HEPiX Spring 2012 Workshop IT/NOTE/2012-007

Prague, Czech Republic, 23rd to 27th April, 2012 Version 1.1

Contents

Highlights 1

Introduction 2

Site Reports 2

IT Infrastructure 5

Storage 9

Computing and Benchmarking 11

Grids and Clouds 12

Business Continuity 13

Networking and Security 14

Vendor Presentation 16

Social Programme 16

The Spring 2012 edition of HEPiX was hosted for a second time by the Czech Institute of Physics, situated on the outskirts of Prague. The meeting took place in the (almost too) comfortable conference room in the Hevrovsky Institute of Physical Chemistry. The meeting was very well prepared and executed by Milos Lokajicek and his team. The hosts had worked hard on sponsorship which meant an additional cultural evening, free lunches and a free HEPiX T Shirt. Access from town was easy and the attendee pack included a free transport pass for the week.

As usual, these notes are personal and I take responsibility for any errors or omissions. The slides of all the talks are available at https://indico.cern.ch/conferenceTimeTable.py?confId=160737#20120423.

Highlights

- As usual in the site reports, news of several recent or impending changes in the top levels of management at various sites (SLAC, Fermilab, CC-IN2P3, et al). New last meeting and more so this time however were some announcements of retirements of long-standing HEPiX members – we are starting to show signs of an aging society! Plus at least 3 sites used the opportunity to advertise open posts.
- Several sites mentioned plans for, or at least consideration of, the replacement of NAGIOS.
- DESY are looking, like CERN, at who should have accounts and/or access to resources.
- DESY also looking at what is needed for Mac support.
- Several labs reported external reviews; one, Diamond, were able to use such a review to establish for the first time a defined annual budget able to support their plans.

- As usual many talks from CERN. Perhaps because of scheduling most talks in the first two days seemed to be almost only from CERN. Perhaps too much? Certainly too concentrated[1] and happily the proportion dropped in the second part of the week. Co-incidentally, or perhaps not, this was the first HEPiX meeting where some late offers of presentation had to be refused because the schedule had already been overstretched.
- The various working groups continue to be active with the exception of the Benchmarking group which is in hibernation awaiting the new version of the Specmark tests.
- Next meetings: IHEP, Beijing from Oct 15th to 19th; CNAF Bologna in Spring 2013; probably University of Michigan at Ann Arbor in Autumn 2013. And there are future offers from Zeuthen and the University of Manchester.
- With Michel Jouvin's elevation to LCG GDB chair, he is stepping down as European co-chair of HEPiX and after a hotly-contested election from among 4 strong candidates, Helge Meinhard was voted as the new European co-chair.

Introduction

The Director of the Institute, Prof. Jan Ridky, welcomed the 90+ attendees from 15 countries. He explained the relationship of the Institute to the Czech science community and some of the non-HEP activities which take place there, including astrophysics, optics and lasers. He was followed by Prof. Rupert Leitner, chairman of the Committee of Collaboration of the Czech Republic with CERN, who described the various HEP activities, not only in Prague but elsewhere in the Czech Republic. Most of the groups are working at or with CERN, but also at BNL, D0 at Fermilab, Belle in KEK, Auger and others.

Site Reports

Prague: the Institute of Physics is spread across several sites in the city. They have a large server room with about ~20 racks and several smaller remote rooms. Due to continuous changes, there are several generations of racks and hosts. He then listed the various clusters in operation, one of which is a WLCG Tier 2 site for ATLAS and ALICE and which awaits some 600TB of disc space delayed by the recent floods in Thailand.

Fermilab: organisational changes – Rob Roser appointed head of Scientific Computing Division, Lother Bauerdick elected OSG Executive Director and Ruth Pordes elected head of OSG Council. Vicky White remains Associate Director for Computing and the CIO. Email was recently migrated from several diverse schemes to Exchange 2010 and calendaring followed shortly after. They now have distributed redundant network core connectivity. Improvements to the cooling of the main computing facilities have been agreed and work should commence shortly. They have signed a managed services contract with Dell which includes periodic refresh of the desktop hardware. BlueArc continues to perform correctly and a major firmware update is planned for May. They are experimenting with clouds – see later.

NIKHEF: trouble with a data storage acquisition where iSCSI performed poorly under strain so they eventually moved to gluster on 4 Poweredge servers, 30TB per server. They moved from Cisco AP 1250B to Aruba devices for their latest wifi network expansion after a demonstration at a recent major conference in Amsterdam.

[1] An exception was the CERN-only segment where AGILE and its use in CERN was explained in a coherent series of 4 talks.

ManageEngine is used to distribute software to their Windows desktops although there are some problems updating the operating system itself.

NDGF: NDGF is the distributed Tier 1 site of the Nordic e-Infrastructure Collaboration (NeIC) which itself is hosted by Nordfosk. New systems, indeed new clusters, have been installed at the various sites.

Uni Michigan (AGLT2): they have a relatively large Tier 2 centre spread over 2 sites 70Km apart (UM/Ann Arbor which is the principle site, and MSU). In view of service continuity, a lot of work has been put into service node virtualisation, based on the commercial version of VMware, and they have invested in a degree of replication at the MSU site. They have done serious work in multi-service resiliency options. Storage includes dCache inter-site caching and relies somewhat on federated Xrootd usage. His slides are full of detail on the virtualisation challenges they have met and solutions adopted or under test.

CC-IN2P3: the current director will leave in December but his replacement is not yet known. The Oracle Grid Engine (GE) has fully taken over the batch service since January and CC-IN2P3 is leading an initiative to coordinate GE operations and management issues across sites (there was a BOF scheduled for one evening in this workshop – see report later). One activity for example will be to prepare a list of requests to Oracle for future directions of the product. They are training 8 staff to start work on business continuity issues and they are preparing a new trouble ticketing system based on OTRS, chosen after evaluation of 3 products by a small team.

PIC: an LHC Tier 1 site (and also Tier 2 and Tier 3 site) for LHC run by a team of only 10 staff and including 24x7 support. The current computer centre is full and they are experimenting with more efficient cooling schemes. They have moved to Nagios for monitoring; server configuration is managed with puppet. They use Oracle VM (based on the original Sun product) and 90% of the test servers are virtualised as well as more than 50% of production systems. The most recent purchase of CPU servers was won by Dell with C6100 dual-twin systems. The next step will be to implement a single-instance PBS supporting more than 12K jobs. They use FNAL Enstore for tape access and dCache for disc.

GSI: new permanent head of IT Dept is Karin Miers. The mini-cube project (see previous reports) has won an award as project of the month in a national competition. Like Fermilab, they have migrated mail to Exchange 2010 and also created a new hierarchical architecture for their Windows domain. Their mini-Cube continues to grow, now at 2 levels each with 48 racks. A backbone for FAIR has been established, working at 1TB/s.

INFN Tier 1 site: 120 racks and several tape libraries in a 1000 sq. m room. They currently share a 20Gb link between LHCONE and LHCOPN but they will move to separate links at the end of the year. Latest CPU tender was won by Supermicro with its new 2U twin square system. But they are concerned in future that they need to move beyond 1 job per core and would like to discuss with other sites with similar challenges. Worker nodes for production are being deployed for ALICE and Auger. Storage is based on an amalgam of GPFS, TSM and StoRM, known as GEMSS.

CERN: a short talk because many subjects are covered in later talks. Openlab Phase IV has started with all the previous partners plus a new contributor. The Thailand floods are affecting also CERN's planned storage upgrades. Initial experience with the internal repair service to replace vendor warranty service is favourable. The Vidyo service is running well, once the conference has been set up correctly and the users gain experience with it. On networking, almost all of the Force 10 routers have now been replaced by Brocade. The networking group is studying virtualised networks and IPv6 tests continue. The migration to Oracle 11g is progressing well and databases are being consolidated on NAS hardware. Single sign-on is being extended to AIS applications. Service levels are being harmonisation between AFS and DFS. The newly re-launched IT web site is now based on Drupal. The CVI service has been extended by bringing in-house the large virtualised park of BE Dept services and now comprises 2500 virtual machines. X-win32 has replaced Exceed for Windows X emulation and the Microsoft

Campus licence now covers servers. Smartcard tests continue for strong authentication and tests are also being performed on facebook and google lightweight accounts. VoIP Softphones are enhancing the standard phone system and they are looking into Unified Message on Exchange to deliver unanswered calls as e-mail.

UK South Grid: 8 sites centred on RAL as the Tier 1 site. The speaker described briefly all the sites, how they are connected with plans for future upgrades.

ASGC: no new CPU procurements but problems with the addition of 300TB of storage. Moving to cloud-based virtualisation and they hope to support some 1184 VMs on a total of 132 blade servers. They have adapted the virtualisation model proposed by the HEPiX working group.

SLAC: Randy has been appointed interim CIO but they have hi-jacked the NASA CIO to take on this position permanently. Amber Boehlein has taken over Scientific Computing along with Scientific Computing Systems. Bob Cowles is retiring from Computing Security head. At a higher level, a search committee has begun looking for a replacement for the retiring SLAC Director, Persis Drell, and there will be a new COO in due course. Long-term data access for BaBar is now in production. The clusters for various post-HEP projects are being gradually built up and the Fermi Gamma-ray Space Telescope has been extended by a year. They are looking at developing a document management system and they will partner with Stanford on a searchable photo storage database. A Data Centre expansion project has been approved and work has just started with occupancy scheduled for 2014. HPSS is alive and well and they have just started tests on Lustre.

RAL: new Director of e-Science Dept will be announced after Neil Geddes's transfer to the Technical Dept. And John Gordon has renounced being head of Division Head in preparation for his retirement later this year. At short notice RAL discovered that they have to replace some switch gear in the main power feeds and there are risks of extended power cuts until the work is complete. Also unexpectedly but more positively, they received a capital-only Government grant which allowed some infrastructure upgrade in their machine room. They discovered a vendor cheating on the benchmarks for a storage tender.

IHEP, Beijing: they have recently undergone a major network upgrade. They run dCache for CMS and DPM for ATLAS. Their CPU cluster is managed by PBS plus Maui and their file system is based on Lustre. They are concerned for their disc farm where 67% of the units are running with expired warranty and they searching for a smooth upgrade path.

Gridka: they have opened a new building this year, mainly for offices but with space to expand their computer centre later. Usual round of CPU and storage purchases in line with LCG and other commitments. They run several middleware packages for the different experiments supported. They have implemented a 24x7 support service with 2 teams, one for storage and one for grid services. They are migrating away from NAGIOS for performance reasons – the second, and not the last, site to report the replacement of NAGIOS. As the GGUS centre they have dealt with some 80,000 questions with the help of 1500 experts worldwide.

BNL: since last report, power and cooling have been stable despite major renovation of the physics building. They will soon implement Synapsense wireless temperature monitoring. RHIC is migrating to Shibboleth SSO with ATLAS being the last group to move. HPSS still in use here also. Important upgrades to their CPU and distributed storage farms. For configuration management, they are relying more and more on puppet plus Git plus Cobbler plus GLPI. Oracle was upgraded to 11g for the ATLAS conditions database and other DBs should follow later this year.

Diamond Light Source: their computing centre comprises some 500 servers and 500 workstations. They have now commissioned a second Lustre cluster with much better performance than the first one so they upgraded the metadata server on that one and it improved notably. They are looking into replacing NAGIOS and are testing

Zenoss. They underwent an external review and the panel judged that “IT was an enabler of good science” and that Diamond should have a “proper” annual budget.

DESY: they celebrated 20 years of the Zeuthen lab recently. DESY has joined HESS and Belle II. Tunnelling work has started on XFEL. In an echo of an earlier report from SLAC, the DESY IT group more and more finds itself offering consultancy and advice to small non-HEP experimental groups who arrive with less coherent ideas of how to apply IT. A recent major network upgrade went badly and took several weeks to stabilise. In a partnership with IBM, IBM Sonas has replaced Lustre in Hamburg with DESY having source-level access. After some scheduling issues with Maui, they are investigating Grid Engine for the Grid resources. Solaris is still alive in DESY although in decreasing numbers.

PSI: they are about to build a light source similar to XFEL, due to open in 2016. They have long relied on AFS for file storage and they plan to add AFS Object Storage in order to implement an HSM. They have also implemented an HPC cluster.

PDSF at NERSC: NERSC is integrating the Joint Genome Institute computational systems and infrastructure. They have completed the migration to Service Now. There have been a number of high level management changes in the PDSF team. STAR is deploying XROOT on their PDSF compute nodes. PDSF uses xCAT (eXtreme Cloud Administration Toolkit) for software management, including node discovery.

JLab: they have introduced a scheme to encourage all staff to take charge of Accelerator Operations, under expert supervision, for 2 weeks at a time, and thus to encourage cross-domain idea sharing. Their newest purchases introduce the first Sandy Bridge systems to the lab. The accelerator shuts next month for 18 months to upgrade it from 6 to 12 GEV and there will be an external review of computing this summer to ensure readiness for when the machine comes on line again next year.

IT Infrastructure

EasyBuild, Building Software With Ease: a scheme developed at the HPC team in the University of Ghent to automate and ease the build and installation of user software, including especially updates, for end-user programme authors. The tool should be able to cope with different compilers and libraries, even different O/Ss or hardware. It should produce reproducible and verified code and support multiple versions. It is written in Python and reads specification files created by the user to describe the software package. The speaker then presented some examples and worked through them. It has been in use for 3 years and the HPC team build all end-user packages with it. It has recently been released as open source and it currently supports some 250 end-user packages.

Database on Demand: a CERN talk by Daniel Gomez. Allows a CERN user to create a database instance on a virtual server and grants the user full DBA privileges – and responsibilities - on that instance. Today it works for MySQL and should soon also support Oracle databases. He showed some screenshots of the different actions such as creation, back-up, restore, deletion, etc. and he explained how each is handled and what options are offered. He ended by showing the internal architecture and implementation and some admin tools.

Database Access Management: another CERN talk, this time by Giacomo Tenaglia. They needed a tool to track accesses to database and middleware servers on a “flat” network of several hundred nodes in small clusters. He showed the internals of the tool (DAM) and how it works. Current usage extends to 500 servers and 2000 accounts.

Infrastructure Services: third CERN talk in a row, this one by Nils Hoimyr. He covered the status of Twiki (124K topics, 2M monthly accesses); version control services, principally SVN with a legacy CVS service with 14 remaining projects; an issue tracking service based on JIRA which is just getting started; BOINC.

Experience with Service Management at CERN: after a coffee break, Patricia Mendez Lorenzo delivered the first of her talks. She presented the fundamental building blocks of the Service Management programme at CERN, namely the Service Catalogue, the Service Desk, the Knowledge Database, the Service Portal and the various ITIL-based processes such as Incident Management and Request Fulfilment. She explained why GS and IT have got together to implement this in a gradual and pragmatic way but in a way which can eventually be expanded to other services. She gave some details on the first two processes implemented and how they are being improved. The ticketing system in use is Service Now but this is now showing serious limitations in its reporting structure and further studies are being researched to improve this aspect. After about a year of operation, half of this in production mode, a review was conducted by an external consultant. His report was presented to the service owners in March and future project plans have been adapted in light of this but in fact some aspects were already judged as ahead of schedule in comparison with other implementers of such processes. Among the more important changes coming is the implementation of Service Level Management.

Hardware Acceptance Test Suite: Eric Bonfillou presented the new HATS burn-in tests. It is currently in prototype and planned to be used to certify newly-purchased hardware, or re-certify hardware after some major hardware change. It is designed to solve two drawbacks of the previous scheme – no remote console, power control or monitoring; and the previous scheme ran in the confined software environment of a live O/S image which sometimes masked complex hardware errors. HATS runs on a dedicated server and it transfer files to and from the test systems via SSH; sets of tests are wrapped in bash scripts; these are run in sandboxes on the fully-configured target systems and each set of tests generates its own logs. It runs on any Linux environment. In addition to testing hardware, it can update BIOSes and firmware and it can execute performance tests. Eric showed the first results and explained how they plan to move to production mode over the summer. They plan to productise it for use elsewhere after receiving requests from IN2P3 and RAL[2].

ARTEMIS (Almost Real Time Environment Monitoring and Information System): developed to help monitor the A/C environment of the new RAL computing room. After some evaluation of open source solutions, they found none which fitted their needs, for example support of multiple vendor monitoring systems. The chosen solution supports temperature, humidity and airflow sensors. Java scripts run on clients and report back to a server. The almost real-time aspect allowed them to vary the cold aisle temperature and they could sit and watch the effect. They estimate to have decreased their PUE score by 0.3 units by increasing the cold aisle temperature from 15ºC to 21º. Future plans include moving to 3D sensor placement, to fully integrate heat map views and to support more of the infrastructure such as the chillers, power meters, etc. He ended with a plea for help since he is alone on this project.

AGILE: the introduction to CERN's AGILE infrastructure was presented by Helge Meinhard. He gave the rationale why CERN had developed some of its existing tools, usually the non-existence of needed tools judged able to handle the scale of the challenge. But does CERN still have the resources to continue with home-made tools when other comparable tools have appeared on the scene? For reasons such as the explosion in the introduction of virtualised services and the imminent introduction of a remote Tier 0 centre, CERN also needs to optimise more processes and introduce more automation where possible. CERN wishes to move to ?aaS where the ? could be infrastructure, platform, application, etc. It needs to make more use of standard tools packaged into toolkits such as one for monitoring, one for provisioning, one for control, etc. From this came CERN's AGILE Infrastructure

[2] And the HEPiX audience

Project. They started with the configuration management area (full machine lifecycle) and virtualisation/cloud deployment but eventually it should cover all aspects of running and managing CERN's Computing Centre(s).

Configuration and Operations Tools: presented by Helge on behalf of Manuel Guijarro. Currently these tasks are handled mainly by Quattor but CERN has no resources to develop this tool any further and they feel that the community support is too small and unfunded. So how to automate this and better integrate the sub-components, e.g. into an IPv6 environment? The most commonly-used comparable open source tool is the puppet toolkit and puppet and chef should be the core of any Quattor replacement. The first challenge is the node count which puppet can handle, especially when one considers the explosion of virtual nodes. The target is to handle some 300K virtual nodes. A first try in September 2011 had mixed results but efforts are continuing. They are currently running OpenStack for cloud software for virtual machines, puppet as the configuration management tool and Foreman as the dashboard. None of the tools are perfect but does CERN always need the best tool or can it survive with tools which satisfy the needs[3]. More recently they have added yum for software distribution and Git for template management. Helge then explained briefly how puppet works. With these tools, CERN is moving towards PaaS, and more automation. Helge's team is now looking for early adopters from among other CERN/IT support groups. In the questions, Michel Jouvin could not resist pointing out that part of CERN's current problem with Quattor was that it never participated in the community Quattor support effort.

AGILE Infrastructure Monitoring: presented by Pedro Andrade. CERN has currently some 30 monitoring applications, 40K data producers and generates some 280GB of monitoring data per day. These range from data coming from the hardware to data from the running jobs but there was little overlap or sharing of data. The first step in unifying this must be to aggregate all the monitoring data in a single store and define a single format for that data. As far as the monitoring technology is concerned, it must be easy to switch tools in the future so the intention is to create a chain of tools with well-defined solutions at each layer. These constraints and examining a number of use cases, led them to create an overall architecture which they believe can handle the challenges such as changing individual tools as monitoring technology is developed and which is scalable to CERN's needs. A key tool in this architecture is the message broker which transports all monitoring data via messaging. The first implementation chosen for the message broker is Apollo and tests have started with this. Lemon has been chosen as the monitoring tool of choice, Hadoop for data storage and Splunk for development of operating tools, dashboards and APIs. Of course any and all of these tools can be replaced as and when more appropriate tools appear.

IaaS: presented by Jose Castro Leon. As Helge said, the target is to support up to 300K virtual machines and deliver Infrastructure as a Service. OpenStack was chosen to deliver the operating platform or toolkit to orchestrate the cloud. Within the OpenStack architecture, they use

- Nova for instance and volume management; within Nova, KVM is the chosen Hypervisor; Nova has 2 APIs, for OpenStack and for Amazon EC2
- Glance for image management
- Keystone as a cloud identity service for authentication, authorisation and as a service catalogue; it uses role-based access control and it integrates with LDAP and CERN's Active Directory
- Horizon as the graphical user interface; this has Shibboleth authentication which interfaces to CERN's SSO.

Different applications consume different IaaS resources and have different QoS requirements so a scheduler is required along with resource management and Pedro showed the scheduling architecture to handle this.

[3] Quote from the speaker

Scientific Linux Update: given by Connie Sieh. Usual graphs showing the progression of the usage base, number of mirror sites, etc. SL 6.2 was released in February and 5.8 was released yesterday; OpenAFS 1.6.1 is in testing for both. SL 4.9, indeed SL 4, end of life was February. Current lifetime of SL 5 is now 2017 and for SL 6 is 2020, although there will probably be no new versions of 5 (Redhat or SL) after 2012 so how it will handle new chips after this is an open question. After the departure of Troy Dawson last autumn, the FNAL team was boosted by the addition of 4 people, some dedicated to projects. They will create a Packages database for use for Quality Control.

Computing Infrastructure Upgrade: presented by Wayne Salter. He gave an update on the work on Building 513 since the last HEPiX meeting. Overall commissioning of the room is planned for the autumn with first machine installation in November. He described the installation of the UPS systems including the small fire which occurred. He then moved to the exercise to contract a remote computer centre and explained the procedure which had been followed including the technical specifications and the service levels demanded. These assumed a gradual expansion in capacity at the host site over the coming years. He then summarised the replies to the invitation to tender and how they were analysed. A proposal was made to CERN's Finance Committee and this was approved at its March meeting and a contract is currently being established with the Wigner Data Centre in Budapest, formally KFKI, and an existing Tier 2 site. The equipment will be housed in a new computer centre under construction and Wayne described the new facility. One small disappointment is the rather higher-than-expected PUE[4] value of 1.5. Meetings are about to start to define detailed working procedures and first hardware installation is scheduled for 2013. The goal is to have a production Iaas service operational by Q1 2014.

Computer Room Experience of a Tier 2 Site: the site in question is Oxford. He described some cooling and fan speed problems he had met in his local computer centre just outside Oxford and the solutions which had been applied. They are currently looking at enclosing their cold aisle retro-actively in order to improve temperature control.

Procurement Trends at CERN: presented by Eric Bonfillou. CERN/IT has a central procurement team which conducts (almost) all tenders for IT systems for the computer centre, from writing the specifications to executing the evaluations of test systems and preparing installations with the Operations team. They have noticed that increasing disc sizes have revolutionised disc servers and hardware RAID has been replaced by software solutions. Studies have been initiated to identify use cases where SSDs make financial and technical sense. As previously reported, CERN is turning away from vendor warranty support since it is so variable in quality and a contract has been established which defines a local CERN repair service based on part replacement from a stock supplied at purchase time by the vendor.

Monitoring at GRIF: the architecture must take account of the fact that GRIF nodes exist at multiple sites, each with its own rules on firewalls, the use or not of ssh and, especially, the wide variety of node types. There is a total of 1000 nodes and there are 16 sys admins. The speaker listed the various packages in use. Pakiti is used for Security. Nagios is a central part of the monitoring architecture and while they acknowledge its power, it has its limits, for example slow restart after a crash unless a particular patch is applied, and the number of active checks is limited by the fork capabilities of the server. Monitoring is an ongoing process, there is always more which should or could be monitored but there is only so much monitoring data which can be understood by a fixed number of sys admins.

Quattor Update: talk given by Ian Collier of RAL. He started with some history, including why Quattor was attractive to some sites, namely the pan language, modular architecture and pre-deployment verification, all of which remain true today. Aquilon is the third generation of configuration database of Quattor. The main change is

[4] The ratio of total facility power to power used for the IT equipment itself.

the introduction of a broker daemon which has overall ownership of the system and with which all users must interact. If I understood Ian correctly, Aquilon was developed by Morgan Stanley and the Quattor working group are working to allow the QWG templates to work with the Aquilon schema and they are almost there. RAL has Quattor thoroughly embedded across their Tier 1 and they have developed a nice dashboard which they will make available in sourceforge. Without directly referring to the CERN talk which previously explained why CERN was moving away from Quattor, Ian noted that there is an active Quattor community which is, as he put it, "not large but very active" and he described some of its current activities. Despite his reticence to make this reference, a question was asked from the audience: "given CERN's tendency, why stay with Quattor", to which Ian repeated that the original reasons for adopting it remain true. Ian estimates that there is about 1 FTE at RAL working on Quattor support, spread across multiple staff. Michel Jouvin joined the debate, noting that Quattor was attractive to small sites largely because of the availability of sharable configuration templates, and in fact a similar sharing could be done with puppet. Michel further noted that there is currently little demand for additional features and so the support load is rather limited, being mainly maintenance; for example when Morgan Stanley adopted it, they put in 8 to 10 FTEs to implement it for them but have since scaled this back to 1-2 FTE maintaining it along with their templates.

Storage

Data Storage Strategy at CERN: delivered by Kuba Moscicki. Another replacement speaker or, as Kuba put it, Presentation as a Service. At CERN DSS group deals with many data sources from physics data to file backup. Today's offering includes -

- AFS for user file services and possibly large user workspace for end-user analysis data; low latency, general purpose file system
- CASTOR for long term bulk storage
- EOS, low latency storage for data analysis, entirely disc based
- TSM, the backup system

The group also tracks hardware technology, moving more towards JBOD storage and SSD caches. They are working with the new openlab member Huawei to investigate storage for cloud systems. HSM appears to be reducing in importance and the group distinguishes between long-term storage of physics data (CASTOR) and on-disc storage of active analysis data (EOS). For the future they will introduce m+n block replication which should allow them to define or adjust the QoS parameters of file containers (directories). They expect all the current systems to still be there in 2015 but much improved. They are ready for cloud storage but there is no pressure for it as of now.

CASTOR Tape Services: first of two talks by Alex Iribarren. He described recent service improvements such as the introduction of writing buffered tape marks. The next step is to improve the writing of meta data to tape. The hardware itself is kept as up-to-date as possible within budget constraints. In this direction they are testing LTO drives from SpectraLogic, perhaps for use with less often-referenced data. They perform continual tape verification, reading back and repacking old data.

Backup Infrastructure at CERN: Alex continued with a description of the service of which he currently the service manager. He listed the impressive variety of files which are backed up. The increase in stored data is continual although they have been stricter on what is backed up and growth has been reduced from 60% annually to 38% last year but there is still a daily traffic of 57TB per day. The tape setup is quite complex and the fault tolerant architecture actually creates more problems than it solves. It is also expensive and hard to maintain, requiring

special hardware uniquely for CERN. For AFS they have two head nodes cross-linked to two TSM servers and a recent upgrade to new servers delivered a notable performance boost, simpler maintenance and a simpler procurement procedure because they are now using standard CERN/IT-wide systems.

Lustre and Infiniband at GSI: this was the first appearance of Infiniband at GSI. Installation and configuration was made easy by the use of Chef. Even booting can be done over Infiniband but there is an independent Ethernet connection for IPMI modules for emergency booting. They have two independent heating measurement systems. They continue to be very happy with their Lustre clusters and will continue to expand them.

Building a Czech National Storage Facility: the speaker works for CESNET whose mandate extends to providing data storage facilities. The aim is to build 3 distinct centres offering some 15-20PB of data for a variety of science and research organisations. It is co-funded by EU structural funds and the Czech Government. It should offer long-term storage with high redundancy using geographical replicas. It should also offer backup and archive for institutes but also for individuals. This implies a very broad usage pattern. The tender for the first site was won by SGI and the site is now installed with 3.8PB. As a Grid Storage Element, they offer access via dCache and they would like to add xrootd for ALICE but they have concerns on how to interface it efficiently to tape. They have also not yet decided how to implement inter-site replication: file-based, block-based or an FTP backend for HSM? Finally, they are also concerned if each site tenders separately in case of different winner vendors being incompatible.

What Comes After RAID: an INFN talk on data protection technologies. “RAID 5 is dead” claimed the speaker because increasing disc sizes makes rebuild times impossibly long. A 2PB file system would stand a 50% chance of being lost in a period of 5 years. RAID 6 remains feasible now for up to 10PB file systems but the long-term trend is similar. How about “erasure coding”: data is divided into m fragments and recoded into n fragments where n > m; then objects stored in n fragments can be rebuilt from any m fragments. Historically this algorithm has been too CPU-heavy to be useful but with multi-core processors this is no longer true. On the hardware side he proposed RAIN – a redundant array of Independent nodes. On this, RAID would be implemented across the nodes rather than across the disc arrays. A first implementation of these is Permabit and a second is from a new startup – StreamScale. And DataDirect Networks has introduced a system known as Web Object Scaler. Kuba pointed out that in fact EOS tries to implement erasure coding.

Evolving AFS at CERN: a second talk by Kuba. Today it covers 32,000 home directories on 55 file servers and 900 discs. 300M files are added annually and they just crossed the 1B file count. There are 10,000 CERN clients and 5,000 off-site clients. 300M read/write operations every day. Users ask for 1GB home directories, faster response, quotas, etc. Partially in response to this, CERN plans to move to SAS-based storage units which should be more reliable and offer better performance via the use of SSD caches. It would also permit the AFS team to benefit from the standard procurement methods used elsewhere in CERN/IT and described the previous day by Eric. These new storage units offer also better pricing. On the other hand, the introduction of them was not easy although they are settling down now as the team gains more experience with them. On the plus side, users now get 10GB home directories, 100GB workspaces, and SSD read caching. The major remaining issue for 2012 is the occasional shortage of threads; when the limit is hit, it is immediately visible to the users and CERN has implemented throttling methods to counteract this. First attempts to simply increase the thread limit failed but perhaps newer versions will support this.

Dynamic Federations: this was a dCache/DPM demonstration by Patrick Fuhrmann who acknowledged that Fabrizio Furano and others had done most of the work, partially under the EMI banner. Currently data largely resides on islands of storage but strict locality has risks, for example a single missing file could be fatal, so failover to another island would be advisable and storage clusters should be seen as a global and seamless file base. The various requirements lead to an http-based solution, in particular HTTP/DAV. The plan is to federate data

repositories that are compatible in name space, permissions and content; there will be transparent and dynamic discovery of meta-data. Xrootd federations are offered also. He then ran through a prepared demonstration with a federator in DESY and DPM in Taipei accessed from Firefox running in Prague.

NetApp Experience at CERN: given by Giacomo Tenaglia on behalf of Eric Grancher and Ruben Gaspar Aparicio. He started with a slide showing the growth of Oracle from its introduction to CERN in 1982, "when I was young" he said[5]. He showed the current network and filer topologies. Comparing NetApp with SUN-based systems he showed how many fewer incidents were reported on the former in an 18 month period 2010-1. He described some of the major features of NetApp and how they are used. Referring back to the "RAID is dead" talk several days before, he noted that CERN's Oracle database totals less than 10PB and that that they really benefit from RAID 6. The snapshot facility is mainly used for backups because a full restore from tapes could take around 58 hours, if all goes well. In the future, taking advantage of the so-called Ontap cluster mode could offer the possibility to perform non-disruptive version upgrades. A standard NFS 4.1 client is coming in a future version but this is not yet natively supported by Oracle. In summary, they are very happy with NetApp, its reliability and flexibility and the fact that it scales to CERN's needs.

Computing and Benchmarking

CPU Benchmarking at GridKa: Manfred Alef described the newest in the AMD and Intel chip lines. On these he has executed the HEPSpec06 benchmarks and he presented the results in some detail – see slides. He also measured the power efficiency where he notices an improvement over the previous generation of chips. He also noticed that running the benchmarks under Scientific Linux 6 (SL6) gave better performance than under SL 5 by some 5-15%.

HEPSpec06 on the latest CPU Generation: Michele Michelotto then presented a broadly similar benchmark run on the latest systems installed in INFN Padua. Once again the interested reader is referred to the slides, accessible from the conference web site. It was noted that the results were not affected by the latest update of the Specmarks, as discussed in the Vancouver meeting last year.

Hardware Evaluation 2012: a talk on a similar subject from a local speaker. They found that their new Sandy Bridge chip systems showed a 50% better performance than the previous generation. They also evaluated the claims for data de-duplication made by the vendors and found much lower factors in reality compared to the claims, which simply re-emphasises the need to perform benchmarks using real data. They also tried to answer the question of how many discs are needed for a 64 core worker analysis node. Their tests showed that 3-4 drives were usually sufficient and that adding more than 8 contributed nothing more.

Migrating to Oracle Grid Engine at CC-IN2P3: a small team was created to execute this migration and they made a list of milestones and of technical problems to be solved, for each of which someone was made responsible. Test clusters were created and guinea pig users nominated. He then listed some of these points such as AFS token renewal, disc space limits per job, etc. From the go decision in June 2009, the migration was completed in December 2011 and BQS was decommissioned. They have used Oracle support since the summer of 2011, with mixed results – good reactivity but a lack of efficiency – and the speaker listed some positive and negative examples. In summary, the batch support team at CC-IN2P3 has been reduced from 3 to 1.5 FTE and they now have a single farm for all types of jobs but they note a loss of job information, uneven spawning of jobs and a general lack of service stability. CC-IN2P3 has decided to create a user group to focus on GE management and operational problems (there are already fora for technical discussions) and the first face-to-face meeting was held at HEPiX the previous day. Nine sites were represented, some running GE, some considering running it and some simply wishing to know more about it. The discussions centred on –

[5] Hell, even I was young when VXORAC was installed but it's all relative.

- Which version to use
- What are the alternatives (!)
- How to benefit from the experience of all sites running GE
- How to contribute to the development

Further meetings will follow, trying to enlarge the group beyond HEP and they intend create a wish list to present to Oracle.

Upgrading Grid Engine at DESY: Sun GE (they still call it thus) is only one flavour of GE at DESY, there is also the UNIVA version running at Zeuthen as well as Torque/Maui. The open (to all DESY users) batch resource at Hamburg runs under SGE as does as the National Analysis Facility, also at the Hamburg site. SGE offers fair share scheduling and GPU support. When faced with a required upgrade, DESY Hamburg considered moving to a commercial version but could not see any justification for moving away from an open source solution. They eventually chose to move to Son of Grid Engine (SoGE) and are currently working to make it more robust. They also need to improve GPU support and graphical access to SoGE, and work on scaling issues. Meanwhile it will be used for the open batch resource (BIRD) and maybe with NAF and they will evaluate its use with the Grid.

Grids and Clouds

Virtualisation Working Group Report: presented by the WG convenor, Tony Cass. There has been considerable work in several areas, for example there is now an agreed proposal for EC2-compatable contextualisation policy. This also works with OpenStack and OpenNebula. The agreed method of passing machine information has been implemented for LSF at CERN, for PBS at NIKHEF and there are discussions with IN2P3 for a GE implementation. Work on image exchange is less advanced but a recent face-to-face meeting at CERN agreed on the next steps. In summary, the WG has made good progress in establishing policies for VM image exchange, and also on delivering a distributed catalogue of endorsed images and work continues in the remaining areas of the group's mandate.

Working Group Image Transfer: presented by Owen Singe. The scope of Owen's area of interest is to support images running on multiple systems and make it easier to manage images. He went through (again) the image endorsement method including the meta data security features. In the past 6 months, they have added 3 new optional fields in the meta data and, more importantly, changed the image format. He has adapted dCache to create DISH (DESY Image Sharing) to publish the images and the VMs. He currently supports quattor and Puppet+Kickstart images. He went through the steps to subscribe to an image from an image list.

Virtualisation and Cloud Computing at RAL: they use the Hyper-V platform in their Tier 1 but progress in making it high-availability has been slower than planned due to technical issues with the test hardware and the delay in procuring production hardware. Meanwhile some production but non-critical services have been moved to this platform and this has been smooth so far. They have ongoing issues with Windows admin (as previously reported) mainly because the team are Linux-based. RAL has built a prototype cloud for the e-Science Department but after a successful beginning, work has stalled waiting for new staffing. It is based on StratusLab. The SCT group in RAL is currently deploying the JASMIN super data cluster which uses virtualisation to offer a computing service for the climate and earth system modelling community.

Multiple Linux Environments at NERSC: PDSF must support multiple workloads and they created CHOS (CHroot OS) to provide the different environments. It was originally written by Shane Canon (now at Oak Ridge) in 2004 and was presented to HEPiX around that time. NERSC looked at alternatives and virtualisation was an obvious candidate, such as KVM. Linux containers such as OpenVZ or LXC) was another option. Both have severe administration overheads, either in the number of VMs to administer or the number of containers. So why not

simply use CHroot and its packaged tool, CHOS. NERSC estimates that CHOS deals with most of the use cases for virtualisation in HPC in a simpler manner. He gave some examples and the benefits for users and sys admins. He then went "under the hood" and showed how CHOS works.

Fermilab Cloud Update: a lot of the work described was performed in collaboration with KISTI, sometimes written by KISTI staff working at Fermilab. The original mission of Fermilab Cloud, production IaaS in support of the scientific community[6], has been extended. The service has added X.509 authentication, monitoring and accounting and the deployment of a distributed SAN. They are investigating automatic provisioning mechanisms. They have calculated an economic model for the cloud and offer different levels of guaranteed deployment of a VM with a price for each level, 24x7, 9x5 or opportunistic. Their model also compares very competitively with standard Amazon EC2 costs. Full detail in the slides. By re-using grid X.509 authentication, VMs can be launched with the user X.509 proxy and can thus make secure access via openSSL to an external repository. Using experience gained from Fermigrid, they believe the Fermicloud is largely fault tolerance. To replace the interim monitoring in place, they are going to decide between Nagios and Zabbix and there is a joint study with KISTI on this. There is also a study into virtualised storage as a service and a summary of the first results, taken from a recent ASGC conference[7] were presented. By design, excess Fermicloud capacity can be applied to Fermigrid and Fermicloud can work in hybrid mode with Amazon EC2 clouds.

EGI Federated Clouds Task Force: the primary objectives are to integrate virtualised resources within EGI's production infrastructure, to provide feedback to technology providers, identify early adopters and make recommendations on issues that need to be addressed. They should produce a blueprint for resource providers and users, they should install a testbed and they should promote engagement of resource centres. 23 institutes participate (CERN is not present) in this 18 month study, split into three phases. After the first phase, a testbed now exists and was demonstrated recently. A first blueprint document is also now available.

Helix Nebula: presented by Tony Cass on behalf of Ian Bird. Initiated by ESA for European space scientists, now expanded to the EIROforum science communities – the aim is to provide a cloud computing infrastructure for European science. They should define a light-weight governance structure and identify a short and medium-term funding scheme. There should be a pilot phase in 2012-4 and a full-scale cloud service by 2014. The consortium includes resource providers and demand-side companies and organisations. A long list of topics has been identified for study during the pilot phase, including some legal aspects. The first 3 use cases have been identified, including a Monte Carlo production task for ATLAS, with a target for initial proof of concept by summer of 2012. ATLAS jobs show similar success/fail rates to running on the grid; wall clock times are about double but does this matter when one considers the cost of provisioning. An EU-funded project has been proposed to tie this work to business.

Business Continuity

CERN Business Continuity Overview: this is a new HEPiX track and the first talk was presented by Wayne Salter who started with a definition of the subject. He listed the kinds of incidents which affect business continuity in increasing levels of impact, including all maintenance activities. It is vital to understand all the dependencies of individual services. One has also to consider what level of business continuity is required, from instant failover from restart from cold in a defined time. Wayne then listed various ways to perform disaster recovery including multiple data repositories, redundant hardware, good security, etc. He listed some of the measures adopted by

[6] It is being used not only by Fermi-based teams; i n usage plots shown, there were blocks of time used by ATLAS and ILC, both via OSG.

[7] Keith Chadwick promised to link the ASGC talk to this session in Indico.

CERN, the most important elements of which are redundancy in the form of RAID, multiple power supplies, three computer rooms, UPS, etc. and he gave some examples – batch, databases, Active Directory, mail servers and DFS. CERN has assessed the remaining risks such as a fire in the main computer room or the network fibre room or an extended power outage and Wayne described what steps were planned to alleviate these, including the suggestion to implement a remote networking hub. He also emphasised the vital importance of testing all solutions.

Business Continuity at Fermilab: while many labs, including Fermilab, are moving towards ITIL, much of the work described in this talk pre-dates this trend at Fermilab. They have multiple power sources and have successfully switched between them as needed. They have multiple computer rooms although these are not directly used to offer redundant systems, more independent systems (but see later). They have a detailed and documented load shedding plan in the event of power problems in the grid computing room. Network access is triangular in tandem with Starlight in Chicago and Argonne. Fermilab and Argonne have some level of backup of each other's most vital services in the financial sector. Fermilab also has a plan for almost complete disengagement from the Internet, for example in the event of a major security attack; a minimum level of internet presence is defined. More recently, using virtualisation and the FermiCloud described above they have built high-availability between the two main computer rooms and this is occasionally tested.

System Perspective of Business Continuity at CERN: presented by Patricia Mendez Lorenzo. Two ITIL processes are involved, Service Continuity Management and Availability Management. She explained the purpose of risk management where the first steps are to identify the elements of risk. A risk equation must be built depending on the assets (in some cases their cost), their vulnerability and the threat. She showed the formal method which the team expects to use to assess the risk and propose counter-measures and she illustrated this with example of the Service Catalogue.

Change Process at RAL: users demand continuous service but service providers require some update time and communications between the partners is vital, even within teams in case of dependencies. Starting in 2009, RAL's Tier 1 support team pushed for a more production culture in the team than previously. The philosophy is to be risk-aware not risk-averse. A change form was introduced at the risk of increased burocracy. Changes should be approved by consensus and authority should be delegated to the correct level of authority. They created a flow diagram and the whole process seemed to improve quite quickly but there was a lack of consensus on risk levels and uneven appreciation of possible impacts of changes. ITIL measures were introduced to improve the workflow and a more recent review appeared to show increased coherence among team members and they all see the benefits of the process. The ideas are beginning to percolate outside the Tier 1 team. RAL now believe that they have achieved the initial goals although gradual improvements will always be needed, for example based on feedback. Risk management has long been part of their project management method and they will next try to formalise impact assessment when considering changes. Their philosophy is to try to detect problems early rather than invest in disaster management, although they do have escalation plans for handling disasters.

Networking and Security

IPv6 at Fermilab: US Government mandate demands that public-facing services must be IPv6-capable by end-Sep 2012 and their internal client services by the end of FY2014. Fermilab believe that the 2012 targets are achievable with modest effort, the 2014 goals are not so clear. They created a small working group who meet regularly. Although scientific computing appears to be outside the scope of these demands, the Grid and Cloud Computing (GCC) Department expects the users to demand it, and probably at short notice so they have established a testbed for IPv6. At the current time, the external DNS servers are now IPv6-capable; the central web service has

IPv6 enabled; the e-mail gateway is contracted (to an outside supplier) to be IPv6 in the next 60 days, and there is a test system for testing from outside the lab. The network team is investigating suitable options for an IPv6 DNS service. For the longer term, they are now looking into the creation of a framework for 2014 deliverables and the GCC Dept has its own list of services to check.

HEPiX IPv6 Working Group: report by its convenor, Dave Kelsey. June 6th 2012 has been officially declared World IPv6 Launch Day, from when ISPs, web companies and others should turn on IPv6 services and leave them on. Participation in the WG has grown, including from among the LHC experiments. The testbed will be described in the next talk. Meetings are held regularly. An "asset" survey is underway to determine the state of IPv6 readiness of the various commercial and home-grown software packages, including applications, middleware, etc. Already some problems have been identified. Best practices are far from clear although some trends are emerging, for example large sites appear to prefer non-automat address allocation. Security is a major area of concern due to the vast scope of the standard and its newness. So there is a very long way to go and the WG will continue its work in the short and medium term, at least.

IPv6 Testbed Experience: report by Francesco Prelz. Seven HEP labs are participating in the testbed with at least one node each. All are running at least one Gridftp server. He listed the four prime questions he thought need to be asked.

1. Does the service break/slow down when used with IPv4 on a dual-stack host with IPv6 enabled?
2. Will the service try using (connecting/binding to) an IPv6 address, when available from DNS?
3. Will the service prefer IPv6 addresses from DNS, when preferred at the host level?
   Does this need to be configured? How?
4. Can the service be persuaded to fall back on IPv4 if needed?

CMS performed some file transfer tests simply to test the reliability. When the firewall was moved from software to hardware, a low level of failure appeared and this is still not understood. They also performed tests with FTS and Oracle 11. Oracle appears to work over IPv6 although it demands separate ports for v4 and v6 listening. They have functional FTS IPv6 transfer agents but the Globus FTP client is always IPv4 by default and a GGUS ticket has been opened for this package to create a run-time option to enable IPv6. Finally Uberftp, a popular gridftp client tool, does not yet support IPv6 and this is an issue. On the other hand, FTS services on IPv4 do not break on dual stack hosts. Among the more serious remaining issues is the fact that commercial software suppliers impose their own timelines; and the presence of functional IPv6 support does not always imply that IPv6 transport can be switched on by default. Work continues …

Federated Identity Management for HEP: report by Dave Kelsey. Some IM federations already exist, for example Grid X.509 certificates for WLCG are managed by the International Grid Trust Federation. Several national and supra-national education authorities have federations. A research collaboration on this was created in mid-2011 and HEP was one of the sciences represented. Some workshops have taken place and more are planned; a requirements document has been written and a first use case has been defined. Recommendations have been made to the research communities, to the technology providers and to the funding agencies and a common vision has been hammered out. Some federated IMs already exist within HEP and could be quickly adopted and the WLCG Security working group (TEG) is going to study this, trust being the main issue, not technology.

Computer Security Update: a talk given by Remi Mollon who had flown in especially for this. He started with the story of the recent DoS attack based on PHP hash table collisions where the first fix was much worse than the bug. There have been several vulnerabilities exposed in Linux but Macs did not escape free either. Although Macs are less vulnerable than PCs, some attacks prove that they are not immune. Remi described some of the activities of Anonymous and its various sub-groups.

Cybersecurity Retrospective and Future Directions: this may be the last HEPiX visit by Bob Cowles so he took the opportunity to look back and make some predictions for the future. He started by relating his HEPiX appearances and some of the scariest stories from those days, including the furore he created when he published some of the clear-text passwords he had hi-jacked during the week's sessions (without the usernames of course). Coming up to date, he noted that little has changed except there are now many more devices to be hacked, such as mobile phones, hand-held devices and so on. Mobile devices have the added complication that they are often not owned by an enterprise and so are much more difficult to update and control. Then come PLCs, which were never designed to be on the network but they now all are and they are vulnerable as we have seen. Plus, as explained by Bob at the previous HEPiX, the introduction of IPv6 is a whole new ball game. Bob is sceptical that we can continue as now, we need to re-think IT services to avoid being drowned in attacks. We should -

1. perform an inventory or business impact assessment
2. create cyber policies
3. follow best-practices for IPv6 architectures, when they are established
4. increase the use of virtualisation and short-lived service nodes, perhaps create a private cloud service
5. accept a variety of authorisation methods
6. create better and continuous security training

Embedded computers are everywhere so the chances of mis-configuration are everywhere also. Hacktivism is bound to increase, along with increasing denial of service attacks and credential leakage. Organisations need to be more pro-active to spot attacks, for example by making more use of log scanners.

Vendor Presentation

Why Near-Line SAS: a presentation by Western Digital, one the meeting's sponsors. He started by explaining how WD had been badly hit by the Thai floods but they are now back to about 67% of normal volume and expect to be at 100% by the summer. The effective acquisition of Hitachi Global Storage Technologies is agreed but the merge cannot be made "visible" to customers for 3 years because of regulatory bodies-inflicted restrictions. He then moved into his topic first starting with a definition of SAS, Serial-Attached SCSI. He compared it with SATA, Serial ATA and he believes eventually near-line SAS will "rule the world". Largely this is due to the decreasing mark-up of SAS compared to SATA. Turning to solid state discs, he expects "enterprise" SSDs to appear; they will be expensive but very fast in relation to hard drives. He emphasised that his talk concentrated on near-line drives and not at all on desktop systems. Pressed on hybrid drives (SSD + hard drives), he seemed to say that they were watching the market but were not going to produce a competitor in this space in the short term.

Social Programme

There was a very extensive social programme this year, largely due to the level of sponsorship achieved. There was a welcome drink on the Monday evening and lunches were free. On the Wednesday evening, we were entertained by one of the Czech Republic's most popular folk music groups who played a variety of music, some modern and some old, even some ancient, pieces. This was followed by what the organisers described as "light refreshment" but which was copious enough for most delegates. On the Thursday evening, a bus took delegates to Prague Castle from where guides led the party through the Castle grounds and down across the Charles Bridge to the restaurant for a most excellent conference banquet.

Alan Silverman
7 May 2012