Helge Meinhard / CERN
V2.0
30 October 2015

HEPiX Fall 2015 at Brookhaven National Lab

After 2004, the lab, located on Long Island in the State of New York, U.S.A., was host to a HEPiX workshop again. Access to the site was considerably easier for the registered participants than 11 years ago. The meeting took place in a very nice and comfortable seminar room well adapted to the size and style of meeting such as HEPiX. It was equipped with advanced (sometimes too advanced for the session chairs to master!) AV equipment and power sockets at each seat. Wireless networking worked flawlessly and with good bandwidth. The welcome reception on Monday at Wading River at the Long Island sound and the workshop dinner on Wednesday at the ocean coast in Patchogue showed more of the beauty of the rather natural region around the lab. For those interested, the hosts offered tours of the BNL RACF data centre as well as of the STAR and PHENIX experiments at RHIC. The meeting ran very smoothly thanks to an efficient and experienced team of local organisers headed by Tony Wong, who as North-American HEPiX co-chair also co-ordinated the workshop programme.

Monday 12 October 2015

Welcome (Michael Ernst / BNL)

On behalf of the lab, Michael welcomed the participants, expressing his gratitude to the audience to have accepted BNL's invitation. He emphasised the importance of computing for high-energy and nuclear physics. He then introduced the lab focusing on physics, chemistry, biology, material science etc. The total head count of BNL-paid people is close to 3'000. A new light source, replacing an earlier machine, is in the approval process, which is progressing as expected for a start in 2017. RHIC is a flagship facility for nuclear physics for studying the transition from a nucleon superfluid into a nucleon-hadron gas. The programme is in its 15th year now, still producing a lot of valuable science, with important questions ahead for the two experiments. BNL is involved in all three frontiers of HEP – energy with ATLAS, cosmics with DES, BOSS and LSST, and intensity with DayaBay, Minos, uBooNE etc. Concerning computing, the facilities for RHIC and ATLAS are managed together, and are built on top of OSG. He then mentioned the requirements for the reconstruction of the DUNE liquid-argon TPC, which requires a new paradigm as it requires high performance and high throughput at a time. HPC resources play an ever increasing role, including for ATLAS, for which HPC provides some 6% of the ATLAS Grid resources. Provisioning is being reviewed now, as the dedicated facilities play an ever decreasing relative role for the computing of the experiments, hence very flexible provisioning is becoming very important. This implies a change of the software the experiments are using for distributing jobs, which needs to become much more fine-grained; ATLAS have made good progress with this, as shown by the successful exploitation of the AWS spot resources. The latter requires a good integration of AWS into the networks used by HEP, in particular ESnet. The future is clearly combining the potential of grid, cloud and HPC. Michael then turned to the potential of the new light source NSLS-II, which is 10'000 times more powerful than the previous one (NSLS). Data rates are expected to reach 90 TB per day, which requires new analysis paradigms and very comprehensive simulations, the latter scaling to 1...100 million cores. BNL has accepted to become a leader in data-driven discoveries, the basics of which he explained.

Replying to a question from the audience, Michael clarified that AWS will not replace the facilities at RHIC, but rather absorb peak demands. The experience with AWS spot resources is good, the availability rather high provided the user is willing to accept a large number of different configurations. A broker is used to decide which offer to take up at any given time; the broker is designed such that it can also manage the dedicated resources and facilities at universities and HPC centres. The equivalent for storage does not exist yet.

Site reports

CERN (Arne Wiebalck)

Arne started by reminding what CERN is, and the basics of CERN-IT, emphasising that there are two data centres used that are 1'000 km apart from each other; the two existing 100 Gb/s links will be supplemented by a third one this year.

The home page will change soon to the new TLD http://home.cern. He discussed changes in the data centre capacity and referred to https://meter.cern.ch. He then covered the status of the three main areas of CERN's Agile Infrastructure project: monitoring, configuration and computing, mentioning the functionality of the monitoring, the growth of the cloud infrastructure now running on 4'600 hypervisors, 40% of which are at Wigner; the OpenStack instance is in the process of being upgraded to Kilo. The configuration area now manages some 18'000 nodes in 240 top-level hostgroups. The Puppet masters were upgraded recently to CC7 and Ruby 2, which has made the compilation time go down to 1/3. Then Arne gave a short overview of the storage services Castor, EOS, OpenAFS, DFS, CEPH and CERNbox. In addition there is NFS, CVMFS, … CERNbox has gone into production, using EOS as the back-end storage; it supports a wide range of client platforms and provides several access methods. The batch service provides some 691k HS06 on 4'200 worker nodes running LSF 7 on SLC6 for local and grid submission; most of these are virtual worker nodes. An upgrade to LSF 9 is scheduled for 13 October (the day after the presentation). The HTCondor phase-in is progressing well; the instance uses the HTCondor CE and currently runs on 1'300 cores; a new Kerberos ticket renewal system is being implemented in collaboration with the HTCondor team and PIC. Hadoop has gone into production; the database-on-demand service has been re-designed and is in the process of being automated further. Concerning desktops, Windows 7 will remain the mainstream version, Windows 10 will soon be available as a pilot. The change of mobile phone operator in June worked quite smoothly; Lync has been renamed to Skype for Business, replacing fixed-line phones by Lync is currently being discussed.

In reply to a question from the audience, Arne explained the status of job efficiency at Wigner; currently there are no known issues.

ATLAS GLT2 (Shawn McKee / UMich)

Shawn explained the current status: the compute power is now reaching some 70k HS06, the services are mostly running on VMware. He showed the monitoring, which is based on OMD and a customised display. For managing the configuration, they use Cobbler governed by CFEngine 3. They use HTCondor and the HTCondor CE. They are prepared to take multi-core jobs from ATLAS with both static and dynamic capacity, and provide a queue for large-memory jobs (4...6 GB). They serve as one of three calibration centres world-wide, which required some changes to the database schemas and replication processes. The services use VMware and iSCSI as well as a DAS backend. Then Shawn explained details about the site resiliency, taking advantage of the two participating sites, which are connected with each other by two 40 Gb/s links; the uplink to Chicago is 100 Gb/s. They are working on deriving alerts from PerfSONAR data, the results of which for WLCG Shawn presented. They have been selected for a grant for software-defined storage research based on CEPH and SDN. Future plans include to participate in SC15, test OpenStack and CEPH, OpenFlow, and dual-stack IPv6 for all their worker nodes, which awaits the proper routing to be set up by the network engineers.

Replying to a question, Shawn clarified that there is no special scheduler (such as Fallow for example) used for the multi-core capacity. Requirements for more than 4 GB per job are hopefully exceptional and restricted to special applications, but the general tendency is that job memory requirements increase.

CNAF (Andrea Chierici)

Andrea started by talking about an incident at CNAF on 27 August, when one of the two main power lines burnt. The fire was extinguished immediately, and power was restored by a Diesel engine. More than half of the extinguishing gas reserves were used, even though the area affected was very small. When the Diesel power was cut, one storage system hosting LHCb data suffered from multiple (65) disk failures, causing some data loss. Additional issues arose due to conditioning systems losing power due to wrong cabling, which has been fixed meanwhile. On 30 August the Diesel generator ran out of fuel, which was not detected due to a sensor failure. It took about the full month of September to completely recover from the situation; LHCb job submissions were re-enabled by 5 September. The link to KIT and IN2P3 has been moved from LHCOPN to LHCONE, as the Geant-based latter one provides better performance. Storage is handled by GPFS and now provides 17 PB. Worker nodes are running diskless. New storage procurements use an IB FDR backbone. The tape storage holds 22 PB and is based on GPFS and TSM; they finished a migration campaign from T10kB and T10kC to T10kD. In order to achieve full wire speed, the C state needed to be disabled in the BIOS of the

tape servers. The computing resources amount to 190k HS06; the 2015 tender resulted in Lenovo blades, which created some issues due to incompatibilities with SL6. Future extensions will probably use facilities outside CNAF; they have run a pilot with an Italian cloud provider, and are in touch with an Italian bank for potentially using nightly cycles. 20k HS06 will be provided by a computer centre in Bari. The batch system is being upgraded to LSF 9, which introduces some issues mainly on accounting. The configuration management is being migrated from Quattor to Puppet. Tests on low-power systems are going on, and will be reported on in the next meeting.

Replying to a question for the audience, Andrea explained that the project with the Italian bank is not very far advanced; discussions have just started.

FNAL (Rennie Scott)

Rennie, in his first site report to HEPiX, briefly introduced FNAL. He has been working for FNAL for eight years now, being responsible for SL and architecture management. Fermilab runs three computer rooms (Feynman, Grid, Lattice Computing Centres). The reliability of the data centres was very high in 2015 so far, which was one of the best years. The ISO 20'000 and ITIL certification is making progress, a re-certification is scheduled for later this month. Technical Scope-of-Work documents have been written or are being discussed with the various experiments they support. The Kerberos-authenticated certification authority will be phased out in September 2016; Web services are moving to single sign-on, Grid services are evaluating CiLogon certificates. Work is focusing on Web site security by moving to central Web services, identifying content owners, and responding to DoE requests to reduce unnecessary Web sites, helped by systematic port scans. They have rolled out a WordPress service; the organisation's Web site is being moved to a HDS NAS system. Plone will be retired by 31 October this year. For distributed computing, they use an HTCondor cluster with 26'000 cores on site as well as opportunistic resources. Access to a central Bluearc NFS server is being restricted because of performance issues; they are moving to CVMFS, leaving the NAS server for dedicated purposes. Monitoring of batch processing uses Grafana. The lattice cluster is moving to ZFS storage. Rennie then gave details on the status of various experiments. They have started the HEPCLOUD project in order to transparently run on commercial and community clouds, grid federations and HPC centres, focusing currently on CMS and NovA. Concerning scientific software, Rennie mentioned art, a generic event-processing framework used by an increasing number of neutrino and muon experiments; there is artDAQ, a compatible framework for DAQ systems as well.

Oxford and SouthGrid (Pete Gronbech / Oxford U)

Pete said that they run two clusters for particle physicists, one for the Grid and one for local usage, both managed with Puppet. They have entirely migrated to HTCondor and ARC CE, and have de-commissioned the CREAM CEs. The current capacity amounts to 6k HS06 and 680 TB; he gave details about the forthcoming procurements. He showed measurements showing that with very minimal impact on performance, the E5-2630-v3 Xeon provides much better power efficiency than the E5-2650-v2. RAL PPD has also migrated to Puppet, actively killing off cfengine-managed SL5 nodes. JET continues to provide some capacity. Birmingham supports ATLAS, ALICE and LHCb; the site is about twice the size of Oxford. In addition, they provide a Tier-3 site. Pete then mentioned Cambridge, who are considering introducing Lustre, even though it may be a little over the top for their requirements. Bristol uses Univa GridEngine as the batch system, deploying cgroups, which they are very happy with. Lustre is being used at Sussex as well; they are moving to 2.5.3. The Oxford machine room needed to be fitted with an additional CRAC unit due to an incident they had. The incident was considered rather serious, as it affected critical University financial services as well. Pete gave some details about the incident and the measures to mitigate it, pointing out that automation is needed to react to temperatures above threshold, as any human intervention will come too late.

PIC (Pepe Flix)

Pepe focused on the changes since the Oxford spring workshop. They have moved the CPU capacity entirely to the main room, and are installing a GCR Carnojet system immersing the CPU in oil for cooling. They are building up an HTCondor-based test bed. The tape system runs Enstore 5.1.0-1 (latest version); five T10kD drives were added, which passed the tests successfully. In addition they are running two generations of LTO drives and T10kC drives. Pepe then covered MCFLOAT, a mechanism to enable multi-core jobs, which is in full operation at PIC now and assures good utilisation of the farm (some 65% multi-core, 96% utilisation). However the CMS multi-core jobs appear not to be fully

efficient. Concerning storage, they are investigating using ZFS with dCache, which does not work well with CentOS 7; they are now looking at FreeBSD. In addition they are considering SSD caches via BTier. They run four dCache instances in total for different purposes. FreeNAS is being evaluated for NFS areas of small projects at PIC. Two new firewall appliances are in place (Fortinet 1500D, max. 80 Gb/s); their connection to LHCONE is now supporting IPv6. WAN traffic has been increasing steadily, sometimes exhausting the 10 Gb/s bandwidth; they are considering deploying meaningful QoS approaches. HTCondor is being considered to replace the current Torque/MAUI batch system; a number of features have already been tested successfully. As CE they consider ARC. Pepe pointed out the high availability of PIC in the WLCG context. He then turned to PAUS (physics of the accelerated universe survey), for which they have set up the data management. In that context they have set up CosmoHUB, an analysis framework for astronomic data via the Web, for which Hadoop has been chosen as the backend.

It was pointed out that the poor performance of the SSD cache could be due to a misconfiguration.

PDSF (Tony Quan / LBNL)

Tony started with an overview of the numerous systems at NERSC. PDSF is a data-intensive cluster for serial, high-throughput processing. It uses UGE with fair-share setup. It is used by ALICE, ATLAS and STAR as well as other experiments such as DataBay, Majorana, … James Botts has now taken over as PDSF lead from Iwona Sakrejda, who retired in June. The cluster is migrated to SL6 and is managed by ansible (was cfengine 3 before). NERSC is scheduled to move back from Oakland to the new building at the lab in 2016; details about this move were given, including how to minimise the impact on storage. Following the final step of the migration in February 2016, remaining systems in Oakland will be de-commissioned rapidly. The choice of ansible was influenced by CRAY, who use it for some of their products in use at NERSC.

RAL Tier-1 (Martin Bly)

Martin explained that they are adding 100k HS06 and 10 PB (for CEPH) to their hardware in the process of the 2015/2016 procurement. A long-standing LAN issue was fixed with a firmware upgrade; the Tier-1 is now connected via 40 Gb/s to the RAL site core network. Next, firewall bypasses and OPN links will be re-established. A lot of work has been going on around IPv6, which is completely isolated from IPv4; they run an IPv6 test bed with a number of dual-stack services. Their CVMFS services include a Stratum 0 for EGI infrastructure and a Stratum 1 for a number of VOs; they are testing options for larger storage. RAL contributes to early testing of new CVMFS versions. Martin then described details about their CEPH system. Their cloud offers SL6, SL7 and Ubuntu images; spare capacity is being used up by creating virtual batch worker nodes. The underlying CEPH storage is proving very stable; the cloud service is moving towards a full production service. Most of the Tier-1 services are running on virtual servers with Windows Server 2012 R2 and Hyper-V. The batch system is running HTCondor, with CVMFS to provide the middleware. They are investigating migrating services from VMs to containers managed by Mesos; concerning monitoring, they are considering InfluxDB and Grafana as a replacement for Ganglia.

IHEP (Qiulan Huang)

Qiulan explained that since the last meeting, storage on the local cluster has been increased to 5.3 PB; they also revamped the internal network of the data centre, separating the data network from the management one for better stability and performance. Concerning WAN, they deploy 10 Gb/s links to Europe and the US. For WLCG, 940 TB storage is provided to ATLAS and CMS. They are using HTCondor (8.2.5 now, which will be upgraded to 8.4.x soon) on 1'080 CPU cores, which they plan to ramp up to more than 10'000 cores. She then described the issue of supporting multiple user groups with a single HTCondor instance. The monitoring is fully integrated with the site monitoring and shows resource usage by user group. Their Lustre system provides a total capacity of 4.5 PB; some more details were given, including a read performance of 800 MB/s. Hardware and software have been upgraded, replacing traditional RAID 6 arrays by Dell DDP systems, and upgrading Lustre from 1.8.8 to 2.5.3. They have experienced issues with Lustre, where the MDS got stuck by large-scale lock time-outs. They also run Gluster with a total capacity of 734 TB (recently upgraded from 347 TB before). Their AFS system provides 19 TB; they have recently added monitoring via Nagios for replica consistence. They run a cloud based on OpenStack Icehouse with one control node and 20 compute nodes providing 229 VMs. They are running a testbed on the WAN for SDN; performance data will be released soon.

She then showed performance figures for their Tier-2 contributions to ATLAS and CMS with good availability and reliability. A number of issues occurred that needed to be fixed one by one. Their monitoring included log analysis via Flume, ES and Kibana; they use it for syslogs as well as logs from the storage systems. In future, more information will be collected and fed into Hadoop. IHEP has now joined the eduroam community, supporting username/password for now; certificate authentication will follow. She then described the service desk, Vidyo at IHEP, and IHEPbox.

A comment from the audience remarked that the worker nodes procured for WLCG are poorly equipped with RAM and disk with respect to the powerful CPUs. Qiulan explained that the separation of the HTCondor instance by user group is not a static partitioning of the cluster.

Grids, clouds and virtualisation

Simulating 5 Trillion Events on the Open Science Grid (Martin Purschke / BNL)

Martin explained that PHENIX, after 15 years of running, will come to an end. A new project, sPHENIX (previously SuperPHENIX) was designed, for which he is the data acquisition coordinator. He referred to a talk he gave back in 2006 at CHEP in Mumbai. The sPHENIX experiment is based on the ex-BaBar magnet. The detector is simulated in detail by extensive G4 simulations, which are taking some 1/3 of RHIC computing. Much more statistics is needed… hence the question arose whether they can simulate 5 trillion events. Earlier cuts are not possible, the full simulation chain is required, of which only 1 out of 1000 events need to be retained. Even two years ago he would have said that this would not be possible, but… He listed the numerous challenges linked with this idea; full automation is really key. One of the important questions is how to get the job to the remote node and back to RCF. In the end they settled on HTCondor file I/O. A job needs some 1'550 MB in executables and shared libraries; everything compressed is about 253 MB. One issue encountered was the diversity of worker nodes, some jobs crashed due to lack of standard system-level libraries. He packaged another 60 libraries and added the package's download to the workflow if needed, which brought the failure rate down to much less than 1%. So far the project has delivered 153 years of CPU. He expects that there will be a large push for more simulations; they have meanwhile managed to set up a CVMFS area, and he is trying to get the production out of the 'expert-only' mode.

In the discussion, it was pointed out that at SLAC a package has been created that contains a number of frequently missing libraries.

End-user services and operating systems

Scientific Linux update (Bonnie King / FNAL)

Bonnie started by noting that the team has been enlarged in terms of number of people, but not in terms of FTE, as they go for a devops approach. SL 5.11 is the last minor version of the 5 series and will stop being supported in March 2017. SL 6.7 was released on 11 August with driver updates, a newer OpenAFS version etc. SL 7.1 was released in April; it comes with the OverlayFS as a technology preview. The focus is now on HEP community site configurations, a contextual framework, and SL Docker images. SL is now part of FNAL's release and change management and fully integrated into their ITIL procedures. For the future they plan to include OpenSCAP (Security Content Automation Protocol), and to explore HEP use cases for Katello (Satellite 6 upstream) and container technologies.

Replying to a question, Bonnie clarified that they are focusing on Docker as container technologies, because that is upstream's choice, but they would be willing to consider other technologies if that was required. It was noted that the Overlay file system still has a number of issues. Despite the name, the team is reluctant to put a lot of scientific software into the distribution in order to keep it as close to upstream as possible.

Linux at CERN (Arne Wiebalck for Thomas Oulevey / CERN)

Arne started by explaining that as production OS, CERN supports SLC 5.11, SLC 6.7, CC 7.1 and RHEL. Infrastructure is in place for package distribution, AIMS, and Koji. ZFS on Linux has been rebuilt and added to SLC 6 and CC 7. A new project is to use Puppet to configure desktops in order to replace the legacy Quattor-based tool, leveraging as much as possible on the Puppet effort in the computer centre. It will be a master-less Puppet, maximally re-using modules already written. CERN CentOS 7 contains the upstream RPMs without CERN-specific customisation, plus some CERN-

specific additions; updates are delivered in a staged way, contrary to upstream CentOS with its continuous flow of upgrades. Then Arne covered the community build service (CBS) that CERN has set up for centos.org; all SIGs of CentOS use this service. SIGs CERN is interested in include the cloud, software collections, virtualisation, Atomic SIGs as well as those on alternative architectures. The CERN team appreciates the support by upstream (both CentOS and Red Hat), and the collaboration with the community. Areas for improvement include better integration with EPEL, SIG roadmaps, and more non-Red Hat contributors.

There were questions about the difference between CC 7 and SL 7 in view of supporting point releases. It was pointed out that the only real difference is that CC7 does not support updates of previous point releases; the same is true of SLC 6 and SLC 5.

Software collaboration tools as a stack of services (Borja Aparicio Cotarelo / CERN)

Borja explained the current offerings of the team around SVN and Trac on the one hand, and Git on the other hand. For better support of code review and collaboration, Gitlab has been set up. It is a service hosted at CERN similar to Github, and is not at all intended as a competition, but to cover use cases that require confidentiality or close integration with the CERN environment. Gitlab offers code review, issue tracking integration, merge requests, integration of CI engines, well-defined roles within a project or a group of projects etc. CERN holds a licence for Gitlab Enterprise Edition, which provides full integration with LDAP e-groups. Currently, after six months of production, there are 1663 projects and 1540 users, not far from the statistics of SVN. The team has developed the Kerberos integration and has contributed it upstream. Continuous integration has become very important for all large software projects; the team offers a Jenkins platform delivering a Jenkins master and taking care of common tasks; some 45 platforms have been requested, which shows that there is a real need. Issues with CI include the fact that each Jenkins master requires a dedicated VM that needs to be provisioned and orchestrated; automation and flexibility are not as advanced as they would like them to be. That's why the team has been considering a real PaaS-style CI service providing freedom for user customisation, scalability, and self-service. The team is currently looking at Openshift as a candidate solution, which looks promising. Borja then covered SVN and Trac; even though they are not recommended for new projects, many existing projects very heavily use them; it would not be conceivable to migrate all users to gitlab any time soon.

Answering a question, it was clarified that there are no concrete plans to de-commission the SVN service, but discussions with major user groups have started. While Trac can in principle be used with git, this is not the recommended solution. The existence of Gitlab at CERN is not meant to invite people to migrate from github unless they need the data to be hosted at CERN or the tight integration with the CERN environment.

PPDcloud (Chris Brew / RAL)

Chris, representing members of his team, explained that their use case was to provide cloud access to files stored at RAL, and to provide a simple equivalent to Windows Offline to Linux and Mac users. They considered buying integrated storage with cloud interfaces as well as software providing cloud interfaces on top of existing storage solutions. After a short comparison of available solutions, they decided to take a close look at ownCloud, which is free. They designed an architecture to provide access to DFS space. Reverse proxies are implemented on two independent VMs. They set up two Linux VMs with ownCloud, Apache and Samba, with sessions shared via memcached and accepting connections only from reverse proxies. Databases required for ownCloud were set up with PostgreSQL. The storage backend is their DFS space without any specific configuration for ownCloud. The service is in the late stage of pre-production, they are now working on the documentation. Issues found so far are that the file locking and conflicts are not handled as smoothly as in Windows Offline; the tablet client requires files to be cached individually; there is no Windows phone client. In future they will consider branding, Office 365 integration, and local storage and sharing.

Self-service kiosk for Mac and mobiles (Tim Bell / CERN)

Tim explained that the number of PCs and Macs is pretty stable at CERN, while mobile platforms are taking off sharply, in particular Android. Some mobile devices are sometimes used for professional purposes. Commercial solutions for managing mobile devices exist, but were found to be way too expensive; in addition there is no single product covering Android, iOS and Mac. They looked at JAMF, the implementation of which Tim explained. For Macs, the MDM pro-

vides an interface. At CERN it has been set up with software packages (open-source ones such as Gimp, LaTeX, OpenAFS, commercial ones including Office 2016, Parallels, Anti-virus, and CERN customs applications, for example CERNbox) as well as configurations (disk encryption, printers, …). iOS was found to be quite more difficult, because the range of use cases was much larger. They set up a community 'self-help' support, but then wanted to support CERN-developed applications as well. A professional iOS use-case is the fire brigade. Administration of managed iOS devices is delegated to user groups. Access to hardware parameters, owner information etc. is granted, but not all functionality for central management is available. Then Tim covered the VPP, the Apple Volume Purchase Program, as a way for CERN-IT to acquire applications and manage the related licences. About 10% of the Macs are now contacting MDM.

The objective is to provide the same applications as on the dedicated DFS area. The scripting language is similar to bash. When a licence is taken away from a user/device, the application will not be de-installed.

Facilities and business continuity

NERSC Plans for the Computational Research and Theory Building (Elizabeth Bautista / LBNL)

Elizabeth explained that the building is a four-story 140'000 square feet building costing aobut 143M USD, far more than was anticipated; the building uses entirely free cooling. Supply to the building amounts to 27 MW of redundant power with 1.0 + 0.5 MW UPS backup. The PUE is expected to be below 1.1 thanks to the fact that the temperature is pretty much around 75 deg F all year round. The exhaust heat from the centre will be used for heating the building. Because of the beautiful views the building, although not fully finished yet, is already much demanded for meetings. Elizabeth then explained the mechanism for seismic protection and presented a video of a test scenario, in which the racks survived a simulated earthquake very well. Part of the PDSF systems have been moved in, as has Cori-P1, one of the large Cray systems. 24x7 operations have started.

Replying to a question, Elizabeth said that the warmest period is about mid September to end October, with temperatures up to 95 deg F, which is still appropriate for cooling IT systems.

Miscellaneous

From engagement to employment – growing our own system administrators (Ian Collier / RAL)

Ian explained that the initiative grew out of an apprenticeship programme. RAL has a lot of experience with successful public engagement through open days, teachers' programmes, programmers' days etc., for which he showed impressive statistics on the number of people reached (even though not all of them visited RAL physically). On the other hand, RAL has faced increasing difficulties recruiting and retaining staff. Graduate recruitment programmes are seeing ever fewer applicants. At the same time, university education in the UK has become much more expensive, with tuition fees of 9'000 GBP per year; vocational training and apprenticeships are a hot topic, with some funding available. Quite some 16...18 years olds are met by public engagement, a number of whom have the right skills to be system administrators or developers. The idea emerged to train apprentices on computing. Ian looked at apprenticeship qualifications, which seemed over-complex. He then turned to degree-type education and found a local college (in Abingdon), which was very keen on collaborating and adapting. When discussing the idea at the lab, other departments were expressing an active interest as well. In June 2015 it was agreed to proceed and recruit four apprentices funded by lab's sources rather than government funds; they are now sorting out the details of rotating the apprentices between departments at RAL. In August they selected the first three apprentices, who started in September… who will hopefully show up at HEPiX some day reporting on their work.

Responding to a question, Ian explained the structure at RAL for public engagement, allowing for reaching out to 243'000 11-18 year olds in 2014.

Site reports

LAL and GRIF (Michel Jouvin)

Michel started by stating that after two years of production, the experience with the new computing room at LAL is positive, it has allowed for smooth running. The room hosts 30 racks providing 250 kW (with further upgrades a maximum of 400 kW could be reached) to IT equipment. Having redundant chillers has been very beneficial. The cooling system is based exclusively on rear-door heat exchangers. An extension is planned, with ensured funding, for a total of

900 kW. At GRIF, there is a decrease of person-power by 5...6 FTEs. At LAL, they are perhaps at the beginning of a better time, they just hired a system administrator for grid and cloud resources, and have had an apprentice starting. The hardware is fairly old, and securing funding for renewal has been challenging. For servers they standardised on dual twins. The almost 10-year old DDN arrays were replaced by DELL MD3xxx iSCSI or SAS boxes with 100 TB per server; it performs well and is reliable. Most systems run SL 6, but they started deploying some EL 7 machines, considering CentOS as well (but they are not happy about the non-support of previous point releases). They moved to a unified authentication infrastructure based on Windows Active Directory and Kerberos, which required sorting out inconsistencies between Unix and Windows accounts. Windows on the desktop mostly uses Windows 8, but Windows 7 is still very important as well. It is difficult to convince the users of Windows XP systems to get out; Windows 10 will soon be supported. They are involved in a number of projects including Cyclone (enabling technologies for cloud federations); IRFU is participating to Indigo data cloud with similar responsibilities. He then turned to cover GRIF, a federated Grid site in the Paris region comprising six labs. The federation is already more than 10 years old and is, like everybody else, suffering from tight budgets and manpower. Three of the six sites have moved to HTCondor, one of which moved to ARC CE, the others stayed with CREAM. LAL has been involved in cloud activities for a long time. The resources inherited from the Stratuslab project were turned into a university-wide resource; the required doubling of the resources was funded by the university. LAL are considering virtualising the worker nodes. At LAL they also have the P2IO project, an initiative across HEP, nuclear physica and astrophysics, the facility of which is located at Ecole Polytechnique. In that context they have been doing R&D work on CEPH; current performance is not at the expected level quite yet.

Tuesday 13 October 2015

Site reports

Jefferson Lab (Sandy Philpott)

Sandy explained that in 2016 JLab will invest 1M USD into HPC for USQCD. They are currently investigating Intel Phi/Knights Landing, NVIDIA Pascal, and Broadwell CPU servers. Factors to consider are hardware availability, networking capabilities, benchmarks, and available configurations. Concerning storage, they just finished upgrading to Lustre 2.5.3; 2014 disk hardware is in production (but there are stability issues possibly due to server memory), 2015 hardware has just been received. Dell MDS systems have been put into production; 2009 storage kit is going to be retired. They are carefully watching the CEPH space. The IBM tape library has been re-located, replacing some low-capacity frames by fewer high-capacity ones; a number of tapes have been reported missing since. The facility is being upgraded with the goal of reaching the DoE objective of a PUE of 1.4; the power and cooling capacities are being doubled. On the roadmap for the next 12 months are an update to CentOS 7, deploying the first new HPC cluster, installing a second tape library, a project on data management and data mining etc.

Replying to a question, Sandy explained that the target rack density of 18 kW is achieved with air cooling. The Knights Landing version considered is the mainboard one.

Wisconsin CMS Tier-2 (Ajit Mohapatra)

Ajit explained that they operate two machine rooms with 17 racks and 650 kW power supply. The Tier-2 is running SL 6 on 6'700 cores, to which 1'200 are being added. In addition, typically 1'500 cores are available opportunistically from the HTCondor HTC pool at the university. For four years they have been running Hadoop with a usable capacity of 2.25 PB. CMS accounts for 65% of the CPU usage. The CE is being migrated from Gram to HTCondor; HTCondor is of course used as batch scheduler. A small fraction of nodes is supporting multi-core jobs; more can be configured on demand. Their switches are connected to the routers via 100 Gb/s links; the link to Chicago is also 100 Gb/s. The role of perfSONAR for finding latency and bandwidth issues is much appreciated. 20 Gb/s from and to Caltech have been demonstrated. Work is going on on IPv6; a total of 380 machines is running dual-stack to the outside world without any issues. OSG services are working fine both with IPv6 only and with dual-stack, with a few exceptions, most notably Hadoop that can only run IPv4. Ajit also showed a list of software, services and tools for management and monitoring they use. Their experience with the HTCondor CE is good in general, but there have been hiccups, in particular when upgrading to 8.2.9 (issue will be fixed in 8.2.10). They use MOUNT_UNDER_SCRATCH and HTCondor's cgroups' functionality. Xrootd 4.1 and 4.2.2 have found to have various issues; 4.2.3 (in OSG testing right now) seems to fix these

issues. Future plans include an upgrade of the older machine room for new cooling installation, further growth of the farm as well as the use of opportunistic resources, and a network bandwidth extension, moving the worker node connectivity to 10 Gb/s. While the site availability is good in general, there have been power incidents coinciding with major sports events in two consecutive years.

Following a question, Ajit explained how cgroups are deployed. Containers are not yet used (except that using cgroups is pretty much equivalent to using containers).

Australia-ATLAS (Lucien Boland / Melbourne)

Lucien explained that their efforts are embedded into the Centre of Excellence for Particle Physics at the Terascale (CoEPP) with 100 researchers and 4 sites focusing on ATLAS, but now supporting Belle II and other experiments as well. Funding is assured until 2017, and has been requested for 2017 to 2024 – writing a proposal with a seven-year forecast on budgets for computing that is new and exciting is of course challenging. The Melbourne services are run by himself and Sean Crosby, while Goncalo Borges looks after the Sydney site. Sean spent three months at CERN working with the cloud team; the collaboration was considered successful; they are looking for future opportunities, including inviting experts to Australia. The total farm capacity amounts to 10'700 HS06 and 950 TB storage for ATLAS; they are still using Torque/MAUI, hence some tweaking was necessary to configure for multi-core jobs. Some smaller Tier-3 installations exist as well. Services are running on virtual machines provided by XenServer; they use Puppet, Nagios, Ganglia, rsyslog, gitolite and Dokuwiki. In 2015 they added RT and Gitlab. In 2014 they recycled an IBM iDataPlex system with 80 nodes and Infiniband MPI connections; even though it is out of warranty, it doubles the CoEPP compute capacity. Trying to kick off the installations, they came across some surprises and specificities of these IBM systems, but they are now almost ready to make the capacity available to their users. Likewise, out-of-warranty DPM disk servers have been used to set up a CEPH system in order to benchmark RADOS and CEPHFS as well as to study failure behaviour and recovery. They have been experimenting with AWS resources made available to them via an academic grant, setting up a Belle II site successfully. They are envisaging collaborations with BoRG, Belle II, and the CEPH developers, and plan for a number of infrastructure improvements. Finally, Lucien pointed out that he is interested to hear from other small teams how they handle operational tasks, end-user support, project work, community contributions and others.

Replying to a question, Lucien explained that concerning CEPH development, file system recovery is a topic that is of interest to them. They use EGI software as Grid middleware.

NDGF (Mattias Wadenstein / Umea U)

Mattias started by reminding the audience that NDGF is a Tier-1 distributed over six sites using ARC CE and AliEn as well as dCache (at seven sites). It is run by NeIC. The networking is being upgraded moving to sharing links guaranteeing a minimum rate of 20 Gb/s. NORDUNet has been extended to Geneva; the direct link to SARA will be de-commissioned. The central services have been moved to a redundant Ganeti cluster with DRBD except for PostgreSQL, which continues to run on bare metal. The ARC CE now supports remote cache reads, effectively sharing caches across CEs. Mattias also mentioned the HTCondor/ARC workshop in Barcelona starting on 29 February 2016, and presented a view of the ARC control tower, which shows a high usage level for BOINC from ATLAS. Concerning storage, they are running dCache 2.13.9 in a fully dual-stacked configuration. Endit is a development project for an HSM provider for dCache, allowing to link dCache with TSM that Mattias is working on. In that context, IBM fixed a TSM issue limiting the number of files that can be specified in a 'dsmc retrieve' command. Concerning NDGF, the PDC site in Stockholm has been fully de-commissoned; Mattias also reported an electrical incident in Copenhagen. He finally mentioned new dCache pools at NBI.

Wigner (Szabolcs Hernath)

Szabolcs explained that their link with the community is twofold, they run the large Tier-0 extension as well as a small Tier-2 centre. For the Tier-0, they provide 4 modules with maximum 1 MW each; the fact that the real consumption is much lower is a serious issue, as is the free cooling and the building management. Issues explained in Annecy have partly been resolved (A/B power balance, 3-phase balance), while others are still open (understand under-utilisation, free cooling, improve monitoring). Even for rooms 1 and 2, the consumption is at the 300 kW level, while the rooms

have been optimised for 700...1000 kW. As a consequence, the PUE is somewhere in the region of 1.6, with relatively large variations due to equipment movements as well as to deficiencies of the monitoring, while the target was 1.5. The impact is that the chiller cluster is regulated very poorly; free cooling does not work, time shifting and state transitions are not reliable; several firmware upgrades have not fixed the issues. There has been no correlation between the outside temperature and the PUE, which demonstrates that the regulation mechanism does not work at all. Szalbolcs then explained three major incidents during the past 18 months, two of which were due to human errors, while the third one was due to an exploding IGBT taking a UPS out of service. The monitoring had its share of responsibility for the incidents, as it was not working reliably (false alarms, no alarms when they would have been needed etc.); a new independent system is being set up now, and will operate alongside the proprietary monitoring. In 2015 the power consumption is estimated to amount to some 6.3 GWh; there is an impressive number of tickets as well as hardware interventions for components.

Answering to a question, Szabolcs explained that the computer room set point is 25 deg C, and that the issues with the free cooling are partly due to software, and partly due to the underload of the facility.

KEK (Tomoaki Nakamura)

Nakamura started by reminding the audience about the main KEK facilities in Tsukuba and Tokai at distances of 60 km and 120 km from Tokyo, respectively. They run accelerators for particle physics as well as light sources. At Tokai they run J-PARC, an accelerator producing hadrons, muons and neutrinos used for particle and nuclear physics as well as life and material sciences; he gave some more details about the experiments, which include Super-Kamiokande at about 300 km distance from the neutrino source; Nakamura reminded the audience that the detection of neutrino oscillations at Super-Kamiokande gave rise to a share of the 2015 Nobel prize in physics. He then mentioned the ILC developments and KEKB, the B factory, with the Belle II experiment, for which he gave an outlook of the computing resources, to which KEK will contribute 20...30%. The Belle II computing model is very similar to the WLCG experiments' one. He then explained the current computing setup at KEK with 4'080 CPU cures, 6.9 PB of IB-connected disk storage, a tape archive with HPSS, and EGI middleware as well as LSF 9. The usage level is fairly high (90%), with Grid submissions taking an increasing fraction. They are connected via 100 Gb/s to Los Angeles and Seattle, and via 10 Gb/s to New York, Singapore, and London (2 lines). For August 2016 a 100 Gb/s link is scheduled to LHCONE. Currently a system upgrade is going on, the procurement process is running. In view of usability and flexibility of resource allocation, they are considering cloud resources.

Replying to a question, Nakamura said that a second room will be used for improving the roll-over process from one hardware generation to the next.

KIT (Andreas Petzold)

Andreas started by explaining that the SCC is not only about computing for WLCG, but also for a number of non-HEP use cases. The farm for WLCG provides a total of 180k HS06; 154 new worker nodes have just been added, which are equipped with 16 physical cores and 96 GB of RAM each and configured for 24 job slots per node. They found that UGE was killing jobs unexpectedly; they remedied this situation by letting cgroups do the work instead. MJF (machine-job features) are fully implemented. The fraction of multi-core jobs has been varying widely. KIT participates in the HEPiX benchmarking working group, the WLCG multi-core task force, and the WLCG MJF task force. Their disk storage, amounting to 14 PB, is currently entirely based on DDN kit; an extension by 2.4 PB is in the pipeline, 3.2 PB replacement capacity is foreseen for 2016. A total of six dCache instances are configured; for ALICE they run an xrootd interface. Tape storage is based on TSM, providing 19 PB in three libraries based on T10k technology for the Tier-1. In addition there is one library for HPSS, which is supposed to take over from the existing TSM systems. They are moving their configuration management to Puppet with Giblab and Gitlab-CI, Foreman, and Puppet masters separately by project. They are very interested in ELK and have many ideas what to do with it in order to improve their monitoring, but don't make much progress because of manpower constraints; prototypes exist for dCache, UGE and LSDF. They have just acquired a new HPC machine with compute nodes from Transtec and Dell storage. The new building has been finished, machines are being installed into it right now. Heating and cooling, as well as power, is being installed

campus-wide, replacing a number of dedicated cooling installations. This will certainly imply a major (~ 2 days) down-time of the services, but they hope to be able to keep the disks running.

CSC (Johan Guldmyr)

Johan first covered the interactive shell based on the SLURM scheduler and sshd, which he had explained in detail at the spring 2015 workshop; the service is running stably and with good performance, and has been appreciated by users. The Tier-2 has had a number of issues, many of which were related to certificates. For ALICE, they run a CE in Openstack. The HPC facility is moving from SL 6 to CentOS 7, from bash scripts to Ansible, from NIS to LDAP, from Ganglia to Grafana etc. Johan presented a sample view of a Grafana display.

SLAC scientific computing (Yemi Adesanya)

Yemi explained that he is acting as director for scientific computing services, having served for 17 years in other roles at SLAC on BaBar computing, OO frameworks for data analysis and visualisation. The team comprises 10 FTEs charged with supporting a number of activities around the support of very diverse sciences. They are providing small to mid-range HPC solutions, assistance in the development of computing models, are supporting the Unix-based core infrastructure, and are building a computing community across the lab. He then explained the funding model for the services they provide; standard services receive indirect funding. They provide storage on a pay-as-you-use basis, and currently manage 12.5 PB of storage space, of which 20% is NFS-based. The storage is scattered across many different solutions including Thors and Thumpers, a few Netapp boxes etc., creating many different single points of failure; they plan to replace this by 'storage-as-a-service' based on a large multi-tenant scalable GPFS file system, attractively priced for groups to take advantage of it. At the same time they are reviewing their tape strategy, moving from T10kB and C to T10kD, again introducing a storage-as-a-service model based on an archive tier of storage. The common theme is decoupling users from the underlying hardware. They are looking into OpenStack and started by purchasing a Nebula appliance late in 2014; shortly afterwards the company disappeared from the horizon. Their goal is now a vendor-supported, production-ready OpenStack environment, they are in discussions with Dell and Cisco, but are also looking at the upstream RDO OpenStack distribution. For batch scheduling, they are using LSF across about 19k bare-metal cores; large parts of the configuration originate from the 1990s. They are on 9.1.3 on new hardware now and take advantage of enforced limits on wall-clock times and cgroups. They are now testing LSF's dynamic host support (machines appearing in a certain IP range are added automatically into LSF, without the need for re-configuration). Similarly to disk and archive storage, they are now looking into charge-back for shared compute services, and have purchased LSF Analytics, a turnkey solution for utilisation reports and dashboards.

Following a question from the audience, Yemi explained that they have not seen yet any indication of scalability limits of LSF. They are considering moving to a leasing model (with return of the boxes after the leasing period) in order to cope with the lifecycle of the kit.

DESY (Wolfgang Friebel)

Wolfgang reminded the audience that there are two sites, a large one in Hamburg and a much smaller one in Zeuthen near Berlin. They are preparing EU tenders for computing equipment (desktops, notebooks, monitors, small items as well as compute and storage servers) for up to four years, with the objective of selecting at least two suppliers. They have introduced a classification of Linux desktops depending on the level of administrative rights for the users. In addition there are classes defining the delay for applying updates. At Zeuthen they are running three batch farms, one for general purposes, one low-latency farm for HPC applications, and one Grid farm; they were all recently upgraded to UGE 8.2.1, which supports mixing unicore and multi-core jobs flawlessly. Some 80 accelerators (nVidia GPUS) have been added to the general-purpose farms. They have improved the monitoring of the batch systems, allowing users to inspect running or recently finished jobs. In Hamburg, data rates delivered by the experiments at PETRA require powerful storage solutions; GPFS has been chosen for high performance and throughput; 0MQ is used for the file transfer into the machine room, as file-based mechanisms were found not to be sufficiently performing. The data distribution can use multiple targets with different requirements. As all raw data and precious reconstruction and analysis data are stored, the growth rate of the data volume is very substantial. In Zeuthen, they run Lustre, which they have recently upgraded to 2.5.3. ZFS was considered, but not adopted. DESY have introduced a self-service for SVN; they currently

host some 400 projects. They are also investigating OpenStack. The migration from Exchange to Zimbra has completed, DESY is now an “Exchange-free zone”. Finally Wolfgang cordially invited the audience to the spring meeting from 18 to 22 April 2016 in Zeuthen.

BNL RACF (Shigeki Misawa)

Misawa said that the site was acting as Tier-0 for RHIC, Tier-1 for ATLAS, and in addition supports a number of other experimental activities. They also play a major role in BNL's centre for data-driven discovery, for which they provide support. They need to increase the facility capacity and are planning to build a new data centre in the former NSLS building. The group provides the disk storage for the new BNL HPC cluster, making all data accessible from both HPC and HTC clusters, they also provide archive storage for the accelerator unit. They have collaborated with the networking group to propose a HPC Core network providing high performance for multiple purposes. The core characteristics include scalable connectivity at data centre and campus distances with line rate capability; security is achieved by selectively enabling connectivity between network regions. This allows for example for placing DAQ storage in the data centre rather than close to the experiment, and for flexibly allocating resources to one or the other user group. The first application of this idea was for the CFN electron microscopy group, working on an instrument that is capable of generating 4 TB in 15 minutes. They also migrated the US ATLAS Tier-1 WAN presence to the BNL Science DMZ, and phased out 1 Gb/s switches for Linux farms. In RACF they have some 52'000 cores in 2'300 servers; they have tested NVMe SSD drives. They are running SL 6 and are testing SL 7 as well as Docker. They run an HPSS systems with LTO and T10k drives, actively migrating from obsolete formats; most aspects of HPSS have been upgraded and modernised recently. They also run some “legacy” NFS as well as GPFS systems and a production CEPH cluster.

Security and Networking

News from the HEPiX IPv6 working group (David Kelsey / RAL)

David reminded about the 2015 deployment plan presented in the spring meeting, which included LHCOPN/LHCONE IPv6 routing by August, perfSONAR dual-stack monitoring, significant production dual-stack data services, and the move of important central services. He then showed that North America ran out of IPv4 addresses on 24 September 2015, and that Google statistics show a steep increase of IPv6 traffic, now reaching almost 9% of the global traffic. The group has run testbed activities with gridftp and later dCache FTS3, hitting problems every now and then. In addition, Tony Wildish, who had looked after the testbed for years, left HEP; Ulf Tigerstedt has taken over. For the experiments, understandably since the start of Run 2, IPv6 is not the highest of their priority; LHCb had DIRAC problems, ALICE needed to update to Xrootd 4.1.x, ATLAS is debugging dual-stack CE problems. The aim is to gradually move to dual-stack services, in particular for storage; many sites, as reported during the workshop, have made good progress. It is important that IPv4-based services do not break. A number of issues were found on the way, which shows that care must be taken when services are moved to dual-stack. David then explained what the IPv6-ready flag in the CERN network management means. PerfSONAR monitoring is available to monitor IPv6 traffic. With LHCOPN and LHCONE, the objective is to ensure the same network performance for IPv4 and IPv6; CERN and six Tier-1s have enabled IPv6. Security issues were discussed during the face-to-face meeting at CERN. In future many Tier-1 sites will be IPv6 capable; in 2016 Tier-2s will need special encouragement. Volunteers are always welcome, and can learn a lot from the working group colleagues.

In the discussion it was pointed out that at least part of the issues involving CERN were due to misconfigurations.

Status of OSG compliance with IPv6 (Edgar Fajardo / UCSD)

Edgar emphasised that the work he will present is the result of collaborative effort across labs. They understand compliance to mean that the software still functions if both client and server are dual-stack, and that it works correctly if one is IPv4 only. Tests have been run between FNAL, Wisconsin and UCSD. Out of 44 packages, 23 were tested and compliant, while 12 were not. Storage services are compliant except for Hadoop. File access has issues with the Bestman and dCache clients; in the area of authorisation and authentication, VOMS, GUMS, gLexec and EDG-mkgridmap have the same issue. The CEs and job submissions work. Then Edgar turned to the GlideinWMS case, explaining the main concepts. He successfully tested with HTCondor and Gram; for the other CEs compliance is claimed, but not tested. In the HTCondor Vanilla universe, the communication between startd and the central manager still had issues.

Replying to a question, Edgar clarified that the file transfer tests did not include any delays. They have used tcpdump to be sure whether IPv6 or IPv4 was used for a given connection.

Network infrastructure in the CERN data centre (Eric Sallaz / CERN)

Eric started by explaining that CERN's requirements included support for 40 GbE and 100 GbE as well as 10 GbE, which means optical infrastructure compatible with LC connectors as well as copper. For management and PDU networks, copper is of course fully sufficient. For high availability and reliability, all advices and best practices on installation methods must be followed; the cabling must be properly organised; the right operational methods need to be used (inspect and test before you connect). The current infrastructure is rather old (1995) and mostly copper-based; since then the data centre has been extended at Wigner as well as locally. The racking is a mixture of types and techniques (water-cooled vs. air-cooled). They considered copper for management and PDU; for the data, the need may come to support 10 GbE to the worker node. For optical fibres, the choice is between OM3 supporting 100 m and OM4 supporting 150 m of 100 GbE. For the connectors, the choice is between MPO (mandatory for 40 GbE and above) and LC; for MPO, there are MPO-12 and MPO-24. With MPO there are three different cabling types (straight, crossed, crossed by pair) that Eric explained in detail; if 10 GbE is to be used, the MPO will need to be mapped to LC connectors, for which again various solutions exist. The proposed implementation foresees infrastructure for every 3...5 racks, new copper in new rooms, and a re-use of existing copper for management and PDUs in existing rooms. OM4 72 fibres are proposed with MTP-12 type B.

On a question from the audience, Eric clarified that for the data centre, single-mode fibres are not considered.

WLCG network and transfer metrics working group (Shawn McKee / UMich)

Shawn explained that the working group was created one year ago, rand eminded of the mandate and of the membership. The group has agreed to work on topics via monthly meetings. They started by collecting feedback about use-cases into a single document; they also reported at CHEP (see forthcoming proceedings). Slow transfers can have many different causes; a core task of the group was perfSONAR deployment and commissioning. PerfSONAR allows for discrimination between network and application problems. Network metrics including proximity information are made available via various mechanisms in close collaboration with OSG and ESnet. Once collected, the metrics is integrated into other monitoring information, of which the experiments can also be consumers. Shawn gave details about a sample implementation targeting FTS transfers, which showed asymmetries between ATLAS and CMS; more results are awaited. The data has also been bridged with LHCb Dirac, and is used by ATLAS for finding the right balance between storage and network utilisation. One observation is that the infrastructure has been tuned for long flows and big files – is this the right choice in the long run? Another use case is a support unit that is able to intervene on potential network issues. WLCG-wide meshes for latencies, routes and bandwidth have been set up. The next steps are to continue the collaboration with the experiments, integration of the higher-level infrastructure, integration with analytics data stores etc.

Update on perfSONAR infrastructure in OSG and WLCG (Shawn McKee / UMich)

Shawn started by reminding the audience about the goals and the achievements so far, and showed the status of the deployment, which counts 278 perfSONAR instances now. Standardised metrics allow for a more precise and faster analysis of any networking issue. Standard metrics include packet loss, delays, bandwidth and TCP retries etc. Regional meshes are possible, but are not enabled yet; the right granularity still needs to be decided. Project meshes have been defined, e.g. for Belle II. Shawn then showed the flow of perfSONAR information. Data about OSG are collected entirely since September 14, with the idea of keeping the data in the long term. Then Shawn explained the proximity concept in detail; among the various options, they have chosen GeoIP as the basis, finding that Geant and ESnet are very interested in working on an open-source project to develop this service further. There is the option to configure 3rd party tests between remote instances, which can be very handy for debugging. Then Shawn showed the dashboard and pointed out that the next version will support notifications. Path analysis is supported as well. PerfSONAR 3.5 was released on 28 September; most sites have already migrated. There are various options for installing the package, enabling more or less functionality. A release for low-cost/low-power nodes is being prepared as well; the small form

factor is handy in a number of installations. Running on VMs is possible, but still not recommended; Docker is on the radar screen.

In the discussion, help was offered concerning best practices for deploying perfSONAR on VMs.

Using VPLs for VM mobility (Carles Kishimoto / CERN)

Carles first stated that the CERN data centre comprises 1'000 racks at CERN and 300 racks at Wigner; the interconnect is ensured via two 100 GbE lines. The access layer is implemented as top-of-rack (ToR) switches with single or multiple 10 GbE uplinks into the distribution layer, which relates with the core via 100 GbE links. The network is routed (OSPF); there are no VLANs nor spanning trees; a lot of use is made of ECMP and LACP. Since 2013 dual-stack is fully supported by the network layer. MPLS is deployed as well. Occasionally there is need to migrate VMs transparently to new hardware in different racks, keeping the same IP address. However IP services do not extend across racks. A number of solutions have been considered, for example fibres between switches, a solution that does not scale. Then they considered fibres between routers, which allows for connecting switches logically via VPLS (virtual private LAN service), a way to provide multipoint-to-multipoint communication. Basically VPLS emulates a large Ethernet switch for the data centre. An issue is that the switches concerned will lose external connectivity for some 10 seconds, which was considered unacceptable. They then had the idea of fitting the distribution routers with loop cables and having the traffic pass by the loop before going to the switch. With a bit of priority tweaking in the router, this can achieve the desired behaviour. Carles then described a potential work flow at CERN for creating and removing the circuit. Tests in production will follow.

Computer security update (Liviu Valsan / CERN)

Liviu started by saying that there has been no major evolution apart from the fact that exploit kits have become ever more sophisticated, and that mobile devices are increasingly targeted; malvertisement is getting increasingly prevalent. The market for exploit kits is evolving very quickly, as Liviu showed with the evolution of market shares of various exploit kits. The dominating kit, Angler, is primarily serving krypto-lockers asking some 300 USD per case; the economic model is impressive. All kits target Internet Explorer and Adobe Flash. Email remains the leading source of compromise (see phishing exercise at spring HEPiX); CERN remains a victim of targeted phishing campaigns. Anti-virus programs are often not very effective due to the delays involved for updating the signatures. Liviu then showed examples of very targeted fishing campaigns against CERN, which were followed immediately by rather generic ones hoping that they would remain under the radar. For small sites the situation is worse, as some convenient cluster management products have very poor security, for example HP Insight Cluster Management Utility, which recommends to disable SELINUX and open NFS shares to the whole world! Liviu then discussed Stagefright affecting Android 2.2 to 5.1. Google had swiftly released patches, but device manufacturers followed at very various paces. iOS has also been plagued by infected apps downloadable from the Apple store. Advanced persistent thread activity is increasing rapidly. Recommendations for users include installing an adblocker from trusted repositories, avoiding browsers and plugins with a long history of vulnerabilities, avoiding applications with a long history of vulnerabilities for opening attachments, using alternatives, avoiding clicking on e-mail links, always installing updates automatically, downloading software only from trusted sources, using different browsers for different purposes, using password manager with two-factor authentication, getting phones from manufacturers who are committed to provide monthly security updates, favouring unlocked phones over vendor-locked phones etc. The essence is to understand your adversaries, to treat security incidents as part of the normal operations etc.

In reply to a question, Liviu recommended an open-source password manager and offered documentation on how to set it up on different platforms.

Building a large-scale security operations' centre (Liviu Valsan / CERN)

Liviu explained that this concerns a unified platfom for ingestion, storage and analytics for multiple data access. For analytics, streaming, batch and interactive usage need to be supported in view of intrusion detection, statistical analysis and other use cases. He discussed in detail the architecture of the system. They are looking for scale-out rather than scale-up, and for a close integration with the rest of the CERN IT ecosystem, using commodity hardware as much as possible as well as OpenStack nodes and Puppet configuration. Cisco's openSOC platform is very close to what they

are establishing. A trusted online group has been set up for discussing SOC infrastructure; interested people can contact Liviu directly. Liviu then gave details of Bro, which the CERN team found very useful. They are using, with good experience, the Malware Information Sharing Platform, of which he showed some screenshots. Collective Intelligence Framework is a useful tool as well. Liviu then explained the algorithm they use for handling indicators of compromise. Concerning the future of academic security, it clearly is a global issue with global adversaries requiring global responses through international collaboration and thread intelligence. Beware that the targets are switching – services are no longer the main targets, but users and administrators are. Liviu finally pointed out that the CERN team is very open towards collaborating with other institutes.

Replying to a question, Liviu explained why Hadoop has been chosen over ES and Kibana. They collect about 500 GB of data per day.

Finishing the talk, he offered an analysis of the phishing mail that was sent on Saturday announcing free drinks, which was full of indications that it was not real. Still out of 106 e-mails sent, 49 recipients read the mail, 31 persons clicked on the link, only 18 read the mail and chose not to follow the link; 9 times the link was accessed without loading the image embedded in the mail, which was used to count the users reading the mail. Due to access restrictions from the BNL network to the phishing domain, a number of additional clicks may have gone undetected.

Wednesday 14 October 2015

Grids, clouds, virtualisation / GDB day

Introduction (Michel Jouvin / LAL)

Michel started by explaining how the idea of a joint HEPiX-GDB event came about, and what the Grid Deployment Board is – a body foreseen by the WLCG MoU, but not a real board with fixed membership, nor about deployment, and little about Grid these days. It is rather a forum for discussion between sites and experiments preparing decisions, mostly by consensus. In addition there are topical meetings typically run the day before known as pre-GDBs. GDB is open to all WLCG sites, Belle II sites are admitted as observers. Interested persons can register to a mailing list (ask Michel). Meetings are usually held every second Wednesday of each month (except for the next one, which will take place on 4 November). Meetings are broadcast by Vidyo; there are notes and action lists. Michel then mentioned the WLCG workshop, which will be held from 1 to 3 February 2016 in Lisbon. The workshop will be followed by a DPHEP workshop at the same location (common registration). Next events include a pre-GDB on DPM on 7 and 8 December, the HTCondor/ARC workshop in Europe (Barcelona) from 29 February to 4 March, and cloud and storage services (no date yet). He then mentioned the ARGUS collaboration, which has been established successfully, taking care of the issues observed at several sites including CERN. He finished with a list of forthcoming meetings.

CERN cloud status (Arne Wiebalck / CERN)

Arne reminded the audience about the basics: the cloud service is one of three major components of the AI project. OpenStack was put into operation in 2013 and followed the official releases with about 6 months' delay. The infrastructure is run in two data centres (Meyrin and Wigner) with multiple cells mapping to use cases. The top cell is running on several physical nodes in HA; each cell is managed by a child controller running on a VM. Currently there are 4'600 hypervisors in production, the majority of which is on CC 7; about 2'000 of them are located at Wigner, initially serving mostly compute use-cases, but now increasingly covering services as well. A total of 125'000 cores are in OpenStack, some 15'000 VMs are currently running; some 65'000 cores will be added early in 2016. Every 10 seconds a VM is being created or deleted. There are 2'000 images or snapshots and 1'500 volumes in CEPH and Netapp (the latter is used for Windows VMs). They now provide volume types covering various requirements concerning critical power, location, and I/O capabilities in terms of IOPS. They have put OpenStack Heat into operation, which allows for orchestrating OpenStack resources through templates; the service is integrated with the CERN infrastructure such as SSO and Puppet. Rally, a benchmarking and verification tool for OpenStack, has been put into operation, too. They have automated numerous procedures via Rundeck, which turns operational procedures into self-service jobs; this has allowed for a full integration with the service management tool Service-Now. The previous virtualisation infrastructure CVI, based on Hyper-V, is being phased out; 70% of the VMs are gone already, there are less than 1'000 VMs left. Machines can be migrated from CVI to OpenStack if really needed. They have successfully split a large cell into nine

smaller ones, without any impact on the users; he explained how the team achieved this. A lot of work was invested on performance; introducing write-back cache policy has reduced the I/O wait on the machine. Arne then explained that at some time, the team observed unexpected spreads of performance, which on closer inspection turned out to be a factor of two between equal hardware types; this was tracked down to wrong pinning of the two 16-core hypervisors to one and the same physical CPU, leaving the other CPU idle. They are now looking into containers; at present lxc is best supported in OpenStack Nova, but there is a lot more momentum behind Docker. They also started to look into OpenStack Magnum. Life-cycle management includes retiring hardware that has reached EOL. To transparently do this, live migration of VMs is necessary (relating to Carles Kishimoto's talk the day before).

Replying to a question, Arne clarified that write-back has the disadvantage that in case of a failing hypervisor, data is indeed lost or corrupted; that is the price to pay for the better performance.

Optimisations of the Compute Resources in the CERN Cloud Service (Arne Wiebalck / CERN)

Arne explained that the cloud and batch teams have observed larger than expected overhead of the virtualisation with a strong dependence on the VM size, ranging from 8 to more than 20 percent for 8-core and full-node VMs, respectively. At the same time, there were reports of dependencies of the VM performance with the EPT settings. A number of parameters were identified that impact the performance, including KSM, EPT, pinning, PAE etc. A first round of optimisation succeeded in reducing the full-node overhead to some 12...13%; with these settings, some hypervisors went into swapping, as the 2 GB reserved for the hypervisor was not sufficient due to 1.5 GB claimed by the kernel. This was traced down to the fact that KSM, though disabled, still reserved this amount of memory; the lesson to be learned from this is that when parameters change, the hypervisor better be rebooted. Even once this was understood, it was found that the optimisation effect on various hardware was different, and inefficiencies were still up to some 20% for full-node VMs. A cross-check with Linux VMs on Hyper-V resulted in 3.3%, excluding that the effect was due to general virtualisation limitations and pointing to the KVM configuration instead. Indeed, the difference is that Hyper-V makes the VM aware of the NUMA architecture of the hypervisor, which KVM does not; Hyper-V in addition pins vCPUs to physical NUMA nodes. For KVM, broader support of NUMA awareness will come with OpenStack Kilo. This was verified by manually configuring VMs with NUMA awareness, reducing the loss for full-node VMs to 3%. Moving some of these handcrafted nodes into batch production, it was confirmed that the performance loss is around 5%. As the performance depends on the KSM setting, the impact of KSM needs to be understood in more detail. On this occasion, system services were found to account for 1...2% of the machine's performance. A little later, the team was made aware of jobs running extremely slow, even though the CPU over wall-clock ratio on the VM looked perfectly fine. This was correlated with high system load on the hypervisor, which was eventually understood to be caused by the EPT off setting strongly preferred by the HS06 benchmarks. Arne explained some of the details, including why benchmark results may be better with EPT off. The team then wondered what the right page size is, as huge pages could be a way out. By default two huge-page sizes are supported, 1 GB and 2 MB. As 1 GB is very coarse-grained, the team went for 2 MB, and defined what is now referred to as the 'Kilo-1' configuration, resulting in 3...6% of overhead for full-node VMs. Arne then commented on why the issue of the slow jobs was not discovered during the testing phase.

Automated virtualisation performance framework (Sean Crosby / Melbourne and CERN)

Sean started by stating that estimating the performance of virtual machines is difficult, as users do not have access to the hosting hypervisor. The cloud team does not always have a complete view; indeed the performance loss for large VMs was discovered by the batch team. Hence better and more regular monitoring is required, which could be done in the context of the QA part of the services. Sean explained the main ideas of an automatic testing framework, triggering benchmark execution automatically and driving it through the process using Rundeck. Results are stored in Elasticsearch/Kibana. Multiple flavours need to be considered, including situations with multiple VMs on the same hypervisor, which need to be run synchronously. The framework was designed to support multiple benchmarks including HS06 and a number of fast benchmarks in use by the experiments. Further work includes to isolate one or several of each hardware type and start testing them and making sure that all known issues can be picked up by the benchmark.

An initial evaluation of Docker at RACF (Chris Hollowell / BNL)

Chris started by introducing the basics of Docker, which is fully supported by RHEL 7 and above, and exists for Windows as well; it has proved very popular since its introduction in 2013. With respect to full virtualisation, containers virtualise at the OS rather than the hardware level, which avoids the performance loss by the hardware machine emulation. At RACF, some software is developed for specific Linux distributions, which is a perfect problem for containers. Containers have been around for a long time, including chroot, CHOS, FreeBSD jails, Virtuozzo, OpenVZ etc., all of which came with one disadvantage or another. There is also lxc, which is focused on long-running containers, while Docker is stronger on running single applications in possibly short-lived containers; HTC batch jobs should fit well into this concept. Chris gave some examples of interacting with the Docker CLI. During their tests, they encountered an important ABI issue when running 32-bit software in an SL6 container on an SL7 host, which was traced back to inode numbers exceeding 2^32. Concerning security, Docker runs as root, but due to the fact it uses Unix sockets only local access is possible. It does not currently support user name spaces, but this support is expected in a forthcoming release. Chris then explained their benchmark work, which shows an 11% loss for virtualisation and a 1% gain (!) for Docker based on HS06. HTCondor supports the Docker universe as of 8.3.6; initial functional tests have succeeded, but they found they could not mount arbitrary directories (except the execution one) into Docker.

Benchmarking commercial cloud resources (Domenico Giordano / CERN)

Domenico stated that performance measurements and monitoring are vital for the efficient use of resources; intrinsic variations of the performance need to be dealt with. Activities started in 2011 within the HelixNebula partnership with the objective of supporting CERN's scientific computing programme. In March 2015, they ran a procurement targeting a single VO for simulation jobs; a second procurement has closed now, targeting simulations for multiple VOs. A third procurement, targeting the full chain for multiple VOs, is on-going; the market survey is currently open. Benchmarking is an essential component for defining the technical specifications, adjudication and re-mediation. A cloud commodity exchange based on a measurable value is required. They started the effort about one year ago looking for a benchmark based on open-source with a lightweight installation and reasonably fast running time; it was intended to be reproducible and have a functional relationship with experiment workloads. Given that in the March 2015 activity ATLAS were targeted, they had a close look at the ATLAS kit validation tool (see CHEP paper by Alessandro De Salvo and Franco Brasolin), consisting of bash and python scripts. For the workload, they chose Geant4 simulations of single muons, which runs within a few minutes; events 2...100 are used for the benchmarking in order to cut off the initialisation process. A total of 30'000 VMs has been benchmarked; results were found uniform within 15%, consistent over time, and well correlated with the time required for the real workload. In the current procurement, benchmarking is used to set limits for acceptable and tolerable performance, and to define services credits for underperforming VMs. They have now proceeded to define a cloud benchmark suite in order to compare and discriminate the most effective tests for the HEP use case; Domenico outlined the strategy and the architecture for benchmark running, data collection and analysis. The benchmarks they tried were the LHCb fast benchmark and ATLAS kit validation as well as the Phoronix benchmark, which is not HEP-specific. Some 250k benchmark suites have been executed, and various providers tested, including the CERN private OpenStack cloud. The effect of variable hypervisor load was studied at CERN with a single single-core VM for the benchmarking. They have also investigated the reproducibility of the benchmark across several data centres of the same provider.

A comparison of performance between KVM and Docker (Wakaru Katase / KEK)

Katase said that KEK is moving to a private cloud (OpenStack) for the worker nodes, and explained the architecture and the workflow. They used Rally, the principles of which he explained, to measure the performance of KVM virtual machines and Docker containers. He gave detailed information about the environment used for comparing single-core machines. He presented benchmark results for server creation and deletion (both single machine and 32 machines at a time). While with few concurrent requests, KVM and Docker take about the same time, at high concurrency KVM is 20% faster than Docker. Taking all steps together, a Docker container is ready for networking in less time than KVM, which remains true when the time to create the server is added. For server deletion, the difference is not statistically significant. Most CPU and I/O operations show some advantages of Docker over KVM, while a physical machine is superior or equal in all cases. The study shows that Docker, while not adequately supported in OpenStack yet, is a very

interesting option, but more investigations are needed around security, stability and more comprehensive benchmarking.

Open science Grid – Physics to campus researchers (Chander Sehgal / FNAL)

Chander explained that his talk was about how other sciences could benefit from developments primarily made for HEP. He then reminded the audience that OSG is mandated to foster open distributed high-throughput computing in the U.S. in a partnership between resource providers and stakeholders. The project provides middleware and R&D effort. Over 120 sites contribute; the VO model is used for trust relationships. The project receives funding by the DoE Office of Science and the NSF. The ultimate goal is to extend resource sharing to the largest possible extent in order to maximally profit from available resources. They are approaching 1 billion CPU hours per year, having delivered 91 million hours in the past 30 days. Over the last 12 months 1 billion data transfers accounted for a volume of 185 PB. There has been a very steady growth since 2005. ATLAS and CMS account for about 2/3 of the OSG usage; an artificial VO without resources, intended to benefit from opportunistic resources only, now takes about 10% of the resources. Together with regular VOs such as life sciences and other science, non-HEP consumes about 20% of the OSG hours. The opportunistic resources amount to 140 million hours per year, with a large diversity of users, non-LHC HEP being responsible of some 40% only. It was key to remove the requirement on users to have a Grid certificate, which has lowered the threshold for using OSG very significantly. Chander explained how this works in detail, and that several thousands of researchers have obtained access to OSG in total during last year. Another important element is the introduction of 'facilitators', individuals at sites who can help people find the right solution for their computing problem. He listed five stories, among them Martin Purschke's design studies for sPHENIX, that were enabled on OSG just by using cycles that would otherwise have remained idle. Chander then mentioned storage and data movement, and showed how easy it is for sites to join their clusters to OSG.

INDIGO Datacloud – improving IaaS resources to accommodate scientific applications (Andrea Chierici / CNAF)

Andrea explained that the INDIGO project is embedded into H2020; it was approved in January 2015. 26 partners from 11 EU countries have joined forces. Key challenges are to collect and consolidate evolving user requests, to create a new sustainable cloud competence in Europe for PaaS (similar to OpenStack etc. for IaaS), and to address technology gaps (vendor or technology lock-in, naive resource allocation strategies, inflexible ways of distributing applications, lack of access to specialised hardware such as IB). The project is organised in six work packages, of which one is the focus of the presentation: to extend features of IaaS middleware to support reliable management of a performing virtual computing infrastructure. Andrea explained the sub-structure of the work packages around support for containers, improvements of existing cloud schedulers, integration of container execution into batch systems, and provisioning of a local IaaS site orchestration. They have made a number of architectural choices including to base the project on standards (OCCI, …), which means limited support for native interfaces, and have selected OpenStack and OpenNebula as cloud middleware frameworks and Docker as initial target for container support. Container execution support is to include transparent execution in batch systems. In OpenStack Nova, there is an existing Docker driver; they are evaluating Magnum. In OpenNebula, only lxc is supported; the project will contribute a Docker driver and develop extensions wherever needed for container support. For the batch system integration, there is the choice between integrating directly into the batch system, or less intrusively adding the integration in user space. Existing schedulers are too simple-minded, more functionality towards fair-sharing will be added. The project is called Synergy, of which Andrea gave some details. Spot and pre-emptible instances will be supported as well. The partition director is intended to flexibly administer the sharing of resources between batch and cloud. They have mostly done the design phase and have started implementing the Docker interface for ONE. Synergy is being implemented as an external OpenStack project; discussions about orchestration are being held with the OpenStack community.

First experience with Mesos at RAL (Andrew Lahiff / RAL)

Andrew explained that resources at RAL can be divided into three groups: storage (Castor, CEPH), compute (HTCondor, OpenNebula), and services (Hyper-V, bare metal). This situation gives rise to problems: manual intervention is needed in case of failure, the environment is very static, and resource utilisation is sub-optimal. One approach is to

not tie services to specific hosts, and to manage services using a scheduler. This would address the identified shortfalls. Apache Mesos implements these concepts, removing the view of an individual machine, implementing the payload as containers. A number of frameworks are readily supported in Mesos, others can easily be added. Marathon is the component that looks after scheduling (potentially long-running) services. This concept requires a service for service discovery, as static configurations do not make sense any longer; Mesos comes with such a component (Consul) that provides the service via DNS and configuration files. It is possible to suppress unhealthy services automatically. For external access to services, load balancers such as HAProxy are supported, as is monitoring for resource usage, application-specific metrics etc. Logging is included as well. At RAL they have set up 3 head nodes with ZooKeeper, mesos-master, Marathon and Consul server, and 16 Mesos slaves. Example use cases show how Squids were relocated and replaced in case of disappearance or malfunction, and how a top BDII cluster that was upgraded in a rolling fashion. FTS3 can be run in the same way, too. The next steps are to look more closely at security and monitoring, then to move non-critical jobs and services in order to run them in parallel for a while.

Replying to a question, Andrew clarified that while there is probably some overhead of running servers in containers, the complexity is probably sufficiently small not to have any major impact. Andrew tried already to provide HTCondor services via Mesos, which worked fine.

HTCondor-CE: Managing the Grid with HTCondor (Brian Lin / Wisconsin U)

Brian started by explaining the main functionality of a compute element (CE), mainly to provide for a job gateway and a submission engine into the local batch system. The HTCondor CE is just a special configuration of HTCondor, it accepts jobs via HTCondor C, supports various authentication mechanisms and batch systems etc. The anatomy shows the usual components of a HTCondor system; for non-HTCondor batch systems, a GridManager and a BLAHP component are added. The bulk of the configuration goes into the routing of jobs. Several job routes are possible. Brian showed example configurations for job routes allowing for filtering according to submitter or to machine features. Brian then gave hints on how to debug a configuration of HTCondor CE that does not work as expected, including increasing the log level and activating troubleshooting tools. Reasons why to switch to HTCondor CE include reducing software providers if the site runs HTCondor already; even if not, it has outstanding scalability, flexible routing policy, job traceability and fewer open ports. However there are downsides: the declarative ClassAd language can be confusing, and the BLAHP layer is not feature-complete. Since last year, in OSG the number of installations has grown from 10...15 to 49. Each site CE advertises itself and its resources to the central HTCondor collector in OSG, which is one step of replacing the BDII. Then Brian mentioned the roll-out at a non-OSG site… CERN! He gave some details of the implementation. New features include the support for Docker containers (see his untested example), BLAHP improvements, EL 7 support etc. Future work will be around improvements to job router configuration (adding defaults) as well as improvements to BLAHP, which will be fully integrated into the HTCondor source. They also plan to expand the central collector for supporting long-lived resource ads, and an adapter into the ATLAS global information system (AGIS). Among the people thanked was Iain Steers from CERN.

Running ATLAS, CMS, ALICE workloads on the NERSC Cray XC30 (James Botts / LBNL)

James Botts started by putting into perspective the PDSF capacity with other, much larger systems at NERSC. Their objective was to enable running standard HEP workloads on the specialised machines at NERSC. He described their current supercomputer setup with currently three powerful Cray systems with a GPFS storage backend. One significant difficulty they encountered is CVMFS, as the worker nodes do not show a typical Linux environment and have no local disk. They had to develop a new approach to get CVMFS to the node – local installations at all nodes were inconceivable, as were NFS nodes or rsync to a shared file system. They hence considered containers (Docker); NERSC is enabling Docker-like container technology through a new mechanism called shifter. All external connectivity must be implemented at the container image level, the worker node system will not provide anything. He presented performance measurements showing that using shifter is indeed practical. In order to get CVMFS onto a node, it is installed on a standard Linux node, an empty ext4 file system is created, then a complete snapshot of CVMFS is made and included into the image; before doing so, a de-duplication is applied, and the image is compressed, resulting in a file of a little more than 300 GB, which can be re-done once per day. The mechanism has been successfully tested with simulations of all three supported LHC experiments, a full production chain test is in preparation. Shifter is now released as

an open-source tool; collaboration with other sites is more than welcome. Cray may even be interested in making it part of their software distribution.

Computing resource virtualisation at IHEP (Qiulan Huang / IHEP)

Qiulan started by mentioning the big HEP projects for the lab in the next seven years with yearly data volumes of more than 1 PB each. To have the required flexibility, they established a private IaaS platform based on OpenStack (Icehouse) in November 2014. It allows users to create and destroy VMs on demand and to virtualise the computing cluster running HTCondor and PBS. Similarly to CERN, every user has a small quota (3 cores, 15 GB); they support testing machines with full root access, and a UI node with AFS authentication etc. An extensive list of OS is supported. A non-routed address is assigned automatically, a routed address needs the approval by the administrator. They currently run 24 hypervisors with 464 cores for more than 1'000 registered users. Qiulan then showed the overall architecture of their internal cloud featuring Puppet, Ganglia, nagios, CEPH, Dirac etc., and gave details about the IP address allocation. The network design uses Neutron and OVS in VLAN mode, exploiting just one NIC. Unified authentication is handled via Keystone and LDAP based on the IHEP mail account and password. On the VMs, non-privileged users are not allowed to change network parameters. She then explained measures they have taken to encourage and foster sharing of resources for a better overall utilisation. She then explained the concept of a virtualised computing cluster, which is a 'batch system out of the box' based on HTCondor or PBS. Initial tests have been successful, hundreds of test jobs have been run. They now need more testing at scale, and to roll the service out to the complete list of experiments to be supported. There is collaboration with remote sites to form distributed clusters, of which Qiulan showed successful statistics. They are now planning for the upgrade to Kilo and for optimising the monitoring.

Storage and file systems

CVMFS deployment status and trends (Alberto Rodriguez Peon / CERN)

Alberto started by explaining that CVMFS is a network file system for reliable and scalable software distribution. It uses aggressive caching and on-demand serving. The basic deployment comprises a Stratum 0 as the single source of new data, which gets replicated to one or several Stratum 1 instances. The clients, equipped with a local cache, pull changes from the Stratum 1, perhaps via some HTTP caches. The Stratum 0 at CERN comprises 21 repositories, most of them very active, some of them rather quiet, totalling to some 10 TB of data in 44 million files. The Stratum 0s used to share a single monolithic Netapp filer mirrored via snapshots and backed up to tape. However due to the obsolescence and the high costs of the Netapp filers, it was decided to move to CEPH-based storage, which meant rethinking the deployment model in view of recovering from failure scenarios, easing the roll back, and avoiding wasting resources. They decided to use ZFS, both a file system and a volume manager, supporting file system snapshots, live data integrity checks, live migration and resizing of volumes, easy data replication, and support for several pools per server. They now run an independent CEPH volume per Stratum 0, and do not mirror the files any more, but rather use snapshots at file system level both periodically and on demand; rolling back is indeed very easy. ZFS also includes consistency checking and repair, which can again be run periodically as well as on demand. For the replication they use ZREP, an open-source solution for ZFS replication based on ZFS send and receive; it replicates both content and snapshots and supports locking and fail-over features. However if a replication fails, manual intervention is needed. Alberto then showed how to recover from a server failure. Recovery from disk failures is also very straightforward. The team is reasonably happy with the solution, and is looking into spreading disks over data centre, further automating the procedures, dumping data to tape, and using RAID as a back-end for ZFS.

Answering to a question, Alberto said that the impact of a scrub has been unmeasurably small.

Home directories at CERN (Alberto Pace / CERN)

Alberto started with a summary of the systems currently in use at CERN. For physics data, Castor and EOS are used; user data are stored on AFS and DFS. In addition there are a number of stores for corporate documents such as CDS, EDMS, Indico, Alfresco etc. The two major systems for physics have different roles. Castor is highly reliable and comes at low cost for archives and physics data preservation; it offers high throughput at large latencies. On the other hand, EOS has been designed for high scalability and tunable performance and reliability at 1 kHz access rate. AFS and DFS

are provided for user and project files; they store a few PB, but the number of files is very large. The related requirements are very numerous and partly conflicting. DFS and AFS offer native online access, the latter one global access as well. New opportunities arise with EOS that supports standard protocols such as http and webdav next to xroot. This has been used to provide a backend to ownCloud, implementing the ownCloud extensions directly into EOS. This allows for sharing and bulk download. They think that for user files, access must be provided to the whole storage repository; direct mounts would be useful (which is possible with EOS via FUSE or Webdav on Linux and Mac); programmatic access can be provided via xrootd. The sync functionality covers all desktop and mobile platforms in use, allowing for selectively syncing part of the tree. Sharing is possible selectively as well via ACLs implemented in EOS, which are supported for CERN accounts, e-groups and roles; sharing is possible via a Web interface as well, supporting sharing to non-registered identities via keys that can be given to people concerned. Things are similar for project space except for the fact that there may be concurrent access by multiple users. Then Alberto talked about the physics analysis and showed typical flows, pointing out that analysis via the Web browser is emerging as a new possibility. Hence they are considering a model where a Web client is the only access for the user; this is a collaboration with PH-SFT. CERNbox and EOS have a certain overlap with AFS, hence the AFS use cases are currently being reviewed

On a question by the audience, Alberto clarified that FUSE mounts can support both Webdav and http.

CEPH-based storage systems at RACF (Alexandr Zaytsev / BNL)

Alexandr started giving a short overview of the production releases of CEPH with their respective functional changes. At RACF they started testing at the beginning of 2013 and became more and more confident that it was a reliable and very scalable solution. He mentioned a number of infrastructure challenges and how they responded to them. Their building blocks include Dell PowerEdge rack-mount servers, Sun Thors and Nexsan SATABeast boxes, the latter ones retired from ATLAS dCache service. They have deployed two very similar clusters with usable capacities of 0.6 PB and 0.4 PB, respectively, details of which he showed. They are in the process of unifying the underlying fabric, still maintaining the separation between the two clusters for functional reasons. He then turned to the production experience; the main use case has been the ATLAS event service, which has used only 10% of the performance the old cluster could have delivered. They observed a performance issue with the higher levels built on top of the RadosGW interface. The cluster was also used as data backend for the opportunistic MC generation for sPHENIX design studies described by Martin Purschke earlier during the week. Longer term, they will move to production and support without any major change of architecture, and are looking for (more) collaborations with other groups using CEPH.

CEPH object storage at RAL (George Vasilakakos / RAL)

George explained that their installation, using Hammer, consists of 47 storage nodes with 5'200 TB of raw storage space and three controllers running on bare metal (their experience with VMs was not encouraging). The gateways provide S3, xrootd, and GridFTP interfaces; they only use CEPH as an object store, not as a file system. Kerberos and X.509 support have been added. For monitoring, Nagios checks are in place; they plan to use ELK for logging. They have deployed InfluxDB and Grafana because of the better possibilities to monitor distributed objects. They are currently working on an S3/Swift object store, for which the ATLAS event service will probably be the first user, on xrootd access for the Tier-1 batch farm (xrootd 4.2.0 adds support for CEPH backends), on GridFTP (on pools interoperable with xrootd ones), and on Castor tape buffers (supported as of Castor 2.1.16). He then mentioned erasure coding (EC) for more efficient use of disk space at the cost of CPU time and extra concurrency; they will use (at least initially) 16 + 3, making the probability of data loss in case of four concurrent disk drive failures very small. However EC is not without challenges, as it introduces higher overall latencies and higher recovery loads. He then explained details of the data placement policy, pointing out that with their current configuration, some 40% of their nodes will be involved in each operation, introducing a lot of inter-node communication. They are currently buying network equipment for the cluster backend, and connected the larger 21 nodes (34 OSDs each). EC 16 + 3 did not work, they went to 16 + 2, which is almost pathological, as 18 out of 21 nodes were involved in every operation. After reconfiguration with fewer nodes, things looked much more reasonably. They still plan to continue testing EC, testing the gateways, and establishing a baseline for individual node performance for identifying bottlenecks.

Scientific data storage at FNAL (Gerard Bernabeu / FNAL)

Gerard explained the current offerings referring to the various experiments supported by FNAL. They are using a Bluearc (Hitachi) NAS system that is pretty expensive; they use it for interactive (and still some batch) applications for CMS and FIFE, is is providing 3 PB of usable capacity in total. They are in the process of migrating the data to dCache except for user home directories; the system is not mounted on the local worker nodes. Their Lustre system is mostly used for HPC computing resources; the system provides 1.4 PB and stores 156M files. Access is provided via Posix-like Lustre mounts as well as GridFTP. They are currently upgrading to the latest Lustre version, and move to Lustre on top of ZFS on SL 6. Their EOS system is used for various CMS use cases; it provides 7.5 PB and currently stores 19.5 million files. Access is provided by FUSE, xrootd and SRM/GridFTP. The underlying hardware is identical to the one used for dCache. The system will take some of the current use cases of the Bluearc system; Tier-1 operations will move out of EOS to dCache. dCache and Enstore are used as HSM front-end as well as stand-alone disk-only system for CMS, D0 and CDF. The disk-only part provides 12. 6 PB accessed via xrootd, SRM/GridFTP, dcap, and NFSv4. They will upgrade to a newer version (from 2.2 to 2.13), move CMS merge job output from EOS to dCache, concentrate access doors to a few systems, move the Chimera data base to SSD disks, and open up the storage building block architecture that is currently limited to servers with FC SAN. dCache with tape backend consists of 4 PB of disk cache and 85 PB of tape space with similar access methods as for the dCache disk-only pool. They will upgrade all CMS pools to 10 GbE on new hardware and optimise the data transfer rate. Finally he presented the Enstore small file aggregator, which combines small files on tape migration automatically, reversing the operation on recall.

Accelerating high-performance cluster computing through the reduction of file system latency (David Fellinger / DDN)

David started by observing that traditionally, for large clusters, the focus has been on the cluster itself, moving to lower-latency interconnects, more efficient message passing structures, and higher-performance processors and GPUs. Also, research and studies have been made on true parallel processing. Today's challenge is however the bulk I/O latency – waiting time during which the processors are idle. Other industry approaches are to move data via distributed hash tables, removing the classical distinction between server and client. The question is hence whether a parallel file system such as Lustre is really parallel; some metadata servers usually do take a central role, constituting obstacles for scalability, in particular in view of the fact that metadata blocks are very small. DDN are now testing a layered approach with a cluster of servers connected to the MDS storage that can more easily be scaled; at the same time they suggest to replace the Lustre code on the client. This middle layer is entirely transparent to both sides; the applications just see a parallel file system, such as if it was a pure Lustre system. David then presented benchmarks that show bandwidths that are impressive and very stable over varying I/O size as compared with native Lustre. An implementation with 8 Supermicro servers has been tested at TACC, again with excellent and consistently high results. David remarked that this is a lesson that could have been learned so much earlier from the Internet P2P guys… Finally he invited the audience to get in touch with DDN for testing applications.

Thursday 15 October 2015

Machine learning and failure prediction in hard disk drives (Amit Chattopadhyay / Western Digital Inc.)

Amit started by recalling what machine learning is: algorithms that improve their performance at some task with experience. He then showed a typical algorithm and listed a few examples including recommendation engines, spam filters and self-driving cars, the latter being a far more complex case. There are different kinds of machine learning, supervised and unsupervised learning as well as discrete and continuous one. He then mentioned typical classification problems, not all of which are linear; the functional approximation is more or less difficult depending on the case at hand. The question is also whether the approach can be generalised; for example the order of a polynomial to be fitted must be right. The ROC curve describes the balance between false positives and true negatives. This is related with hard drives that are tested (say a sample of 300 out of a production of 50'000) with an ORT (ongoing reliability test). Traditionally (based on tribal knowledge and/or engineering insight) the result can be interpreted by explicit limits; test the 300 drives, if they pass, sell the 50'000… but with a delay of the test duration. With machine learning on the other hand, things can be accelerated: do a 50-50 randomised split into training and test; each drive goes through detailed

characterisation before getting to ORT; on the training population, use a classifier to choose the top features describing the drive reliability, build a logistic regression model, calculate the failure probability of each drive, and generate a 'health hierarchy'. A cross-check is the sensitivity to key features; it shows that the learning is not perfect, but adequate, hence it can be applied to the test population. Back to the ROC curve, the area under the curve is 85...87%, hence in 8 out of 10 cases the business call is correct. Amit concluded that this analytical technique can provide a robust framework for fleet management.

QoS and DLC in IaaS in INDIGO DataCloud (Patrick Fuhrmann / DESY)

Patrick reminded about the basics of the INDIGO project, a H2020 project approved in January 2015. The project started in April with 26 European partners from 11 countries with a funding of 11M euros. The objective is to develop an open-source platform for computing and data deployable on public and private cloud infrastructures. He then showed the WP structure of the project (cf. Andrea's talk). He then mentioned some details about WP4 about virtualising computing resources and virtualising storage resources, the latter covering QoS and data life cycle, access to data by metadata rather than by name space, dual-access to data (object store and POSIX file name space), and identity harmonisation for storage. Virtualised network resources are covered as well. Patrick then focused on the storage aspects and in particular on QoS and data life cycle. Problems to be addressed include that there is no common way to describe QoS and DLC, nor to negotiate QoS with the storage endpoint; common definitions for QoS would be very convenient in general and are compulsory for the targeted PaaS layer. They are defining a common vocabulary for QoS involving standardisation bodies (RDA, OGF). Independently there are considerations in WLCG to provide a platform layer, partially replacing common parts in experiment frameworks; SLAC wants to introduce a model for pay-as-you-go, which also requires defined storage classes and QoS. However this is not without issues. Is there a complete set of properties? Is access latency and retention policy enough (as WLCG appears to assume)? There are models around with 200 properties, which probably goes far beyond what our community needs. In addition the values of the properties, for example QoS, are not clear at all; the picture is blurred by volume dependencies of some of the properties. They are hence going for canonical properties, for which a discovery and match service is required; this service would need to provide a GUI and a REST API. Federated systems also need to be considered; they come with additional properties that need to be taken care of. Additional problems include how the client provides the storage class to the storage system; the system only provides the class, it doesn't promise the space. Data life cycle is basically just an aspect of QoS. Patrick finished with a review of the current status of the work.

dCache storage system at BNL (Zhenping Liu / BNL)

Zhenping reminded about dCache and the fact that it has been in use at BNL since 2004. They heavily use the replica manager for disk-only data, which they separate from tape data. Most protocols supported by dCache are in use at BNL. HPSS is used as the tape backend for the tape data area. The storage is not accessible from the WAN, as it goes through 18 GridFTP/xrootd door hosts. The head servers are all running RHEL 6.6. She then gave details about the pool hosts running RHEL 6 with XFS or Solaris 10 with ZFS, the latter being decommissioned in favour of Linux. The total disk space amounts to 14.2 PB. Only the SRM and GridFTP servers are outside the BNL firewall for ATLAS data transfers requiring high performance; all other dCache servers are within the firewall. The instance is integrated into FAX (federated ATLAS storage system using xrootd). A FAX reverse proxy service allows for accessing BNL data behind the firewall by clients located outside; there is also a FAX forward proxy allowing the worker nodes behind the BNL firewall to access data (directly from BNL or via the proxy from other sites). They use Puppet, Git, Cobbler and GLPI to manage the installation and configuration. The instance is monitored with Nagios (soon to move to Icinga2), Ganglia, a bunch of maintenance scripts, ELK, Panda job pages, and a DDM dashboard. Issues they encountered recently include storage pool servers running out of memory (they will increase to 256 GB), replica manager performance issues addressed by separating the replica DB to a separate host, and high load upon SRM restart due to a busy pinManager, which was addressed by increasing the memory for the PinManager. They plan to use different storage tiers (in different rooms) for primary and secondary copies of the disk files, which will also help the availability for read requests.

Space usage monitoring for distributed heterogeneous data storage systems (Natalia Ratnikova / FNAL)

Natalia started with an overview of her activities in the HEP area, mostly at KIT and FNAL, where she has looked after data storage administration and data transfers. She then explained the inreasing storage resources for CMS due to the Run 2 requirements (in particular pile-up and the detector studies for HL-LHC); at the same time the computing model has evolved, separating tape and disk storage at the Tier 1, AAA xrootd-driven data federations, dynamic data management, new data types etc. CMS are organising the data in a global name space (rather than space tokens as for ATLAS and LHCb) addressed by a logical file name such as /store/data, /store/mc etc.; data are accessed by physical file names, the translation being done by a site-provided catalogue. While information about data maintained by PhEDEx is readily available, information about other files is only available from dumps of the storage system itself. Monitoring is required for efficient space utilisation, fair-share between users, resource planning etc. Natalia then mentioned a number of related activities and initiatives, and presented the general architecture and work flow of the monitoring solution, the components of which are the site information providers, the site collector, and the central information store. Issues encountered during the deployment include questions from sites about the motivation of the project, authentication problems uploading the the information to the central data service, and the long time some storage systems take to produce the required information. In addition, security and privacy issues were raised.

IT facilities and business continuity

The HP IT data center transformation journey (Dave Rotheroe / HP)

The speaker started by reminding the audience that data centres exist because of a concrete business case. From 2005 to 2009 HP reduced the number of data centres from 85 scattered world-wide to six all in the US. One of the driving forces has been Moore's law, which has applied for forty years and still appears to apply. Data are doubling every year, which is scary. Vendors are continuing to offer more powerful servers; industry-standard servers are converging towards very similar layouts whoever the manufacturer. In addition there are application-specific servers emerging that are designed for a specific workload. For storage the cost per terabyte continues to drop rapidly; SSDs are now viable for data centre use. Networking speeds evolve to more than is needed. Software is also catching up; clouds are real now and allow for better utilisation of the existing resources. However most companies run a mix of ancient to newest technology, including discrete rack-mount servers, old SAN and DAS, some virtual farms and blades etc. - and hundreds, if not thousands, of applications which can be run, evolved, retired etc. - with very different impacts on the power consumption. Different refresh and growth rates have a very different impact on the power consumption of the data centres assuming no external resources. However enterprises now target 50% SaaS replacement, 40% moving to the cloud, and only 10% refreshed legacy infrastructure. With 50% SaaS, the power consumption figures are quite different. This all means that most enterprises will use less data centre capacity and less IT equipment; co-location, Telecom, and managed service firms will continue to grow; SaaS providers will become pervasive; public and managed cloud providers will grow significantly. He concluded by stating that since 2009, HP has closed two more data centres, now only running four.

High-performance computing data centre proposal (Imran Latif / BNL)

Imran, after introducing himself, listed a number of facts of the existing data centre BCF – a building of the 1960s, with an estimated PUE of 1.9, the layout of which he explained; there are multiple rooms for different purposes. The existing raised floor is only 12 feet high and much congested; the evolution over time of the physically fragmented floor plan contributes to inefficiencies; existing cooling systems are obsolete and unreliable; back-up chilled water does not exist everywhere; “chaos cooling” practices are highly inefficient. The total facility power is inadequate for modern computing requirements, headroom exists only in terms of non-UPS power. Physical space is limited as well; new programmes need to be turned away. In addition there is no storage space for receiving and shipping equipment. The current situation is serious in view of upcoming challenges for the lab, which is taken very seriously by the lab management. It has hence been proposed to renovate the previous light-source building with cooling (including back-up) for the initial (4...5 years) and longer-term (15 years +) future, new electrical infrastructure, architectural modifications and life safety. He showed the proposed layout of the facility, which includes storage space and a user analysis

area; there is room for future expansion. The new facility will use free cooling wherever possible (which is not all year round).

Asset management in the CERN data centre (Eric Bonfillou / CERN)

Eric started by explaining that since 1997 Infor EAM has been used at CERN for tangible assets such as racks, PDUs, servers, disk arrays, rail kits, SAS cables etc. Not covered are software, licences, activation keys etc. Tangible assets are characterised by procurement attributes (admin references, commissioning date, warranty duration, status) and financial attributes (starting value, expected lifetime, depreciation method, depreciation value, book value, residual value). About 30 classes are defined currently, most of them being foreseen to handle stocks of spare parts. Each class provides sets of specific attributes. The class for PDUs is particular in that it includes the in-feed power line linking with a different group (electricity support). Assets can be located at Budapest (Wigner) and Meyrin (CC), which is well supported by Infor EAM. Sites, buildings and rooms are not assets, but rather containers for them. The top-level assets are racks, of which 1'300 are registered. The position in the rack is characterised by the 'U position' from 1 to 42...47 depending on the rack type. These positions are translated into functional positions in EAM, which must be unique. EAM structures assets into a tree, which is quite flexible – twin or quad systems for example are not a problem. No asset can ever be deleted from EAM. The inventory goes back to 2010, older equipment being obsolete and retired; this corresponds to some 100 orders. The workflow for getting the information into EAM appeared simple, but there were a number of issues including incorrect or missing serial numbers on the assets themselves or in the data sources. and erroneous locations for a few assets. In addition, there have been cases where the bar code on the label did not match the human-readable printout on the same sticker. Eric mentioned briefly the process to put servers into operation. Hence putting all these data into EAM was a major effort and required multiple verification steps. In 2016 they will handle the stock of spare parts via EAM, integrate with CERN's Geographic Information System, and perhaps streamline the retirement process.

Energy efficiency upgrades at PIC (Pepe Flix / PIC)

Pepe introduced PIC, the largest Grid centre in Spain close to Barcelona, which hosts the Spanish Tier-1 of the WLCG as well as a Tier-2. The equipment is located in the main IT room on the University campus; the university has paid the electricity bill so far. However the cooling, power, and UPS elements are aging, and spare parts are increasingly difficult to obtain. Since 2003 a number of measures have been taken in order to cater for more equipment and to improve the PUE. Works in 2014 and 2015 focused on modifying the existing IT rooms and introducing free-cooling techniques, profiting from the cold winds falling from the mountains at night. Introducing free cooling implied replacing the IT room CRAHs , integrating a chilled water backup system, implementing the separation of hot and cold aisles with hot aisle containment, increasing the intake temperature according to the ASHRAE recommendations etc. Hot air containment meant installing a false ceiling for trapping the hot air. Pepe showed photographs from the installation of the free-cooling unit. The work was completed in September 2014; first measurements indicate that a PUE of around 1.45 as a yearly average is achievable. Electricity cost savings are estimated to amount to some 100k euros per year. They also replaced the old UPS system with estimated losses of some 15% by a much more efficient dual-UPS solution. They are now introducing oil immersion techniques (GCR CarnoJet) with 4 x 46U, each one capable of dissipating up to 45 kW, of which he showed some details. They will start with 10 servers equipped with SSDs (standard hard drives do not work); the servers are pizzabox-like and fanless. They expect to achieve a PUE of 1.05...1.1.

Energy Services Performance Contracting (Michael Ross / HP)

The speaker, apparently unaware that the vast majority of the HEPiX participants came from non-U.S. sites or U.S. universities, talked about contracts federal U.S. agencies such as DoE have concluded, with the aim of ensuring efficient use of U.S. federal data centres and reducing their energy footprint. The basic idea is that infrastructure measures to reduce the energy footprint are paid for by the energy savings achieved over up to 25 years. Once paid, the energy savings benefit entirely the organisation running the data centre.

Basic IT services

Document oriented database infrastructure for monitoring HEP data systems applications (Carlos Fernando Gamboa / BNL)

Carlos explained the usual data flow between Logstash, Elasticsearch and Kibana. Elasticsearch is a document-oriented, horizontally scalable data base with a mapping similar to a schema definition in SQL. Kibana is an analytics and visualisation platform designed to work with Elasticsearch; it supports dashboards showing evolutions over time. He has used these tools for monitoring selected storage services including SRM and GridFTP, S3, the BNL dCache instances etc. As an example he showed the dCache billing monitoring dashboard and the AWS SE Bestman monitoring dashboard. The automatic refreshment every five minutes has created peaks every five minutes of the CPU consumption of the Kibana server. This imposes limits on the scalability of the monitoring. He described in detail the setup of the test ELK server in terms of software and hardware.

ELK, RabbitMQ, Collectd and Freeboard in the wild at NERSC (Cary Whitney / LBNL)

Cary started by saying that this is a report about work done since about a year; initially it was targeted at facilities, but now covering services in the larger sense. The old system was MySQL based and was hopelessly overloaded, supporting a wide variety of protocols and metrics. From the nodes, a collection of monitoring tools reported into the system. The move to the CRT building is the occasion to clean things up; Cary mentioned the most important design parameters of the new building. There are 10 temperature sensors per rack for a total of 1'500 sensors already installed, which could grow up to 10'000 sensors. The nodes send system logs, performance data, Cray environmentals, and external parameters. The goal was to instrument and collect everything without losing any potentially valuable information. They looked at what “big players” do. System administrators, scientific researchers, researchers and management have all very different interests. The hardware of their solution consists of 32 (Supermicro fat twin) nodes. Cary explained the setup of the sensor network that makes use of Power-over-Ethernet in order to reduce the cabling. He then discussed the flow diagram, in the core of which there is RabbitMQ. He showed examples of Freeboard, and explained the role of collectd, a UDP-based collector on the nodes. Data more than one day old are being migrated away from the fast SSDs. Future work will focus on dashboards, stream processing, data cleanup and anonymiser. Finally Cary suggested a workshop or BoF session around HEPiX in Berlin.

CERN monitoring update (Miguel Santos / CERN)

Miguel started by reminding the audience about the design goals of the system. Rather than the 10'000 systems covered with the old system, they are now handling 19'000 systems collecting more metrics per system and log files as well. There are alerts, archive, display, and streaming. The architecture is the industry-standard lambda architecture extended by an additional leg for the alerts. A large set of producers is supported; transport is using Flume (for the lambda architecture) and ActiveMQ (for the alerts). Flume feeds into ESK as well as HDFS in preparation for analytics. Miguel discussed the paradoxon of the data lake – everybody would like to get access to everybody else's data without disclosing their own. He then showed examples of dashboards they have developed; users can create their own and store them for later usage. They run a central public ES instance for monitoring and public logs; for other use cases the team provides assistance to set up additional ES clusters. They are working on providing ES instances via Heat, avoiding a lot of manual work. To summarize, Miguel emphasised the central role of Flume in the architecture. He then discussed the stream processing prototype (currently in closed beta) which is based on Kafka and Apache Spark; he showed a few examples. They are considering detecting anomalies via machine learning, which is not an easy task at all. Finally he showed Jupyter (previously known as ipython). For the future they will follow upstream upgrades, finish the streaming prototypes, and test a number of other tools (for example collectd and cAdvisor, InfluxDB).

Update on configuration management at CERN (Alberto Rodriguez Peon / CERN)

Alberto presented some statistics comparing with the situation with the Oxford meeting (17% more hosts, 68% less catalog compilation time). The latter reduction was due to the upgrade of the puppet masters to CC7, implying Ruby 2.0 which is not backward-compatible with Ruby 1.8. The upgrade also caused a reduction of CPU usage on the Puppet masters from 80% to about 25%, which means that some masters can be de-commissioned. The next steps will be moving to Puppet-server (re-implementation in Clojure and JRuby), which is supposedly much faster, and considering

Puppet 4, which is not backward-compatible. Hiera GPG is being phased out, as the encryption is not very safe – the Puppet masters evaluate the encrypted quantities. CERN has replaced this mechanism by one that decrypts only on the clients concerned. A side effect is a further reduction in catalogue compilation time. The upgrade of the Apache module from 0.4.0 to 1.2.0 was difficult, as the newer one is not backward-compatible, and the module was very widely used at CERN (385 servers in 40 different services). The issue was handled by creating dedicated environments; Alberto explained in detail how this was done. He then discussed hostgroup ownership, which is required (and required to be the same) in Foreman, git and all other parts of the infrastructure. The source of truth has hence moved to Riak, a key-value database; Foreman and git are regularly being refreshed with the updates from Riak. For the package inventory that serves for ensuring consistent setup across a hostgroup, they went for their own solution based on ElasticSearch. They are migrating all code into gitlab, maturing the continuous integration workflow, and automating more procedures with Rundeck. In addition they have written a root password generator service (the passwords expire rapidly), will upgrade Foreman to 19 and CC7, and will improve the HA and load-balancing between Puppet masters.

ITIL Service models in Detector DAQ Computing (Bonnie King / FNAL)

Bonnie referred to the recent history of DAQ computing with mainly the two Tevatron experiments. DAQ has typically been handled by postdocs and graduate students, who are less familiar with best practices. Her current group manages experimental computing facilities. DAQ systems are often pets rather than cattle and are not used in the same nice clean environments as data centre systems. Redundancy also has a different meaning. DAQ computers are sensitive to kernel and software upgrades, out-of-band management may be limited or not exist at all, and expert shifters may need privileges on the system. Then she briefly introduced ITIL, which she described as more being about best practices. Relationships between service providers and users are often described in service level agreements (SLAs) following a negotiation process. Bonnie gave some examples of such SLAs and then turned to practicalities: setting up a DAQ system is a long-term process, and getting DAQ specialists to submit tickets rather than to pick up the phone is a major cultural change. Dedicated Service-now masks have been developed to this end. Challenges include getting involved early enough in the life-cycle, some grey areas of support (legacy systems not under their responsibility, but still expectations for support), and customers who still like to send e-mail directly – they keep telling them to use tickets. They have had a number of successes though, helping with procurement and purchase recommendations, standardised operating environments, and putting systems under configuration management under Puppet, which allows for extremely quick rebuilds of failing systems. The understaffed summer period was covered much more easily this way. Service-Now helps collecting metrics about the services, of which Bonnie showed some examples. Feedback has been consistently positive. In the future they need to continue being successful to be accepted by more experiments. They plan to introduce more general-purpose rather than specialised machines, and work closely with DAQ software developers (artDAQ) to improve the deployment.

Quattor status (James Adams / RAL)

James introduced himself; since September 2013 he is acting as Quattor release manager, but is also working on CEPH. The he summarised the main features of Quattor… an important part is the community! VUB and ULB (Brussels) have two Quattor-managed clusters supporting IceCube and CMS, and two ONE clouds, which support machine creation with Quattor. They are looking into Quattor as a configuration tool for CEPH and into adopting Aquilon. At LAL Quattor is successfully used for some machines, they consider moving 250 more machines under Quattor control. At RAL the migration to Aquilon is in progress; the ONE private cloud, three CEPH clusters and hundreds of service nodes are Quattor-managed. Quattor is now starting to be adopted more widely across STFC, for example by the ISIS neutron source computing infrastructure. The Castor team has retired Puppet, but not have started the move to Aquilon. They started to experiment with services being users of Quattor, the workflow of which James explained. The largest user, with 38'000 nodes worldwide, is still Morgan Stanley, who are including Vmware ESX clusters and filers. The have introduced the concept of personality versions (previous, current, next). UAM Madrid uses it for 250 hosts, and will move to HTCondor and the ARC CE. Common activities include support for private clouds (mostly OpenNebula, some OpenStack), lots of interest in CEPH, the move from SCDB to Aquilon, and the popularity of FreeIPA. Development contributions continue to increase, as does the quality of the code. Systemd is being supported; there is a new CCM

CLI. Development challenges include the large number of repositories with cross-dependencies, the increasing size of releases and how to build initial images for VMs and containers, for which some ideas exist, but nothing concrete yet.

NVM Express (NVMe): Overview and performance study (Chris Hollowell / BNL)

Chris explained that NVMe is a standard for connecting NAND flash storage directly to the PCIe bus, which avoids the overhead of going via SAS or SATA storage controllers. The interface is designed for highly parallel access, and is supported by Linux kernels from 3.3 onwards (it has been backported to RHEL 6). Devices are available from several manufacturers for capacities of over 3 TB (at relatively high cost). Fusion-IO, while following similar concepts, is a proprietary protocol. Chris then reported on tests run with a pretty standard worker node, to which NVMe and SSD drives were added; the tests were mostly using ext4 on SL 6. Standard disk benchmarks, both bandwidth- and IOPS-oriented ones, show the superiority of NVMe over SSDs (usually by a factor two or more); results of SAS spinning drives are even much lower than SSD ones. Comparing different file systems, the NVMe scores significantly lower for writes on XFS than on ext3 and ext4, which is not entirely understood.

Status of DESY batch infrastructures (Thomas Finnern / DESY)

Thomas explained that the presentation covers the farms in Hamburg, which they run with a team of eight people for configuration, user support etc. The focus is on directions rather than numbers. They run three clusters, a Grid cluster, an HPC cluster, and a local cluster. The local cluster runs on Son of Grid Engine, which is pretty much a dead end; the Grid cluster uses Torque and Mysched and the CREAM CE. The HPC cluster is linked to BeeGFS and GPFS as storage backends, scheduling is done via calendar reservation. The current setup ensures that remote users can submit into the farm which is inside the firewall, but cannot get anywhere else within the firewall. Thomas pointed out the many parallels between BIRD/NAF and the Grid cluster, with similar job properties. Accelerators are being moved to the HPC cluster only. He then listed the beneficial features and policies for batch systems such as project fair-share, resource groups, Kerberos support, multi-core jobs, authentication and authorisation, accounting, monitoring and a number of other features. They now plan to introduce SLURM for the HPC facility, which will co-exist with the calendar reservation, and HTCondor (with ARC CE) for BIRD and NAF with a common support team for the latter two. SLURM is being tested on one management node and six compute nodes with a BeeGFS backend (also for home directories); the scheduling addresses entire nodes only. Logging and accounting is activated, but Kerberos and AFS are not supported. For software distribution, there is an NFS share. There are partitions for private resources, GPU nodes, and standard shared nodes. Concerning HTCondor, they are awaiting the Kerberos and AFS features currently under development. Monitoring is activated via Ganglia; they do not currently see how to transfer their fair-share mechanism from the previous systems to HTCondor. Docker support will be tested a little later. Then Thomas showed the statistics and explained why they need fair-share. They target moving to HTCondor and ARC CE (with Kerberos support!) in 2016, starting with the Grid. SLURM will be introduced to the HPC cluster in 2016.

In the discussion it was pointed out that with the commercial version of GE, 500'000 queueing jobs were not an issue.

Status of benchmarking (Helge Meinhard / CERN; Michele Michelotto / INFN Padua)

Helge started by explaining the background: The need for CPU benchmarking is well established in view of resource requests, pledges, installed capacity, accounting and procurements; HS06 is the standard tool, but there is some latent feeling of uneasiness, as became very obvious during a discussion of GDB on 9 September, which was a little confused, as it was mixing very different issues. An attempt to structure was presented to the WLCG Management Board (MB), suggesting four areas of further work: the CPU power of a job slot as seen by the job, the whole-server benchmarking, the lack of trust in accounting numbers, and the deployment of the 'machine-job features' (MJF) mechanism. On the MB's request, Helge is forming a small group to plan concrete steps and milestones in these four areas.

Michele reminded the audience of what HS06 is, what it has been intended for, and how it is defined. He also mentioned the need by the experiments for a fast benchmark running within the job slot, and compared results obtained with an example proposed by LHCb (small Python script) with HS06, which does not show perfect scaling; other candidate solutions show similar weaknesses. Finally he commented on the future of HS06. Being based on SPECcpu 2006, the community is eagerly awaiting the release of the next SPECcpu suite, which is expected over the next few months. The HEPiX benchmarking WG is hence being revived now.

Friday 16 October 2015

Managing heterogeneous HTCondor workloads (William Strecker-Kellogg / BNL)

Will explained the concept of partitionable slots in HTCondor, which allows for easy accommodation of multi-core or high-memory nodes. Issues include competition and fragmentation; the latter they have addressed by changing to a "fill depth-first" strategy. For provisioning, they use hierarchical group quota; ATLAS is split into analysis and prod, with additional sub-groups. Surplus sharing is enabled (groups can take unused slots from their siblings). For the management they have created a Web-based interface (flask app) that is available from Will's github area. If surplus is enabled everywhere, single-core jobs will take all available slots, hence not enough room is left for the multi-core jobs. He addressed this by automatically re-balancing surplus depending on demand, which requires having access to a metric of demand (site-specific) – they just look upstream into Panda. He described details of the algorithm that ensures a depth-first traversal and showed example scenarios and the resulting CPU usage, surpassing solidly the 95% mark. The algorithm works for any situation where work is structured outside of the batch system; it addresses a provisioning rather than a scheduling problem.

Answering a question, Will clarified that they are currently not pre-empting jobs. They run the script every 10 minutes, but don't change the status of a node before one or two hours.

Upgrade to UGE 8.2: Positive effects at CC-IN2P3 (Vanessa Hamar / CC-IN2P3)

Vanessa started by recalling the history: they have used their own system (BQS) from 1992 to 2012, at which time they started with GridEngine, first from Oracle, then having experienced poor support, from Univa. They are now considering what to do for preparing the expiration of the contract. All nodes currently run SL 6 and UGE version 8.2; Vanessa explained the master and shadow setup. Execution nodes are running with AFS, trace files are kept for five days; they have implemented their own AFS token renewal, GPFS access control, and prolog and epilog scripts. Currently some 12'300 cores are in the farm, of which 33% are used for multi-core, and less than 1% is used for parallel jobs; the farm is fully shared among the various user groups according to pledges. They run some 110'000 jobs per day, with typically 12'000 jobs pending. Taking advantage of improvements in 8.2.1, they decoupled read-write and read-only threads to the batch system, which has caused a speed-up of job submission by a factor 5. For job accounting, job timestamps are now recorded in milliseconds. They can now also enforce limits on qsub, qstat and qdel operations, again containing the impact on the batch system; users can dynamically specify runtime limits for jobs. Short jobs are supported much better now. They are now testing the GPU integration; they also started testing the 8.3.1 release offering a number of additional interesting functionality such as cgroups directly controlled from the batch system, Docker containers, manual pre-emption etc. Vanessa finished by commenting that with their very complex requirements, they find that the support provided by Univa is at a very satisfactory level.

HTCondor recent enhancements and future directions (Todd Tannenbaum / UWisc)

Todd gave a brief overview of the role and staffing of the HTC centre at the University of Wisconsin. HTCondor is part of this effort; he gave a birds-eye view of the system. Even though there are commercial users, the scientific community is their primary target. His reason of attending HEPiX was to strengthen the collaboration between the HTCondor team and HEP. Current channels include documentation, community support e-mail lists, ticket-tracked developer support, bi-weekly or monthly phone conferences, and the HTCondor week – the 2016 one will be held from 17 to 20 May. There will also be a workshop in Europe during the week of 29 February, covering ARC CE as well. He then briefly mentioned the main enhancements of HTCondor 8.2. In comparison, 8.4 adds encrypted job execute directory, the ability to tune kernel parameters, new packaging, improved scalability and stability (10 scheds managing 400k jobs), IPv6 improvements, containers, and improved job submission features, which Todd showed some examples of. The IPv6 improvements mean that the pool can use IPv4 and IPv6 simultaneously. HTCondor had been supporting cgroups for a while; Docker is supported now; Todd explained the Docker universe. However there are surprises – Docker containers can't access NFS/shared file systems, which affects CVMFS, and there are a few more issues. The former is however addressed in 8.5.1. Among the future developments, Todd mentioned advertising images that they already have, resource usage reporting back to the job, and packaging and releasing HTCondor for Docker. For the more distant future, they consider network support beyond NAT etc. Concerning the Grid universe, the jobs are forwarded to

other systems, but remain visible and manageable from the HTCondor interface. He mentioned mechanisms to overflow into public clouds and a number of other developments including the Kerberos/AFS support developed in collaboration with CERN.

Non-traditional workloads at RACF (William Strecker-Kellogg / BNL)

William explained that BNL have a lot of experience with the typical HEP HTC workload, but there are different workloads at similar scale coming along due to light sources, electron microscopes etc., the communities which are less used to the HEP concept of computing. With new beam lines, users may change on a weekly basis, which requires a fully canned solution. The software situation of these users is chaotic at best; some users require special hardware (accelerators) or GUIs. This is at least as much a communication challenge as a technical one. Will then focused on the question of HTC vs. HPC; there may be middle ground between HEP and fluid dynamics. Zero-order requirements at RACF include embarrassingly parallel processing, x86_64 based computing, shared access to data etc. First-order requirements are Linux (Red Hat), free software, and a "friendly" resource profile. As a case study, Will presented a user with an "outlier" style of workflow, whom they facilitated to run on spare cycles on the RHIC farm. This is very similar to the facilitators at the HTC centre at UWisc; again this is at least as much about communication as about technical issues.

Future of batch processing at CERN (Jerome Belleman / CERN)

Jerome started by pointing out that Iain Steers has done most of the work. The current production system is running LSF 7.0.6 (an upgrade attempt during the workshop week had failed) on 4'000 mostly virtual nodes with more than 65'000 cores and 400'000 jobs per day. He then explained why there are concerns about LSF's scalability… as confirmed this week. In 2013, HTCondor was identified as the most interesting candidate replacement system. A pilot service was set up for Grid submissions only with CREAM CE; a few months later they changed to the ARC CE, as it was much simpler to configure and run. Fair-share and monitoring was set up as well. The pilot has run very well since. At the same time they have worked on the infrastructure for local submissions. There are currently 200 worker nodes in the pilot with some 1'300 cores; workflows have been integrated into the automatic workflow presented at HEPiX in Oxford. The pilot is currently running HTCondor 8.3.8 with a single queue and fair-share. Multi-core submissions are enabled, experiments are testing this. The approaches by Liverpool and RAL have been considered for managing the multi-core workload. The service is fully integrated with the Grid operations. Since spring Brian Bockelman spent time at CERN, during which the HTCondor CE was de-coupled from its OSG dependencies; it hence became a very interesting option for them, two CEs have been added to the pilot. Cgroups have been enabled. Monitoring is based on ELK; health checks have been added, using the solutions developed by RAL. The usage is accounted for like for LSF (but the data are currently being sent to the APEL development server). All worker nodes have been benchmarked with HS06, which has been automated with ZooKeeper. Experiments were invited to test, and so they have done… even one which was not invited! In general things went very smoothly; some issues have been cured or disappeared by themselves. A memory leak in condor_shadow was found and fixed in 8.3.6; an incompatibility between OpenLDAP and ARC was found and fixed. The pilot service will be put into operation early November. For local jobs the setup is more complex; Kerberos tickets and AFS tokens need to be managed. The current system used with LSF has weaknesses and needs to be redone; development is underway by the HTCondor team. Job submissions and queries will be handled by a per-user DNS alias. They would like to enforce group membership, which will be implemented via post-submit scripts. Jerome finally thanked the experiments, lead developers, Brian Bockelman, RAL, PIC, DESY, …

In the discussion, the coupling of CondorCE and BDII was brought up.

Basic IT services

Foreman (Mizuki Karasawa / BNL)

Karasawa set the context by introducing NSLS-II; work on computing infrastructure relies on a very small team of people. They have chosen Foreman to go together with Puppet. Its development started in 2009; it is officially supported by Red Hat. It can be linked with provisioning (Cobbler, Razor, RHEV, VMware) and configuration management tools (Puppet, cfengine, Chef, Salt…). She then explained the architecture of Foreman and its interaction with the tools mentioned before. Of the functionalities offered, they are not currently using BMC and REALM. The model they

looked at was a PXE-based unattended install; Foreman also supports virtual machines via the Hypervisor. The link with the configuration management is very tight, making use of configuration groups and host groups; CA assignment and revocation is supported as well. Other features they found useful are the reporting (dashboard – deprecated now), facts (asset management), history and audit, users and roles, and the CLI and REST API. The team really appreciates the automation of otherwise very cumbersome tasks offered by Foreman. Their future evolutions go in the direction of container-based provisioning, decoupling Foreman from Puppet, and following the upstream evolutions including PuppetServer 4.x.

Gitlab and CI (Mizuki Karasawa / BNL)

Mizuki started by stating that git has become very popular; it is a powerful tool, but being so things can also go pretty wrong (for example by pushing testing code into production inadvertently). A management tool is hence very much needed. Gitlab is a Web-based Git repository hosting service with Wiki and issue tracking features; it is very similar to GitHub, but is open-source and can be installed on premises. The company behind Gitlab has 9 employees and some prominent customers including CERN. An on-premise server is attractive, because it is free, eases backup etc. It supports granular ACL controls and can be integrated with powerful tools. The company has acquired Gitorius earlier this year. Mizuki then explained the Gitlab architecture, which is implemented as a Ruby-on-rails application. Useful features include a powerful code review, git-powered Wiki, issue management, code snippets and Web hooks for http call-back. A very important point for them was the integration with CI, which she explained and showed examples of. She then compared Travis and Gitlab CI – they are using exactly the same decription files. She finished by listing the features that make using Gitlab with CI very attractive to them.

Host deployment and configuration at SCC (Dmitry Nielsen / KIT)

Dmitry explained that SCC needed a solution for deployment and configuration covering a multitude of projects (GridKA, LSDF, SDIL, …) and resources (bare metal, ESX, RHEV, OpenStack). They had some experience with Puppet. One objective was to put all deployment and configuration data into git, for which they chose the Gitlab community edition, creating one repository per Puppet module, and following a strict naming convention for groups and projects. They also went for comprehensively using CI. Dmitry then explained how git branches match to environments on Puppet masters. The configuration is done entirely in Puppet, for which all users can create custom environments and which supports both self-developed and external modules. For complex, but mechanical workflows they have provided shortcut scripts. On a host, there are about 30 Puppet modules that are easily identifiable with functions on the host. Dmitry then showed the overall architecture of the system. They are also using Foreman for host provisioning and discovery and for connecting to virtual resources; they use host groups, parameters for host specifications (ENC), parameters for provisioning templates, but no class assignment. He then explained the workflow when a machine is switched on, which is executed fully automatically. They are currently using Puppet 3.8 with the future parser enabled (this should ease the transition to Puppet 4), Hiera with yaml and eyaml, Gitlab and Gitlab CI, Gitbook, and Foreman 1.8; the upgrade to 1.9 is planned.

Monitoring with InfluxDB and Grafana (Andrew Lahiff / RAL)

Andrew started by stating some of the shortfalls of Ganglia: the plots look dated, customisations are difficult, host requirements are demanding, dynamic resources are not handled well etc. For monitoring their CEPH instance, it was clearly inappropriate. They hence looked at InfluxDB and Grafana. InfluxDB is a time-series database written in Go without external dependencies and an SQL-like query language. It can be set up in a distributed fashion; data can be written in using REST or a Python API. Data are organised by time series grouped into databases; each point consists of a time, a measurement, at least one key-value field, and zero to many tags. Multiple points can be written in batches. He showed some examples such as HTCondor (tricking the Ganglia daemon to send data to InfluxDB as individual points), for which he presented dashboards implemented with Grafana. Other examples shown were dashboards for FTS3 and CEPH. Andrew then mentioned cAdvisor for container resource usage monitoring (currently only works with downlevel version of InfluxDB). Telegraf is a collector for metrics from services that it sends into InfluxDB; out-of-the-box this is supported for system metrics, but adding custom metrics is straightforward. However a simple test case tried just the night before did not work quite the way he intended; Andrew promised an update at the next HEPiX workshop.

Miscellaneous

Wrap-up (Tony Wong / BNL)

Tony started with the usual statistics – there were 110 participants, 40 more than in 2004, the previous BNL meeting! 11 participants were company representatives. The statistics shows a good balance between 32 sites from North America, Europe and Asia. A total of 82 presentations were given amounting to 1'630 minutes. This time, apart from site reports, the 'Grids, clouds, virtualisation' track featured the most presentations. He thanked all presenters for having stayed perfectly on time, and mentioned a few highlights from the various tracks. Other highlights include Ian Collier's presentation on young IT internships, the visits, the reception and dinner, and the phishing attack. He then briefly reported from the board meeting: An announcement for the next meeting at DESY Zeuthen (18 – 22 April 2016) will be sent before Christmas; the fall 2016 meeting will be held at LBNL in Berkeley, presumably the week of 17 October, the one following the WLCG workshop CHEP in San Francisco; there was some discussion how the HEPiX benchmarking WG fits with the WLCG effort; the hepix.org Web page has been moved successfully (sites, please provide URL, short introduction, logo to be included); mailing lists, Wikis and file space will be addressed next; the Zeuthen meeting will mark the 25th anniversary of HEPiX, please provide input for a large photo gallery to Wolfgang Friebel. He finally thanked the members of the local organisation, all helpers at BNL, the sponsors and all participants. Helge Meinhard thanked Tony for his very successful double role as co-chair responsible of the scientific programme, and as chair of the local organising committee.