HEPiX Fall 2008 Meeting

ASGC, Taipei, Taiwan

October 20th to 24th

Introduction

With an attendance of around 85, from both "traditional" HEPiX sites and from many Austral-Asian sites, this meeting marked the first incursion of HEPiX into Asia since its creation at CHEP in Tsukuba, Japan in 1991. The meeting was organized in expert fashion by ASGC staff to the extent that the agenda was full around one month before the event and late-minute submissions could only be accepted by extending sessions and indeed the overall meeting length. This is a huge change for HEPiX where often the sessions are only fully filled in the final two weeks. These notes represent my view of the highlights. I should note that I have not reported on all talks, in particular leaving out a small number of talks were which were more computer science-oriented than is usually the case for the mostly practical experience content of HEPiX agendas. All slides and some formal papers are available at http://indico.twgrid.org/conferenceTimeTable.py?confId=471. Elsewhere on the site you'll find many photos taken by their in-house photographer.

Highlights

- An obvious highlight is the good response from Asian sites; the host, Simon Lin, had invested a certain effort in attracting sites from across the region, not only HEP sites, and he got a very good response, the Asian representatives making up any loss of attendees from the more traditional HEPiX sites. Preceding the HEPiX meeting with a so-called Grid Workshop, including a day devoted to CASTOR, surely helped boost the attendance. As a consequence, many new sites were introduced to the HEPiX community and *vice-versa*. We hope to see them again, and not only in Asia.
- From the site reports, one common theme was the growing interest in ITIL.
- Another hot topic, as in the past few meetings, is the challenge faced by many sites in hosting their steadily-growing compute farms. Solutions range from adding external boxes (e.g. SLAC), to getting more from existing space (e.g. GSI) to adding new or re-using existing buildings (e.g. FNAL). And some sites are simply stuck while funding is searched for or administration is overcome (e.g. IN2P3 and PDSF). In which box should we put CERN?
- Based on the number of talks offered and the time spent on it, storage remains the number 1 subject at HEPiX meetings. The various results presented were interesting and the Lustre community is getting more and more excited.
- Resulting from a talk on an MCU upgrade at KEK, the question arose if the major labs should not invest some effort in making video-conferencing interwork better. Question for Tim?
- On the downside, some of the newcomers to HEPiX pitched their talks at the wrong level, a few got into far too much detail and one or two others felt they had to explain every minor detail. We have been through this some years back when we were expanding across Europe and North America; if the new sites maintain

their contact with us, and we have already one firm and one tentative offer to return to Asia, presumably they will soon find the correct level.

- Working groups: the Benchmark WG feels it has achieved the task set it (establish an agreed HEP benchmark) and has plans to document and publish its results and then wind-up. The Storage WG has established a devoted test bed at FZK and restarted various tests.
- On the non-scientific side, the local organization team was excellent, not only in the pre-meeting tasks of establishing and filling a full agenda but in the execution of the event during the week. A dedicated team to change overheads and switch microphones between talks; more-than-adequate and reliable networking; on-time shuttle busses; full catering on site; excellent receptions. Taipei HEPiX will be a hard act to follow.
- Next meetings – Umea, Sweden in May 2009 (probably week of May 25 with week of May 11 as backup; decision soon), and probably Berkeley in October or November.

Keynote Talk – Cloud Computing by Fred Baker (Cisco)
The speaker was introduced as the Chair of IETF from 1996 to 2001 and someone involved with network standards for many years. He started with a brief review of computing from the 50s to the 2010s, giving his view of how cloud computing is the natural next step for providing computing resources but whereas he considers grid computing to be a way of sharing mainframe resources, his view of cloud computing is the complete outsourcing of needed IT power to an outside supplier. He compared the various models on the market today, explaining the Google model in some detail, as far as is known from the outside. He compared the advantages and disadvantages of cloud computing for small and larger companies, concluding that the smaller firms see more on the positive side. Cloud computing is being driven by the technology and the providers but the different models from different providers has a serious risk of lock-in. Nevertheless, it is a natural evolution in next-generation computing centres and there will be an increasing number of alternatives appearing in the market place.

Site Reports

CERN: review of the major events at CERN this year from the Open Day to the Inauguration Day followed by some highlights from IT services including the Skype pilot, the first appearance of SLC5, the eventually-successful migration of Remedy from a Solaris to a Linux base, the move to water-cooled racks in the Centre, Siemens joining openlab, the transition from EGEE 2 to EGEE 3, the CCRC'08 runs of February and May and the move towards single sign-on. He noted that the currently-installed hardware in the Centre total 4,400 systems (31,500 cores) with almost another thousand on order; 7,200TB on 28,000 disk drives and 21,000TB on 35,000 cartridges served by 120 tape units. Later CERN talks reported on various other changes in more detail.

FNAL: all production systems have been moved to the FermiGrid. They are piloting moving mail and calendar to Exchange; preparing a pilot Sharepoint service; and upgrading their helpdesk tool to Remedy 7. August 2008 marked the tenth anniversary of Scientific Linux. Investigating Oracle Identify Management Suite. He described various self-assessment security tools to scan and verify the security of a personal PC. They operate 3 data centres, comprising 8 computer rooms hosting in total some 6,500 multi-core servers. FNAL has made the decision to move to ITIL, the goal to have a certified ITIL V2 (he could not explain why not V3) framework and ISO 20000 within 2 years (by November 2010).

ASGC: ASGC is a Tier 1 site for WLCG and serves Tier 2 sites across Asia and in Australia. Their current installation is 2.4M SI12K which is expected to more than double by year-end. He listed some details about the major changes and expansion of their services during 2007 and 2008.Their servers are largely Xeon blade servers and the next procurement should be based on quad core systems. ASGC are also investigating adopting ITIL in their user support with defined role and escalation procedures and integrating this with their Nagios-based event notification scheme. He described the participation of ASGC in the WLCG CCRC'08 runs showing that they were able to achieve their targets. He described the layout of the different areas which make up their computing centre and plans for the future.

ICEPP (Uni Tokyo): ICEPP is an ATLAS Tier 2 site attached to the Lyon Tier 1 according to the ATLAS model. They have some 650 quad-core Woodcrest servers, a number of 30TB disc servers and an LT03 tape robot with 32 drives. Among their installation and monitoring tools one finds some familiar names such as Quattor and Lemon. A part of the installation is used as an analysis facility for ATLAS Japan collaborators.

INFN: since the last meeting, more INFN sites are moving from the national cell to local AFS cells to provide services to local users while still using the national cell for common tools and software. Of the different sites making up the LHC Tier 2 federations for the various experiments, four sites have completed major upgrades. A mail working group determined that users prefer a local mailing facility so each INFN site maintains its local mail facility with about 500 mailboxes per site. A network working group is evaluating IPv6 "in preparation for the future". Use of multi-core hardware and more virtual servers has exhausted their LSF licence pool and negotiations are underway with Platform for a new agreement. A planned network upgrade is delayed by budget cuts to the Italian physics research programme.

Uni Melbourne: an ATLAS Tier 2 for WLCG, based on ASGC as the Tier 1. It is a small site (80 processors) with expansion plans which are on hold until LHC starts up. Currently limited in bandwidth because their 620Mb connection is split into 4 pipes and they are only able to use one pipe at a time. This should be greatly eased next year when new links are installed across the Pacific Ocean.

SLAC: strange renaming forced by DoE to become the "SLAC National Accelerator Laboratory" (thus SLAC becomes a word rather than an acronym). Personnel changes also: new COO in charge of operating the computer centre; Richard Mount is now concentrating on extending ATLAS and perhaps other LHC computing at SLAC and the position of Director of Computing Services is now open; Chuck Boeheim has moved to Cornell and is replaced by Randy Melen as Core Services acting group leader. Their GLAST satellite was successfully launched and is taking very good gamma ray measurements but, to the annoyance of SLAC, has been renamed to be the Fermi Gamma ray Space Telescope (FGST). BaBar has now stopped taking data, earlier than planned due to budget cuts, and the small expansion to the Centre is for other sciences, for example the materials and energy sciences programme. SLAC's second SUN black box (still only number 3 which SUN has delivered it seems) works well although they discovered some problems with the emergency shutdown.

RAL: new computing building is almost complete but there have been changes in the heating and power provision due to changing economics; migration to the new building is planned for 1Q2009. A campaign of full-disc encryption of laptops is almost complete, including Macs but not Linux PCs for which no authorized solution is available, the recommendation is to use VMware for Linux access. Tenders are in various states of progress for more disc storage and both capacity and service servers. The dCache service is closing and CASTOR use expanding, with acknowledged

good support from CERN. They have updated their trouble ticket system (RT) and have implemented a full 24x7 call-out service for all critical services, triggered by Nagios alarms.

KEK: their computer group supports two main experiments, Belle at KEK and J-PARC (accelerator research) at Tokai some 10s km away. KEK runs three computer centres, one for the B factory, one supercomputer centre for numerical simulation and a general purpose one, on which J-PARC depends. The general service is based on an IBM e-Server 326 running RHEL 4; the B Factory system is based on Dell servers running CentOS and the supercomputer is rented from Hitachi, a combination of a Hitachi system and an IBM BlueGene. A limited LCG service is available via the Naregi grid on a cluster of 80 IBM compute servers running either RHEL 5 or SL4 and will be expanded next March.

IN2P3: BQS, their home-built batch system is alive and well, running some 70K jobs per day. Their main cluster (anastasie) will double in 2009 from 7200 cores today. Their most recent acquisition was based on Dell Poweredge servers. They will upgrade their HPSS to version 6.2 next year. Successfully tested 4 SUN Thumpers with 16TB each on their AFS cell. On the semi-permanent storage side, they plan to implement GPFS to replace NFS. In collaboration with British Telecom, they are researching 2*1Gb links to Fermilab. They are planning to replace the service monitoring tool NGOP (from Fermi) by Nagios. There are plans for upgrading their computing building but they will reach saturation of the infrastructure next year but planning for a new building is slowed for administrative reasons.

DESY: most activity these days is around grids, WLCG, EGEE, D-Grid, etc, with a natural gradual expansion of the various services offered. On the Windows side, there is more effort installing Windows XP SP3 rather than Vista and they may try to wait for Windows 7. Similarly, they have decided to try to wait for Office 14 rather than upgrade now from Office XP to version 2007.

GSI: as reported last time, they have arrived at space and infrastructure saturation for installing and running their computer services and have no budget for expansion. Instead they have moved to water-cooled racks to make better use of existing space and power although it is more expensive to run. Plus, overnight before the talk, the water cooling failed and the speaker was afraid for how much damage has been done! A number of security incidents, largely but not exclusively provoked by private notebooks, have made security a hot topic, including adding more manpower to the security team.

KISTI (Korea): they support an ALICE Tier 2 site with 112 nodes on site and another 58 on a remote site. They run standard ALICE software, processing some 8000 jobs per month. Future plans in the short-term include testing dCache on their local storage resources and trying to make use of the KISTI supercomputer on the grid.

IndiaCMS: the speaker described another Tier 2 site, this time for CMS, located in TIFR and supporting 4 Indian institutes. They have 80 SI2K of CPU in 1U servers plus some 45 blade servers.

LAL and DAPNIA/GRIF: DAPNIA has been renamed IRFU (*Institut de recherche sur les lois fondamentales de l'Univers*). They have upgraded to 64 quad-core nodes and to 27TB of NFS space (from 6TB) while GRIF has added 20 new worker nodes and 90TB on 5 DPM servers. Tests are going on with virtualization to run multiple services on a single node. On the Windows side, they will move to Vista and Office 2007 next year, initially 32 bit but probably soon after moving to 64 bit once some VPN problems are solved. LAL has bought a 2 node SUN Cluster with SAS

JBOD (based on Thumper) to finally replace their Tru64 cluster service. Work has started on a new global monitoring project based on Nagios.

NDGF: NDGF report a “fuzzy line” between Tier 1 jobs and tier 2/3 jobs, a “question of accounting” said the speaker. They run a distributed dCache service. The HPC2N site has a new computer room and they have installed a new tape library based on LT04 drives. The Finnish Tier 2 (HIP/CSC) now has an OPN connection to Denmark which greatly improves their throughput.

GridKa: steady increase in installed power and more acquisition in progress, all based on Harpertown CPUs. The next installation will replace older nodes with a large net saving in power. They have noticed some interesting power differences in recent purchases and there is a report later in the meeting. There is a similar expansion of the disc storage but the acquisition in progress has forced a replacement of their early water-cooled cabinets by deeper ones, as well as a physical movement of racks.

PDSF: they support many HEP and other experiments, including ALICE and ATLAS. Lots of old hardware has been phased out and replaced by quad-core Intel and AMD-based compute servers and GPFS disc space on fibrechannel SATA drives. They have installed 10G networking to the outside world (NERSC and beyond) and to the head nodes (tops of racks) and they are also investigating 10GE for direct node access, initially for storage nodes. Moving to SL4 and investigating SL5.

Storage

Data Integrity and Security: an invited talk from Infortrend Technology which is based in Taiwan. Errors in data can happen at any point in its lifetime and different protection mechanisms are needed for its storage integrity or while in transit. But current methods do not (fully) protect against silent corruption errors, even single bit errors. He discussed ways to find such errors – comparing to a re-built copy, checking data parity on read, but these may not be sufficient and new methods are needed – an extra check, called DIF or T10, adding 8 bytes every 512 bytes which he explained in some detail. Under DIF the write client marks the data on write and integrity checks are done on write to the disc, read back by the disc and read by the client. There is obviously a small transfer overhead and DIF can be disabled by the client if throughput performance requires faster response. There also is a small 1.5% capacity “cost” (8 extra bytes in 512). He then moved on to security of the data, the kinds of encryption options available, in particular the new IEEE P1619 standard for data encryption.

Solving I/O Performance Bottlencks: another invited talk, this one from a new Japanese firm, DTS, promoting a hybrid disc device with a reliable and fast intelligent cache. He explained their device in some detail and claimed a factor of 10 to 100 performance improvement. The intelligence comes from using an 80/20 algorithm to keep the busiest block address area maintained in cache. DTS is available as a software product or a hybrid RAM disc and there is a 500GB hard disc in the pipeline. He showed an impressive list of customers, one of which is CASPUR which is already testing some of these devices and testifies to their performance.

Storage Working Group Progress Report: although initially, the Storage WG was sponsored by IHEPCCC and supported by individual HEPiX sites, it has proven sufficiently useful, that many HEP site IT Managements have agreed to continue to support it with resources, in FTE or in one case with hardware, despite the apparent demise

of IHEPCCC. Since the last report at the May meeting, over the summer a new configuration was set up in FZK on a 10 node farm with 80 cores in total and a realistic use case was obtained from CMS. The tests started in September. GPFS was dropped from the tests for now, based on IBM's recent decision to charge for it and only Lustre and XFS in various configurations were tested; tests on GPFS may be added back in later. The slides contain lots of detail on module versions, the configuration parameters and the tests being performed. First results show that Lustre is about twice as fast as any of the XFS configurations. The similar behavior of all the latter makes the testers suspicious that there is interference of a local file system and tests continue.

What is Lustre: a talk by SUN's director for Lustre (Peter Bojanic). SUN is very proud of Lustre's open source status and he quoted some well-known sites using it – Oak Ridge, Lawrence Livermore, Sandia, etc. It is used by 6 of the most recent Top 10 supercomputer sites, 40% of the top 100. Lustre 1.8 should be released before the end of the year, 2.0 in mid 2009 and plans are being put in place for Lustre 3 and he listed the main new features of each and also some even more long-term Lustre projects (see slides). Kerberos authentication is included in 2.0 along with some replication features and HSM features should appear with V3 but no timetable. Eventually (5 years?) they target up to 1,000,000 clients (today several thousand, target 10,000 for V3). Another project is a meta-data cluster (today limited to 1 node and this will still be the case for V2 at least). Many of these and other plans were further described in slides in his talk.

CERN Storage Update: Dirk Duellmann updated HEPiX on the architecture review he introduced at the last meeting. He made it clear up front that most of his talk referred to developments at CERN and initially at least only affecting Tier 0 operations. Work is underway :-

- to improve tape usage efficiency (aggregation and clustering of data before writing – one of the main target of Charles Curran's criticism at the May meeting; talk later)
- database deployment and internal consistency
- monitoring improvements.

Aside from this work, there is a lot of effort going into better understanding and support for analysis requirements for the experiments working at Tier 0, as gathered by Bernd Panzer's working group. These requirements include secure and scalable access, quotas, POSIX semantics and the need to minimize resources needed for deployment. The chosen answer was to target a new deployment of CASTOR with xrootd in time for 2009 data-taking, evolved from the current deployment at ALICE plus some functional changes in CASTOR version 2.8.1. He explained why xrootd was chosen and the 2.8.1 changes (see slides). The use of xrootd is one step along the road to making the storage system look more like a file system from a user perspective which would ease access to the online slice of data from desktops and from batch. From a service point of view, most sites run an MSS (e.g. CASTOR) and a shared file system (e.g. AFS). Could we ultimately reduce this to one file system? There are possible candidates (Lustre or BlueArc?) but there is a lot of R&D needed and he posed some questions which need to be answered.

Lustre at GSI: starting from their test cluster in March this year, the service has built up in capacity (now 133TB) and in user popularity (400 clients). Nevertheless, there were some hiccups along the way and the speaker listed a few of them, how they were solved and how to avoid them. One serious problem is to over-tax the meta-data service, for example when the single MDS is asked to serve more than 2500 clients. So although data file access is ok, access to large number of small files (e.g. the ls command, programme development tasks) tend to take very long times.

Finding a Practical Distributed Storage Systems: Uni Michigan, an ATLAS Tier 2 site, has performed an interesting study of various options in this area comparing AFS, NFS, dCache, Lustre and xrootd. The speaker explained briefly the architecture of each along with their strengths and weaknesses. Details are in the overheads. Their conclusion was to adopt dCache although they had found it resource-intensive to get into a production state and the speaker said he may revisit this choice, made in 2007, after what he had heard the previous day about Lustre.

CASTOR: finally in this stream there was a 100% CERN session on CASTOR featuring very competent and clear talks by Sebastien Ponce on Status and Plans; Giuseppe Lo Presti on the SRM2.2 interface and the monitoring tools for CASTOR, the latter today in a prototype phase; Miguel Coelho dos Santos on operational experience; and Steven Murray on ideas for increasing tape efficiency, partly he admitted in response to the rather negative if highly personal view of the problems exposed by Charles Curran at the previous HEPiX in CERN in May; indeed some of Charles's suggestions for improvement have been adopted. Since I assume that those of my readers who are interested in CASTOR know the subject matter presented, I will not summarise the talks here. [In fact they made a nice little 90 minute session on CASTOR; why not schedule a computer seminar at CERN on this and repeat them.]

Computer Centres

CERN: Tony Cass reported on the status of the Centre, starting with the good news that the upgrade to supporting the full 2.5MW load is complete. But since this is not enough power for longer-term expansion, in the short-term there is an aggressive exercise in removing old hardware to leave space for more power-efficient equipment. Also, we are installing our first water-chilled racks. Beyond that, he explained current plans for a new computer centre. The idea is for a two-stage approach, initially 2.5MW, later 5MW. So far we have requested from some firms a conceptual design which would then allow a single tender for the actual fabrication and delivery in the 2012 timeframe of a new building usable as a computer centre and we are currently waiting for replies before selecting one. If more space is required before this is ready, local hosting may be possible but very expensive and working with a Tier 1 may be an interesting alternative.

Data Centre Thermo-Fluid Dynamic Simulation: increasing power means more heat and more dynamic thermal regimes; so Pisa INFN decided to model this in collaboration with Pisa University Aerospace Engineering Dept to study the thermal characteristics of their planned data centre. The first step was to map the power and cooling distribution on to a CAD model (in CATIA) of the room. A classical approach to the modeling is not appropriate because of the dynamic nature of the heat flow. Instead one makes a simplified object, for example a simple box shape, and study this before applying the model to the real object, building up using the simple object. The coding is of the model is complete and the first simulations have started. In the question period, he admitted they use the airflow quoted by the suppliers and they have not yet figured out a way to actually measure this. Nevertheless, the first results from the model have produced expected temperatures to within two degrees of those actually measured. Alf Wachsmann referred to a firm in the US which runs a set of sensors to produce a temperature map of an actually-installed computer room and in fact Berkeley use this.

CluMan: on behalf of Sebastien Lopienski, Miguel Coelho Dos Santos presented this CERN-developed management tool for large clusters. In order to present a lot of data on a single display, it is based round 3 entities – state (e.g. node up or down; represented by colours), flags (e.g. node in maintenance; represented by an icon) and properties (e.g. load level, node temperature; represented by colour variations). The idea came from a presentation by Chuck Boeheim at a HEPiX meeting last year as well as the Gridmap application and it is based on a web application

comparing the results of Lemon sensors to a configuration database. He then showed screen shots showing the state of individual cluster nodes, sub-cluster states, cluster load, numbers of users, etc. Future plans include adding a method whereby selecting an entity will run a reconfiguration script to execute some action on a node or several nodes or a head node; e.g. to reboot one or more nodes.

Operating Systems and Applications

SVN: on behalf of DES colleagues. I presented the Subversion pilot service as discussed at a recent C5. A number of other sites report using it successfully and after the talk, Wolfgang Friebel of DESY/Zeuthen said he was happy to share with us his experiences; I recommend that someone contacts him.

High Performance Cryptographic Computing: the speaker, from the local institute, claimed this is needed to provide security and privacy in a networked world, for example for the deployment of secure DNS. The talk went into great detail of the algorithm used to achieve a new speed record in the area of elliptic curve method of factorization. The interesting part for this audience was a description of the general purpose parallel GPU computer from nVidia used for these tests. Normally GPUs are used for graphics, virtual reality or gaming but they can deliver cost-effective, low-latency and high performance floating point which can be applied to other applications. The performance boost comes from using massive thread parallelism (thousands of threads) to fill the instruction pipelines on its 240 "cores" in place of Intel's more "intelligent" pipelining tricks executed on a limited number of threads and cores. He did admit it needs a lot of programming although nVidia supplies some tools to help. He achieved a value of 933Gflops.

Scientific Linux: usage trends unsurprising, V3 going down, V5 going up and V4 more or less steady. V5.2 was released in June with XFS support (at post-install only) and pine was replaced by alpine, with no user complaints to Troy's surprise. SL4.7 was released in September, again with alpine and with Firefox 3. Aside from clean-ups and bug fixing, he will continue to track Redhat releases of RHEL 4 and 5, and 6 when it comes, expected in end 2009 or early 2010. The meeting confirmed that October 2010 was a reasonable end date for SL4 support. Discussing supported hardware, SLAC stated that they do not support SL on laptops.

CERN Mail Update: on behalf of Alex Lossent, Rafal Otto updated us on improvements in the CERN Mail service since the last meeting. He covered the recent upgrading of spam fighting, the ongoing move to Exchange 2007 and the redesign of the simba mailing list scheme; in the future this should use CERN SSO for authentication, Sharepoint for the mail archive and e-groups.

Virtualisation

Hyper-V and Virtual Desktop: Rafal Otto explained CERN's current ideas on moving the Windows virtual server service to Hyper-V and possibly expanding it to desktops. The current Virtual Server 2005 is limited in scope and becoming harder to manage. Hyper-V appears to offer more features and better performance and work has started to offer a new self-service scheme based on that. After explaining how this would be implemented he showed how this could be extended to create a virtual desktop infrastructure but they are very unsure about the usefulness or cost of this expansion.

Unfortunately I missed the rest of the virtualization talks because I was discussing with the local organizers their plans for CHEP'10.

Networking and Security

Video-Conferencing Services at KEK: they have gone through all the usual stages – from ISDN, then H.323 and VRVS to MCUs today. Their second generation MCU (Radvision ViaIP) has been now replaced by a Codian MCU4220 which the speaker claims works with many more clients. The most interesting part of the presentation was the question period – do all modern MCUs inter-work? Is there a need for more collaboration by sites? Dave Kelsey noted that this was something that IHEPCCC could have attacked but that HEPiX may not be the correct body to fill that vacuum because the correct people do not usually attend HEPiX. A few of us said we would take this question home.

Grid Security Update: Dave Kelsey's regular update session. He presented the slide from Romain Wartel showing the various interlinked security and security-related groups and then David reviewed them each in turn summarizing their most recent and current activities and short-term future plans. One of these activities was the security service challenge (SSC3) and he presented the results of the Tier 1 sites responses.

IPv6 Transition: Fred Baker from Cisco came back to present the IETF view of the IPv4 to IPv6 transition. He says some ISPs are reporting to Cisco that they will deprecate IPv4 as early as 2011. Since he is also the current chair of the IETF v6ops working group, he maintains a strong interest in making the v6 rollout work properly. Translation and tunneling may help postpone the deadline for exhaustion of IPv4 addresses but not forever. IETF's basic recommendation is to "turn it on in your existing IPv4 network". But when? One leading advocate suggests starting in 2009, run both in parallel and start enabling IPv6 accesses; then from 2010 start turning off IPv4 and complete the transition by 2012. Even Baker thinks that may be too aggressive.

Cyber Security Update: Tony Cass presented this on behalf of Sebastian Lopienski. There was a brief list of some recent major world-wide incidents which had made the news, some which had affected and infected various HEP sites. He described some CERN changes – restrictions on contacting external DNS servers, restrictions on running TOR, etc.

Benchmarking

GridKa: Manfred Alef has updated his CPU benchmarks with the latest Intel Harpertown and AMD Barcelona chip results. He used both SPECint 2000 and 2006 and the SPECall-cpp set as recommended by Helge's Benchmark Working Group (see below). SL4 in both 32 and 64 bit mode was used. The reader is referred to the overheads for the detailed results, in particular slide 10 shows the speed up you will get in replacing a Xeon 3.06-based chip with one of the newer chips tested. Similarly, slide 15 shows the power efficiency compared to the Xeon chip.

Benchmarking Working Group: Helge gave a status report. Over the summer they have had regular phone conferences, in particular discussing what is or are the most appropriate benchmark set(s) to use in the HEP world. The integer part of SPEC 2006 seems a natural place to start. The working group, some 15 people in Europe and

North America and including experiment representatives, decided to concentrate on worker node power; they defined a standard environment and over the last few months have begun running the benchmarks on a variety of chips at different sites. The result was to propose SPECall-cpp2006 as the preferred benchmark set to use and the slides explained how these values can be calculated from the SPECint and SPECfp published numbers although Helge and Manfred Alef pointed out that actually running the tests locally is quite easy and a script is available if you have the SPECcpu2006 licence (very cheap). Having achieved their primary objective – an agreed HEP benchmark – the working group will write up the results, try to present these at CHEP in Prague and wind-up.

Energy Efficiency of Servers: Helge presented how CERN approaches procurements and how we apply power efficiency to the selection criteria by adding a component to the adjudication price based on power consumption. Similarly to the results presented by Alef, CERN has seen an increase of a factor of 9 in power efficiency in just 4 years. Each talk in the Benchmarking stream was followed by the liveliest question sessions of the week.

Miscellaneous

Site Monitoring: Julia Andreeva reported on site monitoring for LCG sites. The goal is to help site admins, in particular by offering common solutions. A customized nagios as a base with some grid extensions has been made available to help them but much of the talk was on site monitoring from the VO perspective because it is the VO community which often sees the problems first. A variety of tools are in use by VOs, including wide use of dashboards, and SAM is also widely used. So extensions to the SAM schema towards VOs was a first step, in particular a SAM interface to dashboards to allow easy browsing of SAM results with VO customisation. Work is currently underway to provide a high-level consistent view of monitoring data retrieved from VO-specific monitoring.

RT (Request Tracker): this is the open source trouble ticket system in use at DESY (and elsewhere). They have both mail and web interfaces with the data stored in an SQL database (mysql, postgresql or Oracle in that order of recommendation). It is written in Perl (beware of Perl updates) and runs primarily on Linux; a Windows version is available but not supported. It depends on queues of tickets (cf. Remedy ticket types) but DESY warns of creating too many queues (30+). Lots of use of ACLs to protect queues, etc. Scripts (actually called scrips in RT) are used to perform actions on a condition or event (e.g. on ticket creation); these are also written in Perl. DESY has two instances of RT, for IT and dCache with 43 and 2 queues respectively. One can create workflows, for example to import a new computer system into the Computer Room where the ticket is passed in turn to the various teams involved with the physical installation, software installation and network connection. DESY have developed a mail interface to LCG GGUS. Disadvantages or RT include having only a basic workflow scheme and only basic management of RT users. On the other hand it is easy to install and get working and has good open source community support with a commercial support offer available in the background from the main developers. DESY ticket numbers appear to match those of the Remedy Helpdesk application at CERN, 300K tickets in 5 years equivalent to some 1200 per week.

GPFS at NERSC: high performance I/O as well as low-latency for MPI is required for the cosmic experiment based at PDSF. They have an Infiniband cluster and are linked by a 10GigE switch to the NERSC Global Filesystem. The configurations of the tests are described in the overheads and the resulting graphs can be consulted there. GPFS is flexible enough to offer several solutions (Infiniband, F/C, 10GigE. Gateways) and he gave hints on how to choose one or another depending on the application. For them, Infiniband was chosen.

Grid Interoperation: the last presentation of the week was given by Laurence Field. Multiple grid infrastructures have evolved using a multiplicity of interfaces but VOs still want to share resources. In the short term for interoperability, there will be parallel infrastructures (user joins multiple grids and uses the corresponding client for that grid; sites deploy multiple interfaces; e.g. ATLAS use of WLCG and NDGF), then we will have gateways (e.g. interface to NAREGI) and adaptors and translators built into the middleware but eventually we need standards and common interfaces (e.e. Glue 2.0 and BES). Interoperation needs agreement, careful software releases and emphasis on monitoring, user support, operations and accounting. He noted the work with OSG, NDGF, Unicore, etc. He summarized the main lessons learned, showed a slide on the emerging standards and posed some open questions for the future.

Social Events

There was an interesting social programme to show off different aspects of Taiwanese culture starting with a Welcome Reception on the Monday evening. Simon Lin hosted a magnificent 10 course Taiwanese/Japanese Fusion banquet on the Tuesday for the HEPiX Board and Session Chairs; for those of us lucky enough to be invited, this was a contender for high point of the week. But the main event for all delegates was the mid-week 11 course banquet in a downtown hotel to the accompaniment of a small orchestra playing both Chinese and Western pieces on traditional Chinese instruments. The highlight for me was the “invitation” from Simon to sing along to the tune of Loch Lomond from the orchestra’s playlist! The highlight for the rest of the room was that I was able to find someone close by who remembered the words and tune from her days in the school choir and thus able to sing it properly in my place!

Alan Silverman

24th October 2008