already made use of them via compiler flags. He explained some of the power efficiency features as well as its clock boosting capacity if power is available, for example if some instruction execution units are not being used at a particular moment. Finally he showed the configuration used for the HEP-SPEC06 tests running SL 6.1 and gcc. He was not permitted to show the absolute HEP-SPEC numbers for Interlagos/Bulldozer but he could compare them to the Magny-Cours scores and how switching on the new features boosted the performance. As evidence of how the speaker succeeded in remaining 100% technical, there was a lively question and answer session covering specific processor features, overall system performance and AMD-Intel compatibility, or perhaps lack of.

CPU Benchmarking at GridKa: we had already heard about the new AMD chip and although Intel's new Sandy Bridge was expected soon, it has now been delayed until March; it will be 8 core. As usual Manfred Alef presented many detailed performance plots and the interested reader is referred to the overheads. Manfred reported that in reply to a recent tender, some vendors claimed SL6.1 was needed for Interlagos chips but under pressure they agreed that the benchmarks indeed worked under SL5.6 with updates but with unexpectedly poor performance and indeed Manfred confirmed this result with a borrowed processor from Dell. These were performed with 32 bit checks but moving to 64 bit showed (a) a performance boost and (b) consistent results between SL5 and 6. The reasons are not understood but it suggests that SL6 may be advisable to get the best performance. Turning to SPEC benchmarks themselves, Manfred noted that there is a new version of SPEC06 since September however they appear to have no influence on the HEP benchmarks so no need for a revision in the near term.

[5] The processor is code-named Interlagos and the core is known as Bulldozer.

HEPiX Benchmarking working group: there was then an open discussion on the future of the working group. Since its inception in 2006, it has produced a standard HEP benchmark which is used by many labs and accepted and indeed referred to by many box and chip vendors, as evidenced earlier by the AMD speaker, and later by the Dell speaker. Hot topics:-

- Should the benchmark move to 64 bit, is it time? It appears not difficult to do but would it cause confusion and should there be a conversion function for the32 bit version? Based on his tests, Manfred Alef thinks the changes to be expected are not worth the effort or the risk of confusion.
- The speaker confirmed that the new SPEC release seemed to have no effect on the benchmark.
- Should we produce a whole node benchmark since some experiments want to start using whole nodes sharing the memory among the multiple cores each running a separate job? There is a whole node working group in the LCG context so perhaps it is too early to come up with an answer to this but, at first glance, as for 32 to 64 bit question, the conversion factor from the current data appears to be linear so why create a new benchmark?
- Be prepared for a new set of SPEC tests, perhaps to be called V6 (rather than 2012).

Hardware Failures at CERN: Wayne Salter gave this talk on behalf of Olof Barring. CERN monitors SMART counters to fail a disc before they fail catastrophically. We have noted some 4,000 vendor tickets in the period 2010-1 and we note some 10,000 physical disc swaps in the same period. Out of a disc population of over 62,000 units this works out at some 5 failures per day, giving a MTTF[6] of 320,000 hours rather than the vendors' claims of 1.2M hours. Of course other components fail also and typically have to be replaced, each with its own lifecycle profile. After poor experience with most vendors (30% of failures are not repaired within the target repair time) we have decided to move to a model where the vendor supplies spare parts and CERN has contracted with its own supplier for the part replacement.

TSM Monitoring at CERN: presented by Giuseppe Lo Presti. TSM backs up AFS, DFS, mailboxes, and so on and so on. There is some 3.8PB of backup data, growing by some 1PB per year. Other significant numbers are 50TB of data backed up per day, 1200 nodes backed up and 1.5M files backed up. An in-house development, TSM Management Station, monitors performance and statistics. However this tool is not compatible with the next version of TSM and is not use-case based. In the new version, alarm generation will be offloaded to splunk, a commercial tool which is currently being evaluated in a free trial. The new management station will add admin tools for TSM admins to the traditional monitoring tasks and in the long term perhaps to automate some tasks and error repairs.

Dell Petasale Technology: a Dell speaker, Roger Goff, talked about a 10 Petaflops system being built in a Dell/Intel collaboration. The code name is Stampede, it is funded by NSF ($27.5M) and situated at the Texas Advanced Computing Centre. 20% of the power will be "normal" CPUs, 12,800 Intel Xeon E5 Sandy Bridge processors, and 80% co-processors. One advantage of Sandy Bridge over AMD/Interlagos is that it can perform 8 floating point operations per cycle (interlagos can do 2 if my memory is correct). Can HEP-SPEC codes[7] benefit from that? He described what he could about the Sandy Bridge processors and the new MIC (pronounced Mike) architecture for co-processors but he was limited by Intel in what he could say because neither is announced yet. He described how the co-processors and discs would be integrated and the different configurations which could be built from the basic components. In answer to a question, he showed some HEP-SPEC graphs prepared by Dell in collaboration with various HEP sites comparing Sandy Bridge to older processors.

Grids, Clouds and Virtualisation

Cloudman and VMIC Overview: presented by Belmiro Moreira on behalf of the CERN/IT, ASGC and BARC teams. The objective of Cloudman (CERN and BARC) is automated and centralised resource configuration via a graphical front end and pluggable back ends. It has the concepts of regions to describe the physical location of the resource and zones which describe a set of resources in a region. There is a so-called top level allocation to share resources from one or more zones to a group of users. There are also project allocations where a share of resources is allocated to a project and similarly group allocations. Work started on coding this month. VMIC (CERN and ASGC) is a Virtual Machine Image Catalogue and follows the virtual machine image catalogue model proposed by the HEPiX working

[6] Mean Time To Failure

[7] Once again, after the AMD speaker, the Dell speaker referred directly to "our" HEP-SPEC benchmark.

group led by Tony Cass. It should thus interact with other HEPiX sites to easily import and export image lists from/to other sites. After showing the architecture of the tool, Belmiro demonstrated some different image flows.

LxCloud, Status and Lessons: once again presented by Belmiro. LxCloud has been running since 2009 and has undergone various modifications since then, not only in usage (initially only batch mgmt.. now overall infrastructure) but also in internal configuration, moving for example from XEN to KVM with OpenNebula (ONE) and currently evaluating OpenStack. In a previous talk on this CERN was evaluating both ISF and OpenNebula and the result has been to select the latter, version 3 of which has shown good scalability (tested up to 16,000 VMs on 500 compute nodes) although there is some concern about LSF scalability now. On top of LxCloud are the virtualBatch production service and some early tests of Amazon EC2. The former consists of 48 nodes, 384 cores, 432 VMs and experience is good. The tests with EC2 are in a very early stage.

Hepix Virtualisation Working Group Status Update: given by Owen Synge. One area not yet addressed according to Owen is the understanding of different and differing cloud products; perhaps this is beyond the scope of the WG. He also feels there has been some misunderstanding about the WG outside the immediate HEPiX community; in other words there are still some trust issues. The HEPiX model is a publish and subscribe model but he admits that full interoperability of the various virtualisation services running at various HEP sites has not yet been fully tested or proved. After a few slides about meta-data security, he then jumped (a feature (?) of Owen's talks it seems) to what the working group has been doing over the past 6 months. This includes documentation updates, completion of image cache code, development of a test suite, some bug fixes and some new features. He then gave a long list of what remains to be done, for example contextualisation, clouds and fair share, further investigation of StratusLab.

OpenShift: presented by Troy Dawson who has left Fermilab's SL team and now works for Redhat[8]. OpenShift is a tool to help web application developers. Redhat refers to it as PaaS, Platform as a Service. The idea is that the developer takes care of the code and the Redhat tool takes care of the rest – build, certify, distribute, maintain, etc. Most of the presentation consisted of demonstrations of the so-called Express version and the Flex version. There is also a Power version. The Express web sites created are located in a cloud maintained by Redhat and it supports different web app tools, including Drupal, JBoss, etc. Flex is built over Amazon cloud (only, for the moment) and is intended for more permanent and/or high transaction web sites. Flex comes with a free trial period.

Virtualisation Working Group Report: presented by Tony Cass, a less technical presentation than that by Owen earlier. Tony considers the trusted image generation policy to be perhaps the most successful outcome of the WG. The policy document has completed the EGI consultation process and sent for approval to the EGI council. Turning to image exchange, there has been success in image exchange policies and CVMFS is now accepted as the standard software distribution tool. But there is less progress in delivering a distributed catalogue of endorsed images. More work is needed in this area and a face-to-face meeting is pencilled in for RAL in December to kick start this work back into life. Tony believes that StratusLab Marketplace could be an interesting alternative for a cross-site layer in this area although the remaining lifetime and future prospects for this EU-funded project are unclear. And then there is cernvm ….

Eucalyptus at NERSC: the speaker was assisting users to use this and not one of the authors. Eucalyptus was part of the Magellan project, a large (720 nodes, 5780 cores) cloud facility funded by ARRA[9] and shared by Argonne and NERSC. Eucalyptus was chosen over OpenStack and Nimbus as a cloud support tool to manage users, images and nodes as explained at a previous HEPiX and appeared the "best" choice at that time. NERSC used it with low numbers of image types and instances but Argonne has some scalability issues with larger numbers. A user engagement campaign was launched to attract users for testing and offering lots of help to get started. They were encouraged to install the client on a central system initially because firewall issues made debugging on personal nodes rather difficult. One of the first groups was STAR and ATLAS has built a highly scalable compute cluster on the Eucalyptus cloud. But both of these were serial workloads and the project mandate included investigations into parallel workloads but attempts to adapt some typical NERSC threaded workloads to Eucalyptus were not successful – too much performance degradation. The speaker's summary is that Eucalyptus was not quite ready for production when first installed although it is better now; there is lack of a scheduling tool; use of clouds requires sys admin help, users cannot be left to themselves. There will be a full report on Magellan, including a small part on Eucalyptus. As a postscript, Magellan has ended as a project and both Argonne and NERSC have re-allocated the hardware to other uses. Too bad for user communities which had devoted time and effort to use Magellan!

[8] And from his talk, he is clearly still having trouble making the switch from SL to Redhat Linux, but then it has only been a couple of months.

[9] American Recovery and Reinvestment Act

Virtualisation and Clouds at RAL: presented by Ian Collier. They have been working on services virtualisation for the RAL Tier 1 site. Examining some use cases, they have identified candidates for virtualisation but don't expect these to become cheaper, but hopefully more manageable. They have looked at VMware, Hyper-V and open source platforms and decided on Hyper-V for reasons of good experience elsewhere, management features and low cost. This is working (mostly) well on a handful of production service hosts although the team is Linux-based and they find interfacing to Windows rather a challenge. RAL is also creating a test e-Science cloud using the IaaS (Infrastructure as a Service) model. They are examining (who is not?) both OpenStack and StratusLab. The latter seems somewhat further along but this is still at a very early stage since work began only last month. Initial development is happening within the RAL Tier 1 but the eventual use base is much wider across the lab.

OpenStack: this was a presentation by a firm offering OpenStack support. OpenStack was founded by NASA and Rackspace to power cloud storage, compute nodes and networking. It is now a worldwide open source collaboration. He described the three major components – Nova for compute provision in the form of a network of virtual machines, Swift for data storage and Glance for image storage.

Network and Security

LHCONE: presentation given by Edoardo Martelli. Since the inception of the LHCOPN network for WLCG, the LHC computing model has changed to take account of the lower cost, reliability and higher bandwidth of networks. In this model, any tier 2 site can get its data from any tier 1 and this needs a faster, predictable and pervasive network between the tier 1 and 2s; Bandwidth should be at least 1Gb (minimal), preferably 5Gb or 10Gb. Hence LHCONE – LHC Open Network Environment. Of course LHCOPN remains for the tier 0 to tier 1 links. LHCONE consists of single node exchange points, continental or regional distributed exchange points and interconnecting circuits between these. All must share a common policy. Services include dedicated circuits, a shared VLAN or backbone and monitoring. Although the exact roles and responsibilities are not yet defined, all stakeholders are involved and LHCONE is a true community project, now in prototype.

perfSONAR: this was given by a speaker from Internet2. Ever-increasing network usage implies more and more reliable, higher bandwidth networks. But inevitably problems become more complex and we have to deal with flaws. Most sites already perform their own monitoring but more and more cross domain debugging is necessary. Internet2 encourages the use of existing tools where possible but they have added perfSOMAR-PS as a monitoring middleware which has been developed by Internet2 and US-ATLAS. It is available in the form of a toolkit which provides access to underlying tools and data sources and provides a rich set of APIs and GUIs. The speaker showed some examples of successful problem debugging. As well as deployment in US-ATLAS since 2007, it is now deployed in ATLAS Italy, on LHCOPN, in Canada and beyond. Most questions concerned the fact they base their release on CentOS and are unwilling to build and release – and support – an SL version, even though porting should be rather simple.

IPv6 at FZU: with only a small allocated address space (one /24 subnet), Prague had over-subscribed this three times and were suffering severe routing problems. They have setup IPv6 on a separate VLAN. They want fixed IPv6 addresses on each node. The speaker reported on experience with DHCPv6, autoconfiguration, network boot over IPv6 and said that they plan soon to join the HEPiX IPv6 working group.

IPv6 at CERN: presented again by Edoardo. He presented a few reasons why we need to move. He showed how the 128 bit address splits into site, subnet and host part and he introduced some new acronyms we will require to get used to, for example NDP (Network Discovery Protocol) replaces ARP. Multicast will replace Broadcast and an interface may have multiple addresses. There are various ways to move to IPv6,for example bridging or dual stack but since the first does not scale and has all the disadvantages of NAT, dual stack is generally agreed to be the preferred method, even though bridging may be good for a fast entry. The CERN service will be based on dual stack so every device will have both an IPv4 and an IPv6 address; we will use the same provisioning tools, operate the same security policies and offer the same network services. There will be one or more /64 subnet per physical subnet but IPv6 will only be deployed where network devices are able to run at IPv6 line rates. Addresses will be managed by DHCPv6. The LANDB schema is currently undergoing some major changes, in particular devices will have an "IPv6 ready" flag. CS group have prepared a deployment plan and everyone is concerned, from system managers to developers to operations managers. They are therefore creating a CERN IPv6 Forum with representatives from IT groups, the major experiments and all departments. Finally, they will make a testbed available.

HEPiX IPv6 Working Group: report by the convenor, Dave Kelsey. Since its creation, some 16 groups from Europe and the US and so far one experiment (CMS) have joined the WG and there are some 50 names on the mailing list. He noted that CERN and DESY are probably the most advanced sites in terms of plans and testbeds. The next face to face meeting is planned for December in CERN. They have created an IPv6 VO hosted by INFN and 5 sites have already connected to a HEPiX IPv6 testbed with more sites planning to join. Grid data transfer tests will start next month. If all goes well it is hoped CMS will start their tests in December and workload management tests will follow. EGEE created some interesting IPv6 code testing tools so these can be applied to EGI. In the US, DoE are proposing target dates for IPv6. Another WG activity is to perform a gap analysis, what is the scope of concern (applications, middleware, batch systems, etc), what does "IPv6 ready" really mean? Work on this should commence soon and a survey is being prepared.

Security Update: presented by Romain Wartel. Before starting his review, Romain replayed some old attacks from the 90s which, as he said, could have happened yesterday. Some things never change, except that today many hackers are more professional, motivated by profits. Security is a matter of trust; and yet in the past months two certificate authorities were compromised and one of them has gone bankrupt. Are there others? Romain showed how social networks have been used to trap browsers, often using a major news item as bait, for example recently the death of Steve Jobs was used to trap 26,000 out of 80,000 subscribers to a blog. He gave many scary examples where reputedly trustworthy organisations have been invaded electronically. IOS and Android mobile phones are a particular target. He then turned to recent attacks on HEP sites. There was an Identity Federation workshop at CERN in June and another is planned for Oxford in November.

Is Cyber Security IPv6 Ready: a most entertaining talk given by Bob Cowles. Bob ran the first part of his talk as a quiz, who knew what about IPV6. He then went the through the possible areas of security concerns. For example, since it is so easy to allocate blocks of IP addresses rather than individual addresses, address blocking will be more difficult. Bob expects all the IPv4 errors to be repeated which will allow "the younger members in the room" to "enjoy" the same experiences as he has lived through. The CentOS firewall can be considered IPv6 ready, it does not fail if you send it IPv6 packets; in fact it simply ignores them and passes them on! In Windows, you can turn off IPv6 but you can't turn it on again and you need to re-install! There are already some published security guidelines, see slides for the references.

Storage and File Systems

ATLAS Storage at TRIUMF Tier 1: they have some 2.1PB of usable disc space running under dCache as well as 5.5PB of tape storage. Their first large disc farm consisted of IBM discs and initially they had a 4.5% failure rate but this rocketed to 58% in the fourth year so now they monitor error logs to replace discs before failure. The power supply unit failure rate is also high (22%), blamed on high temperatures, but seems to stabilise after a burn in period in the first year. They are very happy with IBM support. With a more recent SUN disc purchase they see only a 1.6% failure rate but expect to see this rise with time. On the tape side, they bought some high density frames from IBM and suffered some initial problems with library inventory corruption. Although IBM resolved the issue with the first purchase, it reappeared with the next purchase and IBM had to intervene again.

EMI Data: Patrick Fuhrmann presented plans for the second year of EMI. After acknowledging the contributions of other EMI personnel, he started by summarising the main goals of EMI. He explained how the first release happened in time and the second release (Matterhorn) is on schedule. What happens after the end of the project, May 2013? The EU reviewers had urged more effort in this area and current plans are to ook at an Apache-like open source scheme rather than an EMI-2 project. Turning to EMI Data, the objectives are to improve existing infrastructures, integration and standardisation. Current work includes creating a better SE and catalogue synchronisation, for example to resolve dangling references (pointers to lost files). There will be a new release of FTS, a consolidated EMI-Data library and a WebDAV front-end for LFC/SEs. For Patrick, pNFS is a done deal; it is continually under test at DESY under dCache and first production use will commence soon for the photon science group. Redhat 6.2 will come with a pNFS-enabled kernel and SL will follow shortly after. There are still issues with X509 certification and wide area transfers but these are being worked on. In announcing the ongoing changes to DPM, Patrick paid a sincere and warm acknowledgement to Jean-Philippe Baud for his contributions over the years. On dCache he noted the change in release numbering, V1.9.14 becoming 2.0. Future plans to dCache include moving away from the assumption, indeed the necessity, of basing it on a mounted file system. They will add a data abstraction layer and this could interface to something like the Hadoop file system or an object store. He warned that these were only plans and may never see the light of day. Also, they are looking at a new three tier model, adding a layer of cached files, for example on expensive SSDs, in front of the discs to cache files required for random access by analysis jobs. First results of simulations of this model will be presented at ISGC and at CHEP. Asked why an interface to Storm was missing in the last EMI release, he referred obliquely to "issues" in the Storm team and hoped it would be included in the next EMI release.

HEPiX Storage Working Group: Andrei Maslennikov reported on the latest activities of the working group. In the last 3 months some more evaluation tests were performed, principally at FZK, and Andrei presented the running environment. He then presented graphical comparisons of different AFS versions, of different file systems running experimental code and so on and readers are invited to consult his overheads. They plan to continue the current tests, hopefully increasing the size of the test facility. They would like to evaluate OpenStack/Swift. They will re-issue their questionnaire next spring.

Migration from dCache to HDFS: from a US CMS Tier 2 site (Uni Wisconsin). Along with 6 other US sites they have performed this migration while preserving their resource commitment to CMS. First, why change at all? Wisconsin has just over 1PB of usable storage and run some 20K jobs per day, 60K CPU hours. For the past 8 years they have relied on dCache and have been very happy with it. In 2009-10, HDFS appeared, reputed to be highly fault tolerant and designed to run on commodity hardware. It was quickly validated by US-CMS and the first US tier 2 sites started migration. After some hesitation, Wisconsin felt that HDFS made better use of the "crummy" hardware which they relied on for disc storage. While this may seem a little vague, the speaker seemed to think that further explanation of the reasons for change were outside the scope of the talk which is a shame[10]. The rest of the talk was about the steps taken to smooth the migration and continue to offer service during the migration. They deployed an HDFS cluster, mounted HDFS on dCache pools via FUSE and mapped pNFS ids to file names, checksums and meta-data. They then scanned the directories, resolving duplicates and repeated this until all the data was migrated. Storage access for grid jobs was maintained via the dCache service and the only serious service disruption lasted one day during which writing to dCache was disabled to complete the cut-over. Six months later the HDFS service runs smoothly and they are happy with the open source model of support. Outside the US, there are a few early adopters of HDFS.

CASTOR and EOS: presented by Giuseppe Lo Presti. The overall strategy remains to maintain Castor for Tier 0 production activity and move xroot-based end-user analysis to EOS. The CASTOR service continues its rapid growth and moves about 1PB per month with peak writing speeds of 6GiB/s during the heavy ion run last year. A new in-house written Transfer Manager will replace LSF inside CASTOR; stress testing has shown a factor of 10 increase in performance and it is now being rolled for ATLAS, then CMS and then the other experiments. CASTOR tape performance is also improving, both read and write, the latter largely via the use of buffered tape marks over multiple files; this alone shows a speed-up factor of almost 3. This is still under final testing but will soon be available for wide deployment. Turning to EOS, this is disk-only storage of user and group data with in-memory name space. This gives very good file-open latency and focuses on end-user analysis with chaotic access patterns. It is based on the xroot server plug-in architecture and is fully complementary to CASTOR. The access protocol, xroot, is not the critical factor in the speed, rather it is the client caching and this is where the development effort is concentrated. It can run on single JBOD discs with software redundancy on cheap and unreliable discs; or on RAID systems. It has a tuneable quality of service via redundancy parameters; it is optimised for reduced latency; and is scalable. It is designed to be self-healing. Field tests have been completed and ATLAS and CMS use it in production since the summer and pools migration from CASTOR to EOS is ongoing; currently there is 2.3PB of usable CMS data and 2.0PB for ATLAS. Giuseppe ended with some performance charts and a few words about future plans.

CVMFS Status: given by Ian Collier of RAL/SFTC. The CVMFS client offers a /cvmfs/ filesystem area with files coming from a web-server. File accesses are intercepted by Fuse and file operations trigger downloads. There is lots of caching and the whole is transparent to the users and jobs. All data is now hosted on Netapps provided by CERN/IT, the so-called Stratum zero and there are a number of intermediary Stratum-1 servers to which clients connect. CERN/PH/SFT has used it to offer a software distribution service for read-only data such as a software repository and it has been widely adopted and is under constant improvement to respond to the demands of the experiments. There are tools for small groups to set up repositories for CVMFS but Ian says they are not trivial to use so far, contested by Ian Gable of Victoria and others.

[10] I was surprised that no one from the dCache probed this but in conversation with Patrick that evening, he believes some pressure was being applied on US sites to come off dCache rather than for technical reasons.

20th Anniversary

Banquet: the celebrations started on the Thursday evening at the banquet and included a slide show by Alan of photographs over the years from Corrie Kost, Sandy Philpot, Thomas Finnern, Tony Cass, John Gordon and others including Alan himself. There was a HEPiX quiz with Redhat T shirts as prizes, and the presentation of a plaque to HEPiX's oldest active member.

HEPiX From the Beginning: Alan Silverman presented the history of the group. He described how it was created, where (CHEP at Tsukuba), when (1991) and by whom (Les Cottrell, Judy Nicholls, Harry Renshall and Alan). He listed some early meetings and showed at how we have arrived at the present format – Spring meetings in Europe and Autumn/Fall meetings in North America with occasional excurions to Asia. He listed the various working groups of HEPiX with their successes. He ended with some questions such as "is HEPiX worth the money" and what directions we should take in the future.

Networking Retrospective: presentation by Les Cottrell. He first reminded the audience of the protocols, devices and network topologies of 1991. Cabling was coax for terminals, twisted pairs for phones and thicknet or thinnet for Ethernet. DECnet Phase V was being introduced because phase IV was running out of address space. ISDN had been standardised but not yet deployed. Internet had some 1M users and 600K hosts. First WWW servers started to replace Gopher and WAIS. In 1998, 75% of all Internet users were American; today they are less than 15%. He presented some of the initial design goals of the Internet and the challenges it faces, such as address space; mobility and the need to change IP addresses (need for persistence across links); spam, now 97% of email; lack of effective broadcast and multicast. He then showed how the technology has moved forward, the growth in usage and how that is skewed in some places such as Africa and parts of Asia. He noted how the Football World Cup in South Africa has led to the installation of multiple network links around Africa which has not only delivered Internet to more African countries but also reduced the price thanks to competition. Les ended with a live demo of this growth as a series of plots using a Google app.

HEPiiX Experiences: Talk by Thomas Finnern. Thomas admitted that he does not document so much but considers his HEPiX talks as part of his documentation. He claims to write self-documenting scripts. He has attended around 50% of the HEPiX meetings and he recounted some of his highlights, in particular with the early HEPiX and X11 scripts which were a joint DESY/CERN collaboration. After an absence of some years, Thomas re-commenced his HEPiX interactions in 2007 with work on virtualisation. He ended with a slide – HEPiX is helpful, HEPiX is fun and HEPiX is worthwhile to attend.

20 Years of AFS: presented by Rainer Tobbicke. AFS was always projected as a "planetary"[11] file system with all the features we know and love. It has global name space which is still an issue with file access. He described the initial installation of AFS at CERN and related some early experiences with this. He related how we had waited, in vain, for DCE/DFS to replace AFS, how there had been a review into an AFS replacement which had concluded there was no replacement. Currently, we see peaks of 2000-3000 accesses per second on popular volumes. Today AFS has found its niche: it is low volume (100TB) but with 4-6 billion accesses per day, low latency, largely automated management. It has clearly a sound design and has set a reference point and has resisted a serious of challenges. It is alive and well and will probably maintain its place as an infrastructure file system for some time yet.

Computer overview over 20 years: given by Corrie Kost. Corrie compared the cost of performing a million operations 20 years ago, today and what may be possible in 20 years. He noted how disc sizes are mushrooming. He compared time sharing systems of 10-15 years ago to clouds today. He listed some current trends, gathered for example from Gartner reports. Corrie ended with some projections in a 10 year span – massive growth in storage, more efficient input devices that QWERTY keyboards, Android devices dominating Apple devices and so on. He also predicted the end of tape technology and the slowing of Moore's Law. Beyond that, perhaps 20 years from now, AI computing may come to the fore, robotics – but will it be silicon-based or carbon-based.

Alan Silverman
28 Oct 2011

[11] Rainer, and others, consider the word global to have de-valued.