CHEP 2010 Report

CERN-IT-Note-2010-007

Taipei, October 18th to 22nd, 2010

Version 3 (17 Nov 2010)

Authors: Joao Correia Fernandes, Ivan Fedorko, Wojciech Lapka, Giuseppe Lo Presti, Alan Silverman (editor)

Introduction 1
Conference Highlights 2
Plenaries 2
Grid and Cloud Middleware 6
Distributed Processing and Analysis 13
Computing Fabrics and Networking 16
Software Engineering and Databases 27
Collaborative Tools 28
BOF sessions 29
Poster Sessions 30
Exhibition 30
Summary Talks 30
Social Programme 31
Weather 32
Next CHEP 32

Introduction

The 2010 edition of CHEP (Computing in High Energy and nuclear Physics) was held in the Academia Sinica Grid Computing Centre (ASGC) in Taipei, Taiwan. ASGC is the WLCG Tier 1 site for Asia and the organisers are very experienced in hosting large conferences, as was proven again throughout the week. As usual Indico was used for the scheduling and all overheads should be accessible online at http://indico2.twgrid.org/conferenceTimeTable.py?confId=3.

The conference was opened by a spectacular 20 minute drum ceremony in the presence of the Vice President of Taiwan, Mr. Vincent Siew. After the drums, Dr. Simon Lin, head of ASGC and Chair of the Conference presented Professor Wong Chi-Huey, President of Academia Sinica who welcomed the 495[1] participants from over 30 countries, including 25 students sponsored by the CERN ACEOLE programme[2], and presented some of the activities of the institute. In his speech, for which the entire audience were equipped with translators, Vice President Siew paid tribute to the success of the LHC and described Taiwan's important contribution to the world of IT.

[1] This compares well with previous CHEP conferences (Interlaken 516, Mumbai 450-480, Victoria 470, Prague 615)

[2] An EU funded programme based at CERN.

Conference Highlights

- The sweet smell of success: LHC works, LCG works, DAQ works, experimental frameworks work, networks can cope with the load; lots of congratulations all round.
- Clouds are on the upswing of the famous Gartner hype curve. A number of real-world production examples exist but some sites have spare manpower to invent local solutions. No agreement yet on what is a cloud. For some people clouds are "fun" but what is their real place in HEP?
- Is there any site not working on virtualisation?
- More emphasis needed on tier 2 sites (debatable?) and certainly on tier 3s.
- I would guess around 50% of the talks related to LHC experiments. There were many talks from other major centres of course (BNL, Fermilab, DESY especially) but many of these also concerned in whole or in part LHC experiment support.
- In startling comparison to many talks at the subsequent HEPiX meeting 2 weeks later, I heard virtually nothing about the implications of the Oracle take-over of Sun in respect of the apparent Oracle policy not to support Sun software products such as Lustre and SGE on non-Sun hardware, at least not at prices that HEP can afford.
- CHEP organisation: fewer plenaries -> higher quality (in the editor's opinion) and more time for parallel talks; presentation of the opening session (drum ceremony) and receptions (massive banquets and displays of local folk culture) was impressive, sets the bar high for future CHEPs.

Plenaries

Appropriately, given the subsequent preponderance of LHC-related talks, the first plenary was given by **Ian Bird**, LCG project leader, who described the **status of the LCG**, how we got here and where we may go next. He presented some measures of the success – the CERN Tier 0 centre moves around 1PB of data per day in and out combined; it writes ~70 tapes per day; the combined grid supports some 1M jobs per day; and it is used by over 2000 physicists for analysis. He was particularly proud of the growth in service reliability, attributed to many years of preparation and testing and he noted that the LCG Memorandum of Understanding, signed by some 49 partners, has been a valuable tool in achieving this success. On the other hand he did point out that ongoing operational problems, usually unrelated to grid middleware, continue to require constant attention. For the future, Bird thinks we need to be concerned with sustainability, data issues and changing technologies such as global file systems, virtualisation and possibly moving away from grid middleware to more standard technologies such as message brokers. He went into some detail on data management where first discussions on possible new directions have taken place and serious study is underway, including the implications for the required LCG networking. Although there is a pilot cloud setup at CERN, Bird believes that today's clouds do not offer some of the features used and needed by the LCG community (worldwide trust, dispersed resources and people, etc); on the other hand, virtualisation is indeed starting to show some benefits in many sites. He ended with a couple of slides which identified ways in which today's grid middleware could perhaps be simplified and/or improved.

Roger Jones then presented a summary of where the **LHC experiments offline systems** have got to. He started by stating that the first year of operations had been a great success as evidenced by presentations at the recent ICHEP conference in Paris. Partly due to the long lead time, including the unexpected extra year of cosmic data, and partly due to the software validation processes adopted, the software for the experiments has been remarkably stable. Event size is a processing challenge for all the experiments; for example the LHCb processing time is quadratic with event size. Another bottleneck is I/O and he gave an example of how CMS have improved their I/O performance. He paid tribute to CERN's support of the Tier 0 and he remarked that data distribution has been very smooth. Data placement, already identified by Bird as an issue for the future, was picked out also by Jones as one area where more study is needed and he gave an example of work underway in ATLAS. At the tier 1 and 2 sites, operations is still effort-intensive and monitoring could be made more efficient, producing more directed views rather than simply adding more monitoring. Tier 2 sites have come into their own for data analysis jobs but the growing user base creates its own support problems. Databases have also performed well he noted, with higher-than-expected access rates. He concluded that it had been a great year and he congratulated all concerned.

Craig Lee, president of OGF[3] then presented a view of **Scientific Cloud Computing in a larger context**. He listed some of the benefits of cloud computing (better resource usage, provisioning of resources on demand, ease of deployment, etc) but there are issues also (although he spent less time on these). He noted that a number of governments have initiated cloud-related projects and he claimed all major operational grids (he quoted EGEE and OSG) are introducing cloud functionality. Given his title, it is not surprising that he considers that developing standards for clouds is becoming more urgent; something to develop inter-cloud trust relationships was just one example. He believes that commercial clouds will not evolve fast enough to accommodate the spectrum of user requirements and that private clouds will predominate the way that organisations adopt cloud computing for productive use. He then turned to what we can do to evolve clouds and he introduced some current international efforts (US NIST and EU Siena) and in particular work in progress in the OGF on emerging cloud interfaces. His "take-home" message is that the bottom-up way of evolving private clouds towards public clouds takes account of the real world and that we must collaborate and coordinate in order to drive progress.

In the last plenary of Monday, **Sverre Jarp** explained **how to harness the performance potential of today's multi-core CPUs and GPUs.** A modern CPU has three dimensions – the number of pipelines, larger superscalar design and vector width. Add to this hardware multi-threading, multiple cores, multiple sockets and multiple compute units and life becomes difficult when trying to adapt software to make efficient use of this advancing technology. Current software is written at a high level and there is little to no knowledge of what this compiles to in terms of hardware usage. We are blessed with concurrency in HEP event processing but today's software does not expose fine-grained parallelism. We should retain the long-standing event parallel model of processing but we need to find ways to limit memory usage per process. One method could be to use forking of the event processing and another is to write multi-threaded code but still treating events in parallel. He gave examples of different optimisations from ALICE track fitting and from GEANT4 and how the use of forking, as used by ATLAS in AthenaMP, can produce interesting results. His shortlist of recommendations was:-

- Broad programming talent
- Adopt a holistic view, examine the performance of the whole programme and its execution behaviour as early as possible in the development chain and clearly split off the main processing section of the code for optimisation
- Control memory use, e.g. forking can be memory-cheap
- Use C++ for parallelisation performance
- Use good tools for compilers, profilers, thread checkers, etc.

Tuesday was opened by **Harvey Newman** describing a **new generation of HEP networking and computing models**. He started with where we are today, continuing to grow exponentially in capacity and adopting new technology and

[3] Open Grid Forum

standards along the way and the LHC experiments are making increasing use of this as was confirmed at a recent workshop in CERN. In light of experience, the experiment models are being modified, for example more use of pulling data to a job rather than pushing jobs towards the data; or a gradual flattening of the tier structure. At the CERN workshop a network requirements working group was established which has already resulted in new requirements for networking at Tier 2 sites. Harvey believes that in future the HEP community must work with the networking community to define what we need and be prepared to pay for it. The future scenario is likely to be an infrastructure of infrastructures with many players in a federated, open environment. Meanwhile, HEP network usage will continue to expand as international network capacity also grows, will the former eventually outgrow the latter? He fears this could happen around 2013 if nothing is done, e.g. transition to 40G or 100G waves. He described some advanced technology tests and challenges being performed in the US and Europe and he ended with his traditional plea for action to close the "digital divide" where he was able to show some developments in central Europe and Brazil.

Roger Goff from Dell, one of the 7 conference sponsors, talked about **computing paths to the future**. He described some current computing trends, most of which are well-known to most people in the audience. To make use of the changes, he suggested using Intel hyperthreading and/or using SSDs. He then repeated part of Sverre's talk of the previous day about application parallelism, with some emphasis on GPUs and other co-processors. He noted that it is important to balance overall system performance, CPU, memory, I/O.

Kate Keahey from Chicago University and ANL discussed **the blessings and challenges of cloud computing for science**. She started by describing Nimbus (from Chicago University), an open source toolkit which can be used to convert a cluster into an IaaS cloud[4] and she gave examples of its use, for example at Fermilab, BaBar and at BNL's STAR experiment where it is used to interface to Amazon's EC2 cloud. She claimed that cloud computing allows to easily deploy custom-made, user-controlled environments on remote resources; it offers on-demand resources with growth and cost management. There is nevertheless a cost and reliability is often an issue. And it forces users into appliance management in which they are often not skilled so Nimbus are developing tools for this. They performed a proof of concept of the new tool with ALICE and a first release is scheduled for 2011. She ended with the statement that cloud computing is here to stay, and invited the audience to join her. In the questions, Richard Mount asked about the ease and cost of data storage on clouds as needed by HEP applications. She more or less admitted this was indeed an issue but hoped that future innovation would solve it.

Lucas Taylor then addressed the issue of **public communications in HEP**. Recent significant LHC milestones have attracted massive media interest. There is a definite risk of being so open in such a highly complex environment. But Taylor stated clearly that we simply have no choice since we have a de-facto contract with society and a recent European survey backed this, noting that scientists are not putting enough effort into informing the public on what they are doing. Therefore LHC needs a coherent policy, clear messages and an open engagement with traditional media (TV, Radio, press) and new media (web 2.0, Twitter, Facebook). Taylor quotes NASA as a leader in this respect. Media and VIP visits to CERN and access to physicists is to be encouraged and has indeed grown massively in the past few years. He showed numbers demonstrating how working with story makers leads to success as measured in the target audience numbers. He praised the CERN programme of inviting school teachers to CERN during the summer as one tool to improve CERN's image with the younger generation and the recent use of tools such as Twitter and Facebook as another and showed numbers on how the CERN twitters are attracting an ever-increasing readership but he thinks we still have work to do on Facebook. He paid credit to CERN/IT for their work on video transmissions during major events. He also noted major video production efforts undertaken by the experiments, for example ATLAS-Live and CMS TV. He ended by encouraging the

[4] Infrastructure as a Service

audience to contribute where they can – work on strategy or tools for communications, write a blog or article for publication, offer a tour or a public lecture and help build relationships with the media.

This was followed by a presentation of the **FAIR project** being built at GSI, Darmstadt. FAIR stands for Facility for Antiproton and Ion Research and it will reinforce GSI's research in nuclear and atomic physics and other activities including biophysics by massively increasing beam intensity and energy. Construction will start next year and switch-on is scheduled for 2018 and it should produce 1 to 10 times the data of LHC. Construction costs are estimated at €1B and the construction convention was signed earlier this month by 9 states with 4 more preparing to join. It was noted that the budget contains nothing for computing (where have we heard that before?) and only €78M for the experiments planned. Two of these are the size of ALICE or LHCB – CBM for heavy ion physics and PANDA for hadron physics – and there are many smaller experiments in other areas. Triggering is a particular problem and DAQ will have to rely on event filtering so online farms will have to be several orders of magnitude larger than at LHC (10,000 to 100,000 cores) and this is a major area of current research. They are also designing their computing infrastructure, based on a Tier 0 centre at GSI and they are gathering first experience in using clouds. Finally he presented plans for a new (dark) computing centre, the cube structure previously presented at HEPiX. It should support 800 10 inch racks using 6MW of cooling power.

In the second sponsor talk of the week, **Campbell Kan**, a VP of ACER described how "**SSDs are the really cool green drive"** of the moment (his quote and the talk title). He listed some of the performance benefits of SSDs such as higher throughput and reliability and faster access to data and he showed some comparative graphs between hard discs and SSDs. SSDs also offer savings in energy and space and the price gap from hard discs is steadily decreasing. The speaker then listed some adoption trends and went into some detail on the part of computer centre power generated by the actual IT load and how SSDs can lower this.

Ian Fisk then talked about some issues around **Storage Management**, the challenges of evolving technology and the increasing number of users. By 2011, the majority of currently-accessed data is likely to reside on disc and this leads to the need for organising processing of data, e.g. the CNAF setup combining GPFS and TSM into StoRM. For analysis access, there are a variety of solution to data management is the evolving challenge. Data placement is very important. Originally this was based on Monarc when there were doubts about network capacity – send the job to where the data is and there would be multiple copies of datasets. The network now being less of a concern, the paradigm is evolving to one where jobs access data from remote sites. By examining in detail how events are actually written to disc, he showed that simple read-ahead on analysis jobs does not benefit performance. An interesting proposal is to use web caching to use true data caches to serve data. Another is to use wide area file system such as Lustre. For Fisk, a discussion on data storage and management must include a discussion on network and access and some modifications in access can lead to large gains in efficiency.

David South on behalf of the ICFA DPHEP study group then presented probably the most serious effort yet for **data preservation on HEP** – what to do with data after the end of an experiment? With few exceptions, often data of an experiment is "stored" somewhere until eventually it is lost or destroyed. He presented some reasons why preservation is desirable but it needs to be properly planned and this is the purpose of DPHEP. Various workshops have been held and preliminary recommendations published and a longer-term blueprint is in preparation. Some important aspects include:-

- Technology used for storage – follow storage trends, migrating from one media format to the next?
- What data to store? Data itself, software, documentation and publications, meta-data (wikis, messages, etc) and people expertise (most difficult)

He presented some preservation models and use cases where the fourth level should allow full reconstruction of the experimental data. He noted how the Inspire project will help in the preservation of documentation and meta-data of publications from completed experiments. He gave an example of how BaBar data is being used to improve public understanding of HEP. He noted a pilot project at DESY to create a virtual environment to preserve their experiments. BaBar are also looking at virtualisation to help preservation. Long term preservation requires planning and long-term governance and it has a cost but this should be compared to the cost of the experiment which is typically many orders of magnitude greater.

Grid and Cloud Middleware

This stream was opened by **Brian Bockelman** from Fermilab describing **OSG developments**. He defined cyberinfrastructure as everything on a campus which assists the scientist. A campus often has multiple clusters and grids can help share resources based on mutual agreements between users but on a campus level the groups may be too small to see real benefit from this and too much additional overhead. So the OSG proposal is to integrate at the batch level which means no additional level for users to learn. He then explained in more detail how they perform this using Condor, including using Condor glide-in to overlay other batch systems such as PBS, LSF and SGE. Use of Condor offers easy extension to inter-campus grids and eventually, with appropriate X509 proxy, to OSG. Issues – security in the widest sense (e.g incident response in an inter-campus grid), conventions between the different grids (e.g. when to install new software, what software), where is the data?

Next, **Stephen Burke** of RAL with a personal view on **10 years of experience of European grids**. He listed the original elements of grid middleware offered by the European DataGrid and how they have evolved or been replaced in the intervening 10 years. He gave his reasons why most experiments prefer pilot jobs (check local environment before running, jobs not sitting indefinitely in remote queues, jobs controlled by VO and not by site). Although SRM could be considered a major grid success story, it is complicated and yet does not handle the full complexity of storage systems and also it does not always mesh well with underlying storage systems such as dCache. The data management area suffers from multiple re-designs and re-writes and much experience has been lost, it needs to be repeatedly debugged and he believes it includes less functionality now than in EDG days. Middleware development has been hindered by "development hell", often relying on specific versions of underlying tools. He suggests that although there have been improvements over the years, middleware developers tend not to focus on error behaviour. His last 2 slides, general observations and outlook for the future, make interesting reading but they were a very personal view and he made several remarks (for example accusing grids of being passive about security) which the session chairman (Markus Schulz) had a hard time not commenting on.

Rob Quick of Fermilab then gave a similar if less controversial talk on **lessons learned in OSG operations**. Technology changes quickly in IT but one should beware of "shiny objects". He stated that transparency is good, but not until the story (emergency situation) is over. Local support is essential, both financial and managerial. Other important lessons concern maintaining good relationships, open communication, flexibility. And of course there is no replacement for experience.

Synchrotron Imaging Computations on the Grid without the Computing Element by Alessio Curri: They had "almost real-time" requirements which could not be easily satisfied by gLite Worker Nodes due to several latencies (job submission, queues). In addition the required deployment couldn't be done in a standard gLite WN. Thus a hybrid approach was demonstrated.

Visualization of the LHC Computing Activities on the WLCG Infrastructure by David Tuckett: A visualization tool was presented, which displays real-time monitoring data from the LHC experiments within Google Earth. It has several use cases:

- promoting the WLCG (CERN visitor centre, VO control rooms, ...)
- a tool for WLCG experts, thanks to integration with traditional monitoring systems.

Flexible Availability Calculation Engine for WLCG by Wojciech Lapka: GridView availability and reliability numbers computed based on metric results gathered by SAM are accepted as a standard benchmark for measurement of the site's performance in the WLCG infrastructure. GridView engine originally supported only one algorithm for all Virtual Organizations but LHC experiments required more flexibility in availability computations and the new Availability Calculation Engine (ACE) caters to these requirements:

- provides flexibility in availability computations
- each VO can define several algorithms
- improves availability re-computations
- uses single authoritative topology provider

A Messaging Infrastructure for WLCG by Wojciech Lapka: Messaging is a key Technology for WLCG, because:

- It simplifies the WLCG infrastructure.
- It is a robust and effective transport layer
- It is designed for distributed software components.

Messaging solutions are already used by:

- CERN Beams department for LHC monitoring
- Service Availability Monitoring (SAM)
- APEL
- Atlas DDM (DQ2 Tracing data)

As well as this message systems are widely used in industry (telco, finance, ...) and others at CERN are evaluating or interested in using such technologies, e.g.:

- Information Systems (BDII update)
- LFC catalog synchronization
- CERN batch service

Designing the Next Generation Grid Information Systems by Laurence Field: WLCG Information System provides a view of the grid service in the infrastructure. Today there are ~2000 resource, ~370 site and ~100 top BDIIs. Since the number of services in the infrastructure continues to grow, the information system needs to meet the future scalability requirements. Under the new architecture the current "pull" synchronization model will be replaced by a Messaging System (e.g. ActiveMQ) where only changed data will be sent. A Messaging prototype has been developed and looks promising but it needs further evaluation and testing in a production environment.

Patrick Fuhrmann of DESY presented **EMI**, a €24M project, 50% funded by the EU and 50% by the 22 partners. He laid emphasis on the need to harmonise the various middleware suites available today. He presented the planned release timetable, describing it as “challenging”. He then moved into the area of EMI Data, presenting the areas in which they plan to work. As well as harmonisation of the access protocols, they would like to improve and increase the use of standards. For example, Webdav will be very useful for non-LHC communities and will be added to StoRM and DPM after EMI-1 (it is already supported in dCache). Moving to SRM, Fuhrmann considers it a good protocol, just badly understood, and explained that EMI wishes to simplify its specifications, document it better, make it easier to use. In the questions, he admitted that CASTOR and Xrootd are not in EMI but there is a representative of CASTOR in their working group. Xrootd is more problematic.

Oliver Keeble then described **services for grid data management**. The target is to provide a coherent set of storage for data management comprising DPM for storage, LFC for cataloguing, FTS for transfer and gfal/lcg_util for client access. He presented recent developments and current usage in each area. The second half of the talk presented future plans to improve each tool, adapting to new requirements, advancing the use of standards. He described some work planned for LFC to resolve a known consistency problem and, longer term, for FTS to remove some current limitations as well as ideas for a more generic file transfer scheduler.

Standard Protocols in DPM by Ricardo Rocha: The Disk Pool Manager (DPM) is a lightweight solution for disk storage management. It is used in over 200 sites, the largest deployment is 1.5 PB. The recent improvements concentrate on usage of standard protocols:

- https, WebDav
- Network File System (NFS) 4.1 interface.

Use of standard protocols improves accessibility, validation, stability, simplifies the implementation and prevents "vendor lock-in". DPM will provide standard-based solutions for all its use cases (space management, remote data access, Posix data access) in the first half of next year.

Distributed dCache SRM Evaluation at BNL by Ofer Rind: Initial implementation of SRM in dCache suffered from an inability to support clustered deployment. In addition the performance was limited by the hardware and network resources of a single node. They used the Terracotta (open source but non-GPL) platform for Java applications scaling) to horizontally scale dCache SRM services to run on multiple nodes in a cluster configuration. Tests performed at BNL showed that trivial use cases work well. However complex cases don’t scale as expected and thus more work is required. On the other hand the high availability features of Terracotta are still compelling.

Adaptive Data Management in the ARC Grid Middleware by David Cameron: ARC has been used successfully for many years, processing thousands of ATLAS jobs per day. But the start of LHC experiments brings increased and chaotic data access. Following issues need to be addressed:

- FIFO order of job processing
- Lack of priority system
- Insufficient handling of protocols like SRM.

New data management architecture for ARC introduces a three-layer structure. Usage of such a layered structure allows more efficient use of the available bandwidth.

Chelonia - A Self-healing, Replicated Storage System by Zsombor Nagy: Chelonia is a novel grid storage system targeting medium-sized VO's. Features include:

- Replica based robust storage system
- Focus on simplicity and scalability

An implementation exists but requires more work and the team looks for contributors (contact zsombor@niif.hu).

Optimising Grid Storage Resources For LHC Data Analysis by Wahid Bhimji: With the operation of LHC going well, Storage Resource Management (SRM) and local file systems are facing challenges to store and analyse the data produced. WLCG sites need to tune their storage for optimal performance. The tuning can be done on several levels:

- Application
- Filesystems/Protocols
- Hardware
- Coordination.

Future plans:

- TestsInABox: For site sysadmins
- Real workload: Proper comparisons with tests
- Evaluate hot topics e.g. NFS 4.1 and xrootd in DPM.

StorNet: Integrated Dynamic Storage and Network Resource Provisioning and Management for Automated Data Transfers by Shawn Mckee: StorNet is a project to design and develop an integrated end-to-end resource provisioning and management framework for high performance data transfers. It's a system for:

- Management of data transfers
- Co-scheduling of storage and network resources
- Resource reservation and negotiation

It's an interesting concept, under development.

Status of the ARC Middleware Computing Element by Josva Kleist (NDGF): ARC-CE is a well established system provided by the Nordic DataGrid Facility (NDGF). ARC-CE has an open source development model and is deployable on a wide number of operating systems. Several enhancements were done in order to improve the throughput. Motivation for enhancements came from LHC experiments which do I/O-intensive analysis.

Status and Developments of the CREAM Computing Element Service by David Rebatto: Computing Resource Execution And Management service (CREAM) is a service for job management at the Computing Element (CE) level. It was first implemented in the context of EGEE projects, now in the EMI project. WLCG was invited to deploy CREAM-CEs in parallel to their LCG-CEs. At the end it supposed to phase out the LCG-CE. OSG also expressed interest in deploying CREAM. First production release was in October 2008. Today there are around 160 instances.

Evolution of Grid Meta-scheduling Towards the EMI Era by David Rebatto: The process of making the glite Workload Management System (WMS) into a more general service for Job Management has already started. Following functionality upgrades have been done:

- Support for message passing interface (MPI)
- Tighter integration with Logging and Bookkeeping
- Ability to forward parameters directly to the batch system

There's also work towards better support for Pilot jobs.

An Update on the Scalability Limits of the Condor Batch System by Igor Sfiligoi: Condor is a batch system used by many WLCG farms. It has undergone many small incremental scalability-related improvements which have made a big difference, e.g. All CMS requirements are met. Measured improvements include:

- Running jobs: 23k -> 49k
- Jobs per day: 100k -> 400k

It's now much more Firewall friendly. In addition Condor-G added support for several new Grid types (e.g. CREAM).

Grid Interoperation: SRM-iROD Development by Eric Yen: Rule-Oriented Data System (iRODS) is an Integrated Rule-Oriented Data-management System, a community-driven, open source, data grid software solution. The main goal of development of SRM-iRODS is to enable iRODS as a storage resource of gLite.

Stefan Lueders presented some **security challenges in grid computing**. He demonstrated these with his “top 5” security incidents, namely

- ssh hacks exploiting trust
- delays in patching against the latest attack vector
- moving from a tiered structure to a P2P structure increases firewall complexity
- inherent risks in virtualisation, although he accepts that there are many benefits in virtualisation; the distribution of VM images is just one part of this, how/when/how often to patch VM images? Even the CERN philosophy of destroying the images after 24 hours may mean that any forensics of an attack are destroyed before the incident is even recognised
- finally, using Amazon clouds means you must sign an agreement giving away many rights (loss of ownership, loss of guaranteed availability) but you are still 100% responsibility for security.

Adoption of a SAML-XACML Profile for Authorization Interoperability across Grid Middleware in OSG and EGEE by Gabriele Garzoglio: An EGEE, OSG, Globus, and Condor collaboration has released in 2008 an Authorization Interoperability profile and XACML implementation. The major advantages of the infrastructure are:

- share and reuse software developed for EGI and OSG
- give software providers reference protocols to integrate with both Grids infrastructures
- when using the same release of the protocol, enable the deployment of software developed in the US or EU in the EU or US security infrastructures

Improving Security in the ATLAS PanDA System by Douglas Smith: PanDA is a job scheduler for grid systems, built on a multiuser pilot-based architecture. Pilot jobs schemes must address many security issues, e.g. execution of multiple users' code under a common 'grid' identity. Security in the PanDA server was described and usage of glExec ('grid' version of the apache suexec) was implemented.

An Authentication Gateway for Integrated Grid and Cloud Access by Davide Salomoni: The Authentication Gateway for the WNoDeS (Worker Nodes on Demand Service, http://web.infn.it/wnodes/) was presented. The gateway uses an online

Certificate Authority (CA) and it allows users without valid X.509 certificate to get a short-lived certificate for accessing grid and cloud resources in the WnoDeS. Three use cases were shown:

- X.509-based access
- Federated Services (Kerberos, Shibboleth)
- Credit-based access (User/Pwd)

AliEn2: The ALICE Grid Framework by Steffen Schreiner: This framework is used by Alice since 2002 for centrally managed productions and since 2006 for user analysis. It provides a complete GRID environment:

- file and metadata catalogue
- task queue for job execution
- Data transfers using several protocols: fts, xrd3cp, scp.

Recently, new development was done:

- Automatic installation on shared area or worker node.
- The selection of the most appropriate SE for each job.

Scalability of Network Facing Services Used in the Open Science Grid by Igor Sfiligoi: OSG relies on many network-facing services, so understanding how they behave under high load is very important. A dedicated team is investigating it, with all major services (GT2, GRAM5 and CREAM CEs, and the BeStMan SRM SE) being tested. They developed a few open source tools, everyone can reuse them: see Benchmarking & Monitoring Tools: http://sourceforge.net/projects/osgscal/

Migration to the GLUE 2.0 Information Schema in the LCG/EGEE/EGI Production Grid by Stephen Burke: The GLUE information schema has been in use since 2002. In March 2009 GLUE 2.0 become an official OGF standard and work on implementation started in April 2009. The schema interacts with everything, so the rollout must be a gradual process to avoid breaking anything. It's deployed in production on BDIIs but sites are very slow to update. LDAP rendering of the schema was shown, the upgrading of the Grid information system to allow both schemas to be used in parallel. GLUE 1 publishing will be switched off only when everything has been upgraded (2012 or later).

Message Passing Framework for Globally Interconnected Clusters by Sajjad Asghar: A model for providing message passing capabilities between parallel applications over the internet was showed. It is based on the Architecture for Java Universal Message Passing (A-JUMP) framework and Enterprise Service Bus (ESB)which is built using ActiveMQ. They also showed results of performance tests with Apache ActiveMQ.

Jerome Lauret of RHIC/STAR described what happens when **STAR meets the Clouds**. STAR has used the grid “modestly” in the speaker’s words, what about clouds? They have tested 5 models from native Amazon EC2 to Nimbus/EC2 to Clemson/Kestrel and he presented very detail charts of the results. The last tested, Clemson/Kestrel, shows some interesting potential and he explained it in more detailed. Kestrel is KVM-based. The full STAR framework is installed and they found it very flexible. They generated 12 billion events (little I/O) but the speaker believes the model would be suitable also for reconstruction jobs although using a commercial cloud would imply heavy data storage costs.

Bogdan Lobodzinski of DESY talked about **H1's tests of clouds**. H1 wish to re-examine their data for new phenomena. They chose Eucalyptus as the cloud middleware and the CEPH filesystem[5] in the Linus 2.6.34 kernel. He listed the steps to run this configuration and the various, numerous, problems generated by both Eucalyptus and CEPH it seemed[6]. Although not ready for production mode [!] he believes the direction look promising. Management is difficult however.

Igor Sfiligoi of UCSD presented **early experience using Condor glide-in WMS with clouds**. The glide-in factory was presented with an image containing a pre-installed O/S and a bootloader that will get the startup script, start it and then shut down. Should the image be included in the glide-in factory configuration or should we allow VO contributions? No decision yet. Another challenge concerns authentication where grids are GSI-based and clouds are not, today at least. Proxy delegation is also missing in clouds. A prototype exists, submitting RHEL 5 jobs with CernVM FS to Amazon EC2 and to the ANL Magellan Cloud (Eucalyptus) and it seems to work quite well on the former and they are starting to turn this into a production-level service. He ended with some benchmarks running Monte Carlo jobs.

Artem Harutyunyan presented the **CERNVM CoPIlot as a framework for orchestrating VMs running LHC applications in a cloud**. The goal was to integrate cloud resources seamlessly into grid infrastructures. But applications are heavy to deploy and change frequently. Users should see no differences between jobs running on the grid or cloud but grid certificates cannot be used on untrusted environments. Hence the CoPilot which consists of components to integrate cloud resources into grids. The components, which communicate via Jabber/XMPP to the CernVM images on the cloud or on "volunteer" PCs (Boinc), include a job adaptor, a storage adaptor, a job manager and a context manager. He demonstrated this graphically for several use cases from ALICE and ATLAS.

Belle computing: An experiment on CP violation due to run at KEK, targeting O(ab^-1) total data. Mostly centralized computing, very high primary event rate (row data rate ~ 1.8 GB/s, higher than ALICE or HI). They use DIRAC, the framework developed by LHCb, and reuse Grid-based technologies and/or Clouds for the distributed part.

StratusLab: Cloud-like Resource Delivery for Production Grids by Michel Jouvin: StratusLab is a European project (started in 1 June 2010) which aims to provide coherent, open-source private cloud distribution. It combines existing open-source technologies with cloud management solutions developed within the project. Grid and cloud technologies are complementary. Grid middleware provides federation of distributed resources and services. StratusLab allows easier deployment and maintenance of sites by administrators, it makes processes more agile, appealing to new scientific communities. The first release of StratusLab toolkit is expected in a couple of weeks.

Integration of Cloud, Grid and Local Cluster Resources with DIRAC by Tom Fifield: DIRAC was originally designed to support direct submission to the Local Resource Management Systems of clusters for LHCb. Then support for grids was added and the last change was support for Amazon's Elastic Compute Cloud (EC2). During the talk, the results of tests (Cloud-only, Cloud and HPC, Grid, Cloud and HPC) using the 2010 Belle Monte Carlo use case were shown. DIRAC is ready to provide seamless integration of cloud, grid and HPC cluster resources to its users today.

Cloud over Grid for e-Science by Eric Yen (Academia Sinica): The best strategy to use Clouds is to integrate the new technology with the current production Grid. Thus a new framework with virtualization capabilities on computing, storage, network and management features to support virtual machines (VM), virtual services and virtual platforms is under development at ASGC over its gLite middleware.

[5] See Ceph: A Linux petabyte-scale distributed file system

[6] My [the editor's] favourite comment: "the file system CEPH is stable "if not too much files [sic]"

Volunteer Clouds and Citizen Cyberscience for LHC Physics by Artem Harutyunyan: Open source platform for volunteer computing (BOINC) was used for distributing LHC physics simulations to volunteer computers via the Internet. Since HEP code requires a complex environment, CernVM was combined with BOINC in order to run the BOINC client and virtual machines controlled by the CernVM wrapper. This project has several benefits: economic (extra CPU power) and outreach (raising public awareness of CERN and the LHC).

Distributed Processing and Analysis

Maria Girone presented **WLCG Operations and the First Prolonged LHC run**. Intended to be complementary to the subsequent experiment talks, this presentation addressed whether the service had delivered quantitatively (yes, up to and beyond the foreseen level according to metrics such as number of jobs, data transfers etc.) and qualitatively – the main theme of the talk. This was done using the Key Performance Indicators used to measure WLCG service quality and in particular the areas (infrastructure, network, middleware, database and data / storage management) where significant incidents had occurred as well as the time take to resolve them. Although the time to respond to incidents is usually well within targets, many incidents are not resolved within one day and a significant fraction (much higher than the target of a few per cent) take many days or even weeks. Notwithstanding the general satisfaction expressed with the WLCG service, these KPIs not only highlight areas where investment is most urgent but also provide a convenient way of measuring progress over the coming months and beyond.

Dan ven der Ster presented **Hammercloud**, first explaining why it was developed (clear need to improve stress testing of grid sites). There are two use cases – on-demand large-scale stress tests using real analysis jobs on one or many sites; and frequent ping jobs to validate sites by testing all the services needed by the distributed analysis tools. Hammercloud has evolved from job robots from CMS and ATLAS and to date has run some 3000 tests comprising over 8 million jobs. He described the use cases, the workflow, the components and showed the look and feel of the tool. He showed examples of its use in ATLAS, CMS and LHCb (no apparent interest from ALICE he noted during the questions) and described plans for the future – a more generic core to make adding a VO easier, more powerful presentation of the results and RSS feeds to allow people to subscribe to results.

David Tuckett then presented the **experimental dashboards for monitoring**. The architecture is based on a common framework with loose coupling to the data sources to allow for easy modification to applications to different sources and/or different VOs. Users are closely involved in the development process. Monitoring data is exposed to external applications which include generic applications for jobs and tasks, sites and services and specific applications for VOs. He then went into some detail on some specific applications. Results can be displayed at different depths of detail with views for different target audience from 24 by 7 shifters to managers and paper or presentation authors. He believes such monitoring has played a key role in ensuring the quality of LCG activities, on infrastructure quality; it has improved the user experience and it has eased the evolution of the various monitoring applications with time.

Dirk Duellmann talked about **proxy caches in a multi-tier grid environment**. A proxy cache can hide problems due to the latency of accessing remote data from Tier 2 or 3 sites in particular. Users access the proxy with the same rights as the real cache so we can establish a hierarchy of cache servers. Which protocol to adopt? Candidates include http, xroot or client side caches and he gave pros and cons for each. After some initial tests in July, they intend to run some tests shortly to get some practical experience. He showed the current testbed and presented some preliminary results which look promising.

Tony Johnson then presented the **Gamma Ray Space Telescope project**, now is confusingly renamed **Fermi**. It was launched 2 years ago and downloads its data every few hours to NASA who send it to SLAC for processing and onwards distribution, including making it public within 24 hours. This led to a long set of design rules for the processing chain. There are two components, one to catalogue the incoming data, one to pipeline the data to batch jobs via job controllers which mean the jobs can be run at remote sites and can interface to various flavours of batch. Oracle databases are present throughout of course, with Java stored procedures for performance. In fact Java or Java products are extensively used in many areas of the setup.

ATLAS DM Operations by Ikuo Ueda: an explanation of the ATLAS Distributed Data Management (DDM) System, providing some figures of the current operation, which largely exceed targets. Peak I/O rates of 6 GB/s, need to throttle export as that's the physical limit for the Tier0. Monitoring tools designed for the shifters to have an overview of the system and quickly identify issues. Data distribution and usage are notably different from expectations as these are 'first' real data (e.g. more on detector's performances - ESDs - than physics - AODs). Nevertheless smooth operations, the Grid reached 10GB/s 'without a problem'. Looking forward to the challenges of more LHC data.

CMS Computing by Daniele Bonacorsi: a presentation of the CMS computing model, starting from the benefits of the data challenges, in particular STEP'09, to get ready for the 'real' challenge of the LHC data. The Tier0 operations performances exceeds expected/design levels. However, resource utilization is not yet at the target level: LHC 'live' time is below expected level. Sites' readiness defined as an AND of many tests, and monitored over time. Data movement across different Tiers: in particular, exploiting the T2-T2 full mesh. Commissioning it at 7 links/day average over last 6 months, almost completed now. Analysis at T2s: 800 unique users/day and counting. Conclusions: smooth operations, though resource demands will increase soon as more physics is coming - 'stay tuned'.

ALICE Computing by Costing Grigoras: The experience with AliEn and the Grid services, at the core of ALICE computing, is presented. Target in the last few years: decrease the number of services needed at a site. Data movements are performed via xrootd, with 3rd party copy. Data Management Operations: 1st pass at CERN, in 24h; 2nd pass at T1s, ~1 month after; chaotic analysis in the whole Grid with high stability and performance. Overall steady operations. Castor @ CERN has 2.3PB ready now for the Heavy Ion run. Conclusions: Grid operation is routine, simplification of the Grid concept as no difference is foreseen between T1 and T2 sites (see also Ian Bird's plenary presentation).

LHCb Computing by Philippe Charpentier: Peculiarities of LHCb computing: small experiment, small (~35 kB) event size, yet high trigger rate, hence reconstruction cannot happen all at CERN and T1s are needed as well. Typical issues with Data management concern SEs availability and reliability. Evolution of real data in 2010: going for higher luminosity, higher pile-up (2.3 collisions per trigger), hence larger and more complex events than expected, up to 50-60 kB. Need to adapt the computing; nevertheless, managed to cope thanks to available resources, now running ok. Analysis: very little at T2s. Foresee to increase it at T2s, defining 'Analysis centres' and 'Reconstruction centres'. However, main caveat is that data access is the weakness of the Grid. Full reprocessing will start from Nov 2010, and usage of resources especially beyond T1s is being investigated.

Xroot at GridKa by Artem Trunov: A description of the Storage Element setup for ALICE at GridKa: based on servers with (SAN-)attached storage and local GPFS. Main advantage: hardware failures don't result in service unavailabilities as data may be accessed from multiple servers. Tape backend implemented using standard MSS features from xrootd; SRM interface using BeStMan. No performance tests but observed good performance in production and received good feedback from VO (ALICE) - the second largest SE with 1.3PB already deployed.

WebDAV for High Throughput Data Access by Gerard Bernabeu: Experiences at PIC, motivated by moving towards standard protocols (today mostly dCache/dcap accesses). Considered NFSv4; still experimental though it works with dCache. WebDAV is standard and provides both random (POSIX) access and efficient bulk data transfers. dCache chosen as backend as already known and deployed at many sites, as well as at PIC. Performance tests show little throughput overhead wrt native xroot (xrdcp) on large (1 GB) files, and best throughput with small (2 MB) files. Conclusions: HTTP/WebDAV seems promising as a standard protocol for bulk transfers, while for random access it seems more investigations on NFSv4.1 should be foreseen.

PHENIX Analysis by Christopher Pinkenburg: A detector for Au-Au collisions, 800 MB/s RAW data rate. Presenting experience with reconstruction and analysis. Typical issue: inefficient access to tape when trying to run a cycle of analysis (3 weeks, ~200k files from tape, delays) - the so called 'Analysis Train'. Paradigm changed to the 'Analysis Taxi': close access to general users, only 'taxi drivers' have access (i.e. production), users are allowed to jump in once a week provided they run resource constrained (and valgrind-validated) jobs. Web interface for users to monitor their jobs. Storage: 2PB data, backed up on HPSS, moved through the system per month. Other issues include local disk I/O: using ROOT TTree's (without the TTreeCache optimizations) makes disk I/O quite expensive - thinking about trading I/O with CPU (e.g. recalculating values instead of reading them).

CDF Access to LCG by Gabriele Compostella: CDF (Collider Detector at Fermilab) is a FNAL experiment recording ppbar collisions at 1.96 TeV, 8 fb^-1 of data on tape as of today. The main CDF centres are at FNAL and INFN-CNAF but they have moved from a dedicated computing facility to opportunistic usage of the LCG via an extra middleware layer which has been developed to allow the existing software to submit jobs using gLite. Typical issue: jobs are composed by subparts and the time to complete a job may get very large if some subparts are queued for too long. Partially circumvented by resubmitting jobs that don't complete after 1h. Other issues to be addressed include how to use the LCG also for the data flows.

Parallel MC on the Grid by Jon Kerr Nilsen: A framework to run MPI-like parallel applications on the Grid, based on Ganga and called GaMPI. Presented first results, some of them promising as they compare well with pure MPI. But a number of issues to solve, typically around parallel jobs submission.

Data preservation in HERA by Janusz Szuba: A discussion on this hot issue (as also seen at the plenary) and how it is being addressed at DESY/IT. At format level, RAW and mDST data are stored in ROOT format. In terms of technology, proposal to develop a storage solution at DESY to target data preservation, including periodical recall and reprocessing of the data. Another aspect is documentation preservation, including non-digital information: considering digitalisation of ZEUS documents.

ATLAS DDM status update by Vincent Garonne: An update of the Distributed Data Management system for ATLAS: this year they reached 10GB/s data transfer aggregated over all ATLAS clouds. A cloud in this sense is roughly a Tier1 with its associated Tier2s/3s. Architecture based on an Oracle database surrounded by services, Apache as frontend, and a command line API that is dataset-oriented. DDM enforces the ATLAS computing model. Among different features, a deletion service is used to delete a dataset from a site; 4PB/month has been deleted on 400 endpoints. A Tracer service monitors dataset usage over time. Focus now in consolidating and running stably, with an eye on future storage evolutions as they may significantly impact the ATLAS computing model.

Monitoring CMS Computing by Jose Hernandez: Status update of the infrastructure used to monitor CMS jobs running at Tier1s and Tier2s. Reports from the CMS computing dashboard, measurements of the readiness of a site as a function of its jobs' success/completion rate. Typical figure: 80% of the jobs successfully complete at the first attempt, almost all within a number of automatic retries.

MyOSG and MyEGEE by Rob Quick: The problem: there are a plethora of operational tools for the grids, developed by different teams with different intentions. The target is to have a common visual web-based interface, in particular homogenizing the interaction with OSG and EGEE (gLite) grids. Notably, output is provided in various formats including novel ones (e.g. iGoogle, content for mobile devices). However, the risk that this is just yet another interface to the grids is not negligible. A MyEGI interface is soon expected to replace MyEGEE. Experience with users is that mobile content is the most popular form of access.

Computing Fabrics and Networking

Building a High Performance Computing Infrastructure for Novosibirsk Scientific Center by Alexander Zaytsev: The situation with large computer centre infrastructure of Russia improved during last decade but the lack of national broadband is still a reality, especially for remote sites such as Novosibirsk. The project may be considered as a base for regional broadband network later to be connected further east. Heterogeneous hardware and software units have been integrated in NSC (Novosibirsk Scientific Center) and such a solution may be exploited at other regional sites to build common computing and storage environment. Centre hosts 1280 cores, 13.4 TFlops, Infiniband, 2x70 sq.m rooms, 140kVA power input & 120 kW of cooling. More regional centres are connected with ~10GbE supporting large data transfers between clusters. Last integration step will be common job handling environment. Virtualization is based on Xen and VMware. Challenge was to support software bound to the SLC33 at KEDR experiment.

Computing infrastructure for ATLAS data analysis in the Italian Grid cloud by David Rebatto: The Italian ATLAS grid cloud consists of sites with different set-ups and purposes (Tier-1@CNAF, Tier-2s, grid-enabled and non-grid tier-3s). All these Tier levels were discussed. A few highlights: academic network based on 10Gbps backbone and 1-2.5 Gbps links to all INFN and academic sites; with target of 2.5 Gpbs to all Tier-2s by August 2010. All Tier-3s are using GPFS+StoRM as storage solution, successfully tested by HammerCloud. In general, based on last year's experience, the performance of Italian Grid is satisfactory and no bottleneck has been observed in data transfer and user job execution.

The National Analysis Facility at DESY: Status and use cases by the participating experiments by Kai Leffhalm: NAF usage by LHC experiments was described. For details see slides.

First Challenges for the ALICE Tier2 Centre at GSI by K. Schwarz: This Tier 2 centre comprises about 20% of the global ALICE T2 requirements. GSI batch farm is based on 340 nodes (~2700 cores), 1PB Lustre cluster, 300 TB xrootd, Grid cluster, etc. Monitoring is based on the MonALISA tool and integrated with ALICE global monitoring. One of main focus is delivered to GSI cloud solution, based on Debian as OS, KVM as hypervisor, libvirt and OpenNebula for building clouds (16 physical boxes -> 100 VMs). This 'Grid site in a cloud' effort will continue. Various processes run onsite: massive simulations, reconstruction and calibration processing and data analysis. PROOF is used for development and prototyping to get fast response and good statistics (large job run on Grid).   The tool PoD, which allow PROOF to start on any resource management system, was developed at GSI. Permanent effort is put into system I/O optimization. Solid plan for future contains tasks like enabling 10Gb link to GridKa, migration from LSF to SGE for batch system, combine local and grid storage (probably Lustre with xrootd frontend), etc.

First operational experience from a compact, highly energy efficient data centre module by Manuel Delfino: Few facts about the PIC[7] data centre: total electrical power 2000kVA, common cooling for computer centre and offices 1230 KW. 150 sq.m machine room. ~2500 cores, ~3.5 PB of disk storage, 3.4 PB tape capacity (STK 8500, IBM 3584). One of main limitation was/is cooling: power usage effectiveness is 2.3. But cooling cannot be extended or replaced or only at very high cost. Direction of more efficient data centre modules was selected. Solution is based on using 25 sq.m x 2.2m height AST SMART-SHELTER with 2 standard gas expansion A/C 40kW cooling (redundant, no UPS) and 100kVA UPS for IT power. There is a power meter on each circuit collecting environmental data (cooling parameters, temperature, etc.). Cold and hot aisles are completely isolated. In reality few limitation are detected (e.g. humidity variation still under investigation, hard isolated racks). Monitoring implemented via SNMP, HTTP, TCP/IP and SCADA interfaces, using various tools (from in-house scripts up to CENTREON). In future, apart from solving open issues they will implement automation of system reaction on detected problems.

Advanced Cluster management for Large-scale Infrastructures by I.Fedorko: CERN CC is facing a rapid increase of capacity and, together with large variability of configuration, also an increase of management complexity which has to be addressed. All these issues have to be addressed within an existing software infrastructure. The Cluman project is focusing on an advanced visualization of the CERN CC infrastructure and enhancement of administrative job management with integrating existing tolls/software. Cluman is providing visualization of a fabric hierarchy merged with monitoring data. Administration tasks may be launched from a command line as well as from a web interface. Attention was put on high granularity of privileges to allow specific tasks for a dedicated user on dedicated resources. Software is based on the 3 tier (frontend, middleware and database) web architecture based on REST (Representational State Transfer). DB (flavour-independent) backend is holding request states and user information. System is able to process 10k jobs (every job consists from couple of requests as schedule, start, return result etc.). From benchmarking one could conclude that 100k processed jobs is a realistic target with infrastructure enhancement (Django, oracle server pool, etc.). Advance visualization and job management on top of existing Quattor management tools were built and put into production.

Efficient Management of Large Clusters with Minimal Manpower by Tony Wong[8]: This talk was focused on discussion of operational experiences from running the BNL data centre over last 10 years and facing dramatic CPU and storage capacity increases with minimal manpower increase in parallel. From HR point of view flexible and multiple skilled staff is key point of manageability. From software point of view open-source solution is preferred such as MySQL, Nagios, Ganglia, puppet, etc. Operational experience demonstrated strong need of solid monitoring and high level of automation. Demonstration by numbers was made: computing grew from 1999 by a factor of 42, storage capacity by a factor of 8625(!), power usage grew by factor of 47 but manpower only by a factor of 4. Various issues were discussed including cooling problems: power usage grew faster than cooling capacity, so they live without fault tolerance; cooling units were added but not enough to face the additional heat load. Nevertheless the system's reaction to the northeast US power cut in 2003 which was handled by fully operational and successful automatic emergency shutdown was considered a good success. Few operation numbers: on average ~10 trouble tickets/month, 5.7 million batch jobs/month, 4.8 million CPU-hours/month of effective runtime, 3% of ineffective CPU-hours/month due to batch scheduling inefficiency. Average occupancy is ~87% CPU over last 12 months. This all is provided by ~2000 nodes.

Autonomous System Management for the ALICE High-Level-Trigger Cluster Using the SysMES Framework by Timo Breitner: Alice High Level Trigger Cluster was discussed. Consists of 269 nodes, 3 networks (Infiniband, FastEthernet, GbEthernet). System operated by 1 FTE sys admin. Management complexity is increased (on top of hw and sw

[7] Port d'Informació Científica institute

[8] Formerly Tony Chan

heterogeneity) by functional heterogeneity. There is a serious requirement for automation and notification (monitoring) and to cope with the situation SysMES Framework was developed. It is a rule-based tool set for the monitoring and management of the systems and applications targeted over the network. Information is stored in central storage and accessed via ORM. System assumes and provides some roles of a sysadmin and an operator: automatic inventory of devices, distribution and configuration of resources, state recognition and automatic reaction, etc. Global event evaluation is available. All these functional elements were presented. Global events are built from conditions over all nodes in racks or whole cluster, optionally with limited duration (temporal condition). Multiple actions may be executed after correlation condition is matched.

Fabric Management using Open Source Tools by Jason A. Smith: This talk was focused on discussion of software selection and consideration for handling operation in the BNL data centre. As mentioned in previous talk, a lot of attention is put on open-source solutions. Main areas, which were covered by the software selection, were provisioning (Cobbler/Koan), asset management (Fusion Inventory and GLPI), configuration (Puppet) and virtualization (in the near future). Cobbler/Koan is written by RedHat, supports the main Linux distributions (RHEL, SL, CentOS, etc.) and has CLI and web UI. For configuration a fully automated solution was needed after provisioning and later reconfiguration with centralized management with control on application level. Many software solutions were evaluated (Cfengine, puppet, chef, etch bcfg2, AutomateIt). Puppet is simple but powerful (DSL-Domain-Specific language), offers a configuration catalog and dependency resolution, dashboard, configuration visualization etc. Asset management software (Fusion Inventory & GLPI) was described (collecting server inventory, snmp data from network devices, >100 plugins to extend features). In future rack location will be provided in the data flow and they will finished the pending puppet setup, integrating it with Nagios and deploying virtualization management (RHEV, ConVirt, oVirt, openQRM etc.)

Evaluation of the Scalability of HEP Software and Multicore Hardware by Alfio Lazzaro: Very interesting talk. Software should be designed to cope with multi-core and multithreading (SMT). Currently multithreading (2hw-threads/core) is usually turned off by default. Multithreading usage would have also impact on procurement procedures as CPU power measurements are different. Sequential software profits poorly from multithreading but at the same time consumes significant memory resources. The talk continued with a detailed discussion of recent hardware benchmarking: Westmere-EP and Nehalem-EX. For hardware specifications and details of the results the following link should be visited: http://www.cern.ch/openlab.

Dynamic Network Provisioning to Enable High Performance Data Transfers for LHC by Philip Demar: The End Site Control Plane System (ESCPS) project was introduced. The aim is to configure local network for use of circuits, process user requests for such circuit services and coordinate WAN circuit services. Long term vision is to have Federated Control Plane of Dynamic Circuits built from end sites. Typical site network model consist from hosts, devices (e.g. router, switch) and physical links (e.g. interfaces, ports). ESCPS model is based on aggregated flow (unidirectional stream of packets), end entities (physical or logical entity as flow start/end points, e.g. site border), virtual paths (dedicated network path between end points) and rules (configuration units, which are creating a desired virtual path, using traditional network elements as attributes or parameters of these rules). Circuit termination points shall be selectable (e.g. site perimeter or host network interface). The architecture details (various system views, components, workflows, etc.) were discussed. Backup slides even extend details by results and additional information. ESCPS specifications are finished as well as component design with prototyping. Next step is integration of the components into a unified framework.

Powering physics data transfers with FDT by Zdenek Maxa: In this talk FDT (Fast Data Transfer) was introduced as element powering data transfer in HEP. FDT was discussed *per se*, as fdtd/fdtcp and the PhEDEx part (Physics Experiment Data Export, see next talk). FDT is an open-source product transferring data over TCP, running on all major platforms as it is written in Java, using simultaneous TCP connections (gain on distributed file systems), integration with MonALISA, etc. Under the same conditions GridFTP (globus-url-copy), SRM (srmcp) and FDT (fdtcp) were tested and compared. GridFTP managed to migrate test data during 74 min with ~23MBps, SRM during 80-120 min with 14-21MBps, FDT during 41 min with ~42MBps. PhEDEx could profit from excellent FDT performance as well any solution using SRM (such as fdtcp provides an analogous interface to srmcp). There are a few points to note: at the moment FDT is not capable of grid-auth transfers and FDT interface is not directly compatible with PhEDEx (using only SRM and FTS). Some of the issues are addressed by an ftcp/fdtd implementation with the following characteristic: set of files transferred by a single pair of reader/writer processes, transferring within a single transfer job, grid-auth transfers, but no scheduling for now, monitoring via MonALISA (e.g. involved host monitoring, status tracking of thousands transfers, progress of long-term transfers etc.). An integration of FDT-PhEDEx based on fdtcp was discussed. Project is ongoing with well defined short and long term plans.

Improving CMS data transfers among its distributed Computing Facilities by N. Magini: Talk was focused on an introduction of CMS data transfer model beginning with hierarchy description (T0<->7xT1 ~900MBps aggregated; 7xT1<-> ~50xT2 where T2->T1 run with 10-20MBps and T1->T2 50-500MBps). Details of CMS transfer workflow were described with the roles of PhEDEx (Physics Experiment Data Export) agents, FTS, SRM and grdiFTP as well as component interactions. PhEDEx is the scalable CMS transfer management system providing a transfer request interface and a set of software agents designed to fulfil specific tasks as well as monitoring transfers. The FTS architecture was explained with details of FTS monitoring (full daily statistics extracted, reporting etc.) and examples from system operation (e.g. identification of links with low rate-per-stream).

High Throughput WAN Data Transfer with Hadoop-based Storage by Pi. Haifeng: the Hadoop distributed files system (HDFS) was described with focus on performance results from the CMS experiment. Advantages (Java-API, scalability etc.) and potential issues (e.g. incompatible with current multi-stream data transfers -> need of memory buffer implementation as in GrdiFTP and FDT) of HDFS-based storage were discussed. Key components of this system are:-

- the BeStMan (Berkley Storage Manager) SRM service frontend
- GridFTP servers (WAN file transfer)
- namenodes (namespaces of distributed FS)
- datanodes
- a FUSE POSIX-compliant interface (local file request)
- GUMS or Gridmap for authentication.

The data flow in CMS was described. BeStMan scalability was tested by Glidetester with focus on high concurrency to test processing rates and failures. It was observed that the processing rate is significantly affected by the size of the directory; the globus-based BeStMan server is strongly depended on concurrency while new implementation of BeStMan2 (~50 HZ effective processing rate) is not. For throughput measurement between UCSD and Caltech it was observed that multiple GridFTP clients increased total throughput of the system but reduced individual server throughput. Improvements of individual client efficiency will be investigated. It was concluded that BeStMan is scalable and throughput is satisfactory.

Network for the LHC and HEP by **Barczyk Arthur**: the speaker covered in an initial part of the talk the next generation of network technologies. Providers are ready for 40G & 100G Ethernet with current throughput ~25Gbps (40Gps by 2011). OTN (Optical Transport Network) is the baseline of the next network transport layer and, with transport error recovery functionality, a signal may be transported at 40Gbps over 1650km without regeneration. Optical and packet networks will be combined in hybrid networks. The trend is to understand network infrastructure as a service with related requirements of bandwidth guarantees, traffic isolation (secure end-to-end connection), data caching, remote access/control and quality guaranties (availability, reliability, etc.). HEP may profit from dynamic bandwidth provisioning which is based on separation of high impact data from the rest of the traffic in user-specific end-to-end circuits. Various projects (ESnet SDN/OSCARS, AuthBAHN, etc.) and applications (e.g. FDT) based on mentioned principles have been discussed.

Wide Area Networks for HEP in the LHC Era by **Harvey Newman**: a significant part of the talk was focused on presenting a global view of the world network and an overview of continental and transoceanic network infrastructures. There is recognized strong worldwide pressure on bandwidth and traffic performance as key principles of countries' development (video+mobile traffic, Web2&3, SOA, e-banking/learning/health etc.) where HEP networking projects are performing as an engine of inter-regional connectivity. In general, very dynamic growth of broadband internet and connectivity is recognized through all involved continents. 10Gbps links are well-established and activities are moving towards 100Gbps. The progress of dark fibre installation across European countries was discussed. The DYNES project was presented. This has the aim to build infrastructures based on Inter-Domain Circuit protocol for large and long-distance scientific data flows and to extend such connectivity over ~40 US universities, as well as address the impact of expected LHC data transfer growth.

1**00GE in the Wild – First Experience** by **Bruno Hoeft**: the talk was focused on the first experience with 100GbE links between Forschungszentrum Julich and Karlsruhe Institute of Technology (~250km). As the fibre infrastructure used an existing GasLine backbone (~450km of fibres). With additional optical (Huawei DWDM) and IP (Cisco CRS-3) equipment a test environment was built to initially test cross talking waves. The overall goal was to put the infrastructure into production, fill it with IP traffic and perform stress tests and stability validation. In the initial part few issues have been detected, such as observed light between DWDMs, Cisco configuration and interface errors(between DWDM and Cisco), so no PnP installation yet. All major issues are now resolved and some test scenarios were discussed in detail. GasLine is ready for 100GE.

The US-CMS Tier-1 Centre Network Evolving Toward 100Gb/s by **Philip Demar**: the aim of the talk was to present the US CMS Tier 1. A model of network traffic with the link to the CERN T0 was described with details of hardware resources and data traffic capacity. A core fabric upgrade on the US CMS T1 network was recently finished where the major change was to move from CISCO6509 to Nexus 7000 core switches. The new network layout was described as well as the infrastructure and quality of service monitoring. The perspectives of the Tier1 for 20013/14 was discussed from the starting position of the recent upgrade as well as work in progress (monitoring, increasing Gbps throughput, infrastructure, etc.). The conclusions from previous talks (establishing 10GbE, moving to 100GbE, end-to-end circuits etc.) have been confirmed by this work via production implementation.

Integration of virtual machines in the batch system at CERN by Ricardo Silva: Ricardo started his presentation with a description of the CERN Batch system (3.5 k nodes, 25k cores, up to 22k concurrent jobs, large fragmentation → 25 pools). The main motivation of virtualization activities is the decoupling of the job environment from the underlying hardware. The solution is based on clones of the reference node, the so-called Quattor-managed, "golden node" with a 24h lifetime. Using Quattor-managed images ensures configuration transparency for end users and conformity with co-existing physical nodes. A limited clone lifetime enables flexible configuration management. Population control is based on the Infrastructure Sharing Facility (ISF) product (from Platform) coupled with the LSF batch management system. VM provisioning uses ISF and/or Open Nebula, both scalable over ~10k VMs. VM images are distributed by P2P (selected after comparison with SCP). In summary, the present infrastructure is the basis of a cloud with Batch as first customer. The number of VMs in the Batch system will keep increasing and scalability performance will be monitored.

WNoDeS, a tool for integrated Grid/Cloud access and computing farm virtualization by Alessandro Italiano: the Worker Nodes on Demand Service is an INFN-developed architecture and allows to dynamically allocate virtual resources from a common pool. The architecture offers various authentication technologies for user resource request access which is processed by an authentication gateway. The gateway provides certificates for next request processing. The authorization layer is based on ARGUS. Requests are processed by an interface layer (interfaces to grid, cloud, local resources). Resource allocation is based on LRMS. At the VM layer a virtual resource is instantiated. The interface layer was discussed in more details. In the local interface a virtual worker node is instantiated on-demand just to execute a batch job. The grid interface integrates gLite middleware. The cloud computing interface is based on the OCCI API (accessible programmatically or via a web-based application for user convenience). There is an integrated Virtual Interactive pool (VIP) interface which is a command line interface for WNoDeS, mostly suitable for interactive analysis or software development as it provides customized dedicated resource (previous interfaces providing resource dynamically on demand). Operation experiences were discussed. WNoDeS is working and providing shared and optimized computing resource handling.

Virtualization of PBS Jobs by Pau Tallada Crespi: at the PIC (Port d'Informació Científica) institute they were testing virtualization integrated into a PBS system. After some work, they concluded that it is possible and that their computing infrastructure may be shared for standard as well virtualized jobs. Virtualization for analysis jobs of supported experiments (HEP, astro, bio etc.) were tested and the results are under investigation. In general the experience is positive. There is a lot of work ahead for PIC's team as they integrate solutions with GRID tools, pilot jobs and gain a deeper understanding of last year's experience. As hypervisors, Xen and KVM were tested.

CernVM: Minimal Maintenance Approach to the Virtualization by Predrag Buncic: the CernVM project started in 2008 and undergoes regular updates. The main idea behind CernVM is to provide a framework and tools for the simple launching of image on various platforms and fast and simple contextualization. CernVM images are used around the entire world. Images (~250 MB) are based on a minimal Linux OS which has less frequent need of update compared to standard SLC5. rBuilder is used to build these images. Analysis software changes frequently and if it would be part of image, frequent image changes with related certification procedures and redistribution would be needed. The solution is to use CernVM-FS, a read-only, network (HTTP) file system optimized for efficient software delivery. This FS is generally deployable on virtual as well as physical machines. The scalability of such an http-based file system is going to be addressed by deployment of proxy servers at strategic locations. Finally they discussed a summary of the improvements in various CernVM componentss (e.g. security).

Trusted Virtual machine Images by Tony Cass: virtualization offers plenty of advantages (e.g. simplifying fabric management). Nevertheless there are some concerns (e.g. performance penalty) which come with massive usage of virtualization solutions. One frequently-discussed problem is that of virtual images deployment to multiple sites in a trustable way and with proper isolation so as not to endanger local site security. One possible solution is to establish a policy for trusted image generation, to document contextualization methods and to enable exchange of trusted images between sites. HEPix has created a virtualization working group; it is looking for agreement to these problematic questions and establishing a framework of policies and methods to enable the distribution of trusted virtual images. In general, contextualization should be limited to image interfacing to local infrastructure. A catalogue of images may help to simplify image approval by sites (possible cooperation with StratusLab project in this area). Most current implementations integrate to a local workload management system. In principle VM images could connect directly to VO pilot job infrastructure.

Virtualization Provisioning & Centralized management with iSCSI: the presented solution is not coming from a dedicated IT department but from part of the BNL physicist support groups. No cloud and/or virtualization as a general service is provided. They are rather focused on improving existing and future hardware utilization (why is a single service occupying an 8 core node?). First experience with virtualization started 3 years ago (RHEL 5.2, Xen, non-critical services). Currently they are using RHEL 5.5 with Xen & KVM as VM framework. Provisioning is managed by RHEL Kickstart & Satellite, PXE and custom scripts. The system is in place and working but facing problems with configuration complexity and scalability without centralized management tools (e.g. one tool/technology should handle storage configuration for all use cases). The management should be improved by introducing dedicated technologies such as iSCSI for disk array consolidation, Puppet for configuration tasks, Cobber/Koan for provisioning. Technical details of the implemented iSCSI solution (based on Nexsan SATABoy) were presented. In future they expect usage of RHEL6 with KVM.

Virtualization for the LHCb Online system by Zio Renato: the speaker presented the LHCb virtualization project, focused on detector control systems, excluding data systems, with the aim of optimising hardware utilization. Virtualization includes terminal services (SSH gateways, remote desktops, etc.), web services, infrastructure services (DNS, Firewalls etc.) and control PCs (dedicated hardware including SCADA, CANBUS etc.). Virtualization started with an evaluation of Xen, KVM, Vmware and Hyper-V as virtualization frameworks. Finally Hyper-V was selected and deployed on dedicated hardware. The architecture of the solution and the hardware were described. The Linux server installation is managed by Quattor with post-installation enhancements (the machine should be ready in ~10 min). A couple of issues have to be addressed in the near future, such as missing virtualization of PCI cards, licence-dependency on hardware (e.g. for PVSS) etc.. The first goals of LHCb online systems virtualization have been achieved: the system is in production and the hardware consolidated.

Establishing Applicability of SSD to LHC Tier-2 Hardware Configuration by Sam Skipsey: this talk was about a comparison between SSD[9] storage and RAID solutions for HEP computing. HEP computing is characterized as I/O intensive, with single-threaded applications using multi-core hardware, but not multi-threaded data storage. A legitimate question is why not to use use the faster SSD technology rather than HDDs. A test setup of worker nodes at Glasgow was presented. PCs with SSDs have been used because an expensive SSD server would not be adopted by Tier-2s due to budget limitations. They measured disk I/O, seek and throughput rates for various RAID (HHD based) and SSD solutions. From the test results it was shown that the RAID solution is more efficient. Also remote I/O tests showed advantages of the RAID solution. The work finished with the conclusion that it is not yet efficient to invest to SSD in HEP. They will perform a new review of this conclusion in about 2 years.

LHC Data Analysis using NFSv4.1 (pNFS): a detailed evaluation by Eves Kemp: NFS v4.1 (with pNFS clients) usage in data analysis was discussed. dCahce is used for data management. An NFSv4.1 + dCache model was presented from the infrastructure and user points of view. NFS4.1 is a promising solution because it is based on stable foundations with an industry backend, international standards, proper authentication and authorization, clients on various platforms (Linux, Windows, Solaris), etc. A testbed in the DESY GridLab was presented, resembling a 'typical' HEP analysis cluster and network infrastructure and dCache storage. The main testbed characteristics are: server->clients on 40Gbps links, one pool->client ~5-6Gbps, disk RAID->local /dev/null ~300MBps and final stream serv_disk->network->client /dev/null between 1.5 and 2.5 GBps. First simple I/O tests (reading file to /dev/null, no caching) showed that NFS behaves better than dCap up to certain limit (more investigation needed); later stability showed a few problems but in general it behaves satisfactorily. Two additional tests, based on ATLAS HammerCloud (standard ATLAS application to test site performance) and CMS data analysis, have been performed. In the ATLAS test one sees reasonable results but the comparison with dCap has still to be done. In CMS test one sees less traffic for NFS than with dCap but some comparison is still in progress. The last test was based on reading of ROOT files compiled with the latest root version (5.27.06) compiled with dCap support. During the test various scenarios were tested: reading via NFS and dCap, reading with 60MB or 0MB TreeCache, various numbers of jobs running in parallel, read out of all root file branches or only 2. If the files are not cache-optimized

[9] Solid State Disc

NSF brings significant advantage in file readout. The ROOT results are preliminary as they have not yet consulted with ROOT experts. There is ongoing joint effort (CERN, FNAL, DESY) to provide an SL5 kernel with NFS 4.1 (pNFS).

Storage Service Developments at CERN by D.Duellmann: this CERN storage services talk was focused on one year of LHC data-taking and consisted of a review and a presentation of future storage strategy. From 2008 the 40 minor Castor releases and one stable release 2.1.9 were provided and there was the SRM 2.9 release which addressed scalability, new monitoring, etc. Today we see a very steep increase in the amounts of data delivered to Castor. Close monitoring is vital. Castor generates large amounts of log information and new monitoring gathers all information in one monitoring repository, deriving basic monitoring metrics as well as higher level metrics and visualizes them (more details on poster). Some examples of monitoring visualization were shown and data access optimization was discussed. Various studies of I/O requests by analysis jobs (i.e. using ROOT files) lead to the conclusion that client cache logic helps I/O performance and should be part of ROOT. Castor serves as storage of data coming from the detectors as well as the provision of data to analysis clusters and it is in this second area that the main activities will be focused in 2011. The vision is to provide for analysis purposes a separate disk pool (the EOS project) with Castor data replication, i.e. to use Castor as the T0 archive and serve analysis from replica.

Stability, Cost, Performance of BlueArc Titan and Mercury servers by Michael Ernst: BNL laboratory is supporting storage for various projects (RHIC, ATLAS, etc.). For a long time it used NFS SAN storage but with increasing capacity demand (driven by ATLAS) this SAN solution became very expensive. A solution based on BlueArc Titan 3000 series servers was introduced and worked well, but still at relatively high cost. More BlueArc devices were evaluated with the aim to find the optimal ratio between price and performance. Mercury50 and Mercury100 servers were tested. From a performance comparison with Titan servers it was concluded that Mercury provided acceptable performance in similar workload conditions. In general, Titan is very suitable for storage systems but in case of server underutilization (because both Titan and Mercury have connections limited to 128LUNs) they may be replaced with ~50% cheaper Mercury servers, without dramatic impact on overall storage unit performance.

Operating a Large dCache Storage Element at Tier-1: Issues and Solutions by Ofer Rind: the aim of talk was to share long term experience with large dCache storage systems maintained by BNL. Some parameters of the storage system: 4.5PB on ~100 Sun/Nexsan servers, dCache 1.9.4, ~23.5k tape slots etc. The operational issues (and workarounds) were presented: access timeouts due to uneven data/space distribution; disk pools overloading with staging processes; etc. Remaining areas of interest are:

- storage allocation to handle pool growth and data distribution
- availability of heavily accessed files
- help user improve various aspects of usability: the dccp wrapper (user input check and pre-translation before PNFS ID lookup), Local Site Mover (additional site-defined abstraction layer for transfer), priority stager (facilitate the staging of data) and tape handling (file retrieval handled by ERADAT scheduler).

From an operational point of view, the speaker highlighted the key functionality of monitoring (Ganglia and Nagios for alarms) and the highly responsive support (with SMOD). They are implementing development Agile Methodology (Scrum), oriented on results. In future they plan to continue storage performance enhancement in parallel with addressing known issues.

Prague Tier-2 Storage Experience by Tomas Kouba: the presentation started with an overview of Prague tier 2 evolution (supporting D0, Atlas, Auger and Star HEP experiments). From the experience some highlights were discussed: high reliability of the chosen HP EVA 600 system (only 6 disks out of 112 replaced in 3 years); with Overland Ultamus 4800 they managed safe running but reliability was affected by confusing false alarms, some hardware failures (see below) and forced firmware upgrades. Overland Ultamus 1200 was considered as not very reliable as controller failure caused data loss. Throughput measurements for various used storage solution were shown (no specific conclusions). Requirements of latest new storage tenders were discussed. The storage was designed with the goal of building low-cost “home-made”

NAS servers. This was achieved but they learned that the disk has to be carefully. Their conclusion was that it is possible to build low-cost selected but too problematic for larger scale systems. In future they will provide such solutions only on demand for small working groups and for larger tenders they will return to reliable enterprise solutions.

Joining the Petabyte Club with Direct Attached Storage by Stephan Wiesand: in talk the speaker shared his experience from running Direct Attached Storage (DAS) at Petabyte scale. At DESY, Lustre and dCache are used for bulk data (experimental data, LQCD simulation, ntuples) and AFS for the rest (home directories, work group space, etc.). He highlighted the advantages of DAS, comparing it to SAN and NAS: lower cost, acceptable performance, simplicity, possibility of incremental growth and rapid procurement. All their dCache/Lustre/AFS instances are built from standardized 'bricks' (OSS/Pool Node/fileserver + JBOD, running SL5x64). Performance was tested with the ATLAS HammerCloud test with positive conclusions including some fine tuning. System is reliable as over past 3 years only ~15 minor incidents/PB/year and less than 2 serious incidents/PB/year (storage node goes down) but no data loss during this period. Operational aspects were explained: service is provided by a small team of highly experienced staff, using cheap test or spare units to replace failing hosts which is faster than any kind of contract. Their advice is to purchase complete systems from single vendor, including drivers; check controller logs and status regularly (e.g. weekly), keep firmware up to date, use RAID-6.

INFN CNAF Tier-1 Storage and Data Management Systems for the LHC Experiments Data Taking by Elisabetta Ronchieri: first the speaker introduced the INFN Tier-1 at CNAF - currently 66k cores, 6.6 PB of disk space, 6.6 PB on tapes. She then introduced the GEMSS solution developed at the CNAF. GEMSS stands for Grid Enabled Mass Storage system and represents an innovative and complete HSM solution with focus on management effort minimization (currently 2 FTEs manage the full system). Building blocks of the setup are:

- GPFS: disk-storage software infrastructure (enterprise solution from IBM)
- TSM: tape management system (enterprise solution from IBM)
- StoRM: SRM service for direct access to storage resources using various protocols (e.g. grid protocols). In-house development at CNAF to glue together GPFS and TSM.

The current GEMSS consists of StoRM 1.5, TSM 6.2, GPFS 3.2.1-23 and it is now used by all LHC as well as by some non-LHC experiments (e.g. Argo, Virgo, etc.). Various performance data were discussed. The most intensive throughput (hourly peak ~700MBps) from all sites to CNAF since LHC data taking started was in August 2010. GEMSS sustained such pressure well. The most intensive disk activities were run by ATLAS: access from computing farms (GPFS service) peaked at ~900MBps, storage access on WAN peaked at >3.5GBps. TSM HMS tape system served peak throughputs of 500MBps with 99% efficiency. In general, last year GEMSS performance and scalability experience is considered as very satisfactory and promising for future data taking.

How to handle data at Petabyte scale: PIC's experience by Francisco Martinez: this talk focused on role of data storage service for HEP (ATLAS, CMS, LHCb, Magic, PAUS etc) and non-HEP (e.g. neuroscience) projects. The talk consisted of a description of the centre and their experiences. Resource may be characterized by following overview: ~2500 cores, 3.4PB of disk storage, 3.6PB of tape storage, dCache for disk management (~60 disk servers for 7 VOs), Enstore for tape management (28 tape servers, full decoupling disk and tapes as no tape directly connected to disk server, disk-tape interaction driven by pattern. They use Enstore scheduling for as many requests as possible from dCache with the aim that a tape drive is the only performance limiting factor. They use Torque/Maui for computing management (in-house procedures for draining and urgent batch reboot). Puppet is used for configuration management (600 nodes configured, ~30 different profiles, still some stuff configured by yaim).

Tape Archive Challenges When Approaching Exabyte-scale by German Cancio: German described CERN archive in numbers: 5 CASTOR stager instances, 7 libraries, ~26PB of data on tape, ~160M files on tape, average file size ~195MB, etc. Capacity growth was discussed. The CERN Tape Archive was planned for 15PB/year but current rates are higher and in future a rate of ~23-25 PB/year (100-120M files/year) will have to be managed by the archive infrastructure. Fortunately tape capacity will continue at least to double every ~2-3 years with the next drive generation expected in 2011. Three core operations (writing, migration and reading) were discussed. The writing process is well understood and performing well (aggregate transfer rate 3-5 GBps so not a problem) although with high overhead for writing small files. Migration of data from 500GB to 1TB tapes took around a year (45k tapes, ~40 tape drives) at a cost of 1.5 FTE; archive maintenance activities (including tape verification of ~5.5k tapes in 2 months using up to 10 drives) consume non-negligible resources. Readout of files distributed across various tapes is an issue because of the risky and time consuming operation of tapes; (un)mounting must be repeated several times. This decreases the performance of the average read operation to ~35MBps. Policy-driven tape mounting and/or file pre-staging will help in the short term, but long term scaling must be addressed by a new design (e.g. data-set-based architecture) which should replace the current file-based HSM.

Tape Library Virtualization and Tape Advanced Reporting by Dorin Daniel Lobontu: in this presentation the speaker described the GridKa storage system, with its TSM technology as backend for dCache storage management. Also he presented the TSM-dCache link virtualization with ERMM technology and finally their reporting tool. The GridKa centre runs 75 file servers in 2 dCache installations with 2 tape libraries, 8 PB of disk storage, ~10PB tape storage. Without ERRM on the TSM server, one path for every storage agent (link to dCache) and every tape must be defined. When ERMM is used the TSM servers interface to only one library, ERMM takes over all library management, handling the paths from storage agents to libraries and providing dynamic load balancing among the drives. Advance reporting is based on a collector application monitoring the storage system as well as getting information about ERRM events. All information is stored in a MySQL backend with the possibility to build a complete history of the drive access records. This information is processed by a statistics generator implemented in Perl and passed to a plot generator for display as s time series (hour, day, week, month, year).

Evaluation of Cluster File Systems for the LHCb Experiment by Rainer Schwemmer: the file system solution for the LHCb online system was discussed. The LHCb online system consists of ~70 workstation, ~200 control PCs, ~300 readout boards, ~950 (soon ~1500) filter farm nodes. For storage common a SAN connection to disk arrays via Fibre Channel is used with NFS/SAMBA export to Linux/Windows nodes. There are several sources of high load such as writing experimental data and starting the High Level Trigger (a short but extremely high burst of metadata operation). Various file systems (Lustre, Gluster +EXT4, GFS, IBRIX, Storenext, NetApp) were tested against a list of requirements running on already-installed hardware/ The tests included coping with DAQ I/O, High Availability capability, robustness to failures, etc.. It is recognized that data acquisition performances are no longer challenging for modern file systems but the emphasis is more on architecture limitations (e.g. user/group areas are shared with experimental data areas – how to make sure that user I/O does not negatively impact data taking I/O?) and system manageability. It turns out that failing system calls during trigger start cause complication (40k access calls per instance) with which none of the tested file systems can cope. At the moment nothing can be done with trigger software or deployment procedure. The workaround is to use a proxy-like file system, e.g. FUSE. FUSE stores information in cache and the cache responds to problematic system calls instead of the file system. One would conclude that manageability is becoming important file system selection parameter and that that improper software behaviour may significantly deteriorate even a superb file system.

Exabyte Scale Storage at CERN by Andreas Jochim Peters: the talk was focused on a discussion of the EOS project as an alternative to hierarchical storage (HSM) for data analysis needs. An increasing number of analysis jobs requiring experimental data approaches the limitation of a HSM model for such purposes. Alternative tier models consist of separate analysis, archive and tape pools with random + sequential RW access on the analysis pool and privileged sequential read and write-once pools. Such a model is applied in the EOS project. The goal is to provide a disk pool for analysis with POSIX-like RW file access, hierarchical namespaces, strong authentication, data redundancy, dynamic scaling without replacement downtime, etc. In EOS, storage is organized in single disks (JBODs, no RAID arrays) with network RAID within node groups. It is self healing (from a client point of view all files are always readable & writable) and supports online file migration, etc. Currently a prototype is under development which keeps only namespace index views in memory and allows ~20GB/view for 10^9 files (current implementation 1GB/~1M files). Other features such replicas, self healing, high availability, etc. were discussed. Various tests were done with the EOS prototype: file creation (10M files, 22 ROOT clients at 1kHz), file read (100M read open, 350 ROOT clients at 7kHz), scalability, HammerCloud tests on the ATLAS testbed, etc. This project will continue with development and testing on larger testbeds and by the middle of November there should be an EOS ATLAS instance open to ATLAS users.

From detailed analysis of IO pattern of the HEP applications to benchmark of new storage solutions by Jiri Horky: this talk is about an innovative approach to storage system benchmarking for the needs of HEP. Main aim of storage array and file system benchmarking is to distinguish at least between two solutions. But to install for all possible test configurations all HEP software (e.g. ATHENA) is not an easy, if even possible, task. The speaker believes in an idea of a 'synthetic' benchmark tool able to trace the I/O operations of an analysis job and repeat the same I/O sequence on any tested system. Record collection is based on STRACE which collects information about job access to a hard drive. It was challenging to find an algorithm for mapping file descriptors to real files. The output of STRACE is processed by a C-based analyzer with PyQt-based visualization using Matplotlib. Data access analysis results were discussed showing many of the details (e.g. access pattern in ROOT). He demonstrated how to use I/O information: reconstructed I/O sequences have been repeated on the same system as have been recorded (i.e. all calls which failed/succeed in an analysis job fail/succeed as well in a synthetic repetition). It was shown that I/O repetition provides the same behaviour as an analysis job if the number of simulated processes is less/equal than number of cores (i.e. on an 8 core node, the I/O simulation copes with 8 threaded applications). Such simulation was applied also on a testbbed where it was not possible to run analysis applications. These benchmark tests are still under investigation. In conclusion, an easy-to-use I/O profiler was developed and for access to software, the speaker should be contacted.

End to End Storage Performance Measurement by R.Voicu: this talk was oriented on the end-to-end performance monitoring of storage systems which in general may be characterized as very heterogeneous (storage, servers, network, etc). Performance of such environment very much depends on the technologies used and the architecture of the solution. End to end monitoring will provide an overview of system status and help in problem debugging. Monitoring was built within the ALICE environment and is based on monitoring data collection by MonAlisa agents. Such data are aggregated per site (i.e. some organic unit). Aggregation of data over all the sites had to be implemented. By using Fast Data Transfer (FDT) protocol bandwidth may be (and was) controlled between all the ALICE grid sites. Various details of FDT and the FDT hook of MonAlisa monitoring were discussed as well as the measured metrics (I/O, CPU, bandwidth etc.). In future an extension of the test is planned (as the full chain 'local disk->storage->network->storage->local disk' is not yet monitored) including alarm policy implementation for critical situations.

An Analysis of Bulk Data Movement Patterns in Large Scale Scientific Collaborations by Phil Demar: globally distributed collaborations are contributing to LHC data analysis using distributed data centres for distributed computing. Transport protocols like TCP do not cope with HEP-like large, 'bulk' data transfers and thus various parallel data transfer tools (e.g. GridFTP) are used. This talk discussed measurement of bulk data transfer from Fermilab to the top 100 sites from 24 subnets and from these 100 sites back to Fermilab. Single throughputs (transfer via TCP) and aggregated throughputs (parallel data transmission via GridFTP) were analyzed and various patterns discussed. Measurements were done with *Fermilab Flow Data Collection and Analysis System* (Cisco NetFlow + Silk-command line tools for flow record processing; better granularity than SNMP based measurements). 2.221x10^12 packets were analyzed (Round Trip Time in IN and OUT traffic and various packet statistics). Most average throughputs for Sites->Fermi are less than 10Mbps with a minimum of 0.207Mbps. For Fermi->Sites the slowest throughput was 0.135 Mbps. Correlation between average throughput and RTT was calculated and analyzed. In general, aggregated throughput to sites is higher (IN/OUT) than a single flow throughput and transmission from FNAL to the top 100 sites is better in the other direction.

Deployment and Operations of a High Availability Infrastructure for Relational Databases in a Heterogeneous Tier 1 Workload Environment by Carlos Gamboa: this talk was focused on an overview of the database services at BNL and discussion of some operational challenges: remote access to BNL database and database hardware migration. The main purpose of the BNL database service (the one dedicated to LHC) is to distribute and catalogue LHC data as well as replicate and store detector condition and calibration data. BNL, the US Tier 1 site for ATLAS VO, serves ~6PB of online storage using ~1.6k physical CPUs. Various applied cluster topologies were discussed. Service is monitored by Oracle Enterprise manager Grid Control, nagios and ganglia. The most serious operational problem was the long resolution time of a user reconstruction job. Various tests were done with the target to find better client configuration parameters. With fine tuning, a ~50% decrease of resolution time was achieved. With further study a problem was found in the underlying library which was responsible for 4 DB calls/retrieved row. This problem was resolved in cooperation with ATLAS and the COOL team. The data migration part was not discussed during the presentation due to lack of time. Briefly, for a migration to minimize DB service downtime, Dataguard and Transportable Tablespaces (both Oracle technologies) will be used.

Software Engineering and Databases

CORAL: Andrea described how the ATLAS High Level Trigger software accesses the condition DB through CORAL as a middle tier in front of MySQL first and Oracle later on. Motivated by security concerns: Oracle ports are behind a firewall; a CORAL port speaking a CORAL protocol (with X509 auth) can be opened on the firewall. Scalability is insured by the Coral Server multiplexing many clients on a limited number of Oracle connections; Coral Server proxies can be used to scale up the system. A Coral Server was deployed at Point 1 for ATLAS in Oct 2009, and has been working fine ever since.

SciDB: Databases have been largely used in HEP, though they were originally a 'less-than-good' match for many reasons. However, things have evolved. SciDB has been conceived to address the need for a data management system able to handle data arrays, version control, etc. The product is in its very early development stages; some key features include: native support for nested and sparse data arrays, storage in chunks; AQL - Array Query Language; Linear Algebra, etc. A first beta is due in April 2011. There is a pre-alpha version with some real LHC data and functional tests have been perfomed. A major target client of SciDB is LSST (Large Synoptic Survey Telescope, O(100 PB) data with O(1 PB) metadata).

HTML5: An interesting application of HTML5 and WebSockets from Google to implement a web-based interface to the CMS Condition Data in push mode, much faster than classical HTML4 web interfaces that typically pull data from server in a less efficient way.

Google Web Toolkit for HEP: Google Web Toolkit (GWT) is a tool to write "Web 2.0" applications. Features include widgets, rich GUI, browser independency (even across mobile devices). Based on a Java->JavaScript compiler, which turns out to be quite efficient, it leverages existing knowledge and experience on Java programming. Use case: the DAQ GUI for the EXO experiment, which was written in Java, containing dialogs, plots, some drag and drop functionalities, etc. In particular, in the future, plots could be rendered in 3D using WebGL. GWT is also being used in other toolkits, e.g. pyjama, a 'port' of GWT for python. Issues: rapidly evolving, not 100% of JDK is (yet) available.

Collaborative Tools

HEP Outreach, Inreach and Web 2.0 by **Steven Goldfarb, University of Michigan and ATLAS**: Overview about the Web 2.0 initiatives in Internal/External communications in ATLAS. These efforts were driven by the need of fast information updates for multiple users. The information is presented in public portals, using info streams, social networks and site blogs. Usage foreseen to continue to expand and there is a need of optimizing by focusing on effective and high visibility sites and a move to an Open Source Content Management System (Drupal)

CMS Centers Worldwide by Lucas Taylor, Fermilab and CMS: Update about the project (presented in CHEP'09) that consists in a set of standard CMS remote control centres with almost permanent video presence, status displays, detector quality monitoring allowing remote shifts. The 1st remote centre was set up in FNAL in 2007 followed by CMS CC in CERN in 2008. The expansion was quite large going from 16 centres in 2009 to 55 centres in 2010. The typical cost of a CMS centre is about 12KCHF for a generic standard installation.

ATLAS Live: Collaboration Information Streams by **Steven Goldfarb, University of Michigan and ATLAS:** description of a project launched in January 2010 to address a problem concerning the difficulties of the collaboration in finding and/or maintaining important information in the web and in disseminating it externally. The actions to address the issues involved the Installation of monitors all over CERN and the use of the CERN Infrastructure of webcast and streaming. An ATLAS team is responsible for creating the content is that is then disseminated at CERN and elsewhere. All information is integrated automatically with the existing information repositories (namely CDS, Indico, Picasa, etc.)

Physicists get Inspired: Inspire Project and Grid Applications by **Jukka Klem, CERN:** Inspire Project started in 2007 and is the next generation information system that is based in Invenio (Digital Library SW) and Spires (former HEP information system). It provides an information source for HEP for open access full text articles, experimental notes, etc. It currently gathers info from 4 labs: CERN, FNAL, SLAC and DESY. Inspire is part of the D4Sience initiative that inks Inspire and other communities to build a knowledge ecosystem. The main uses cases are: document OCR, Full text editing and bibliometrics (using Hirsch Index).

Perspective of User Support for the CMS collaboration by **Sudhir Malik, FNAL/University of Nebraska and CMS:** CMS is a project with a lifetime of 30 years and represents a huge collaboration formed by 40 countries, more than 200 institutes and 3500 people. It presents several challenges to the user support activity within the collaboration (background, affiliation, etc.). It needs an effort of organization and since there are no dedicated personnel for the activity the key of success is the cycle between people where the collaborative effort is shared.

Glance Information system for ATLAS by **Laura Moraes, UFRJ and ATLAS:** GLANCE is a general framework to access various existing applications to support experiment management and collaboration members with the main goal of decentralising the ATLAS activities. Information that can be accessed in the framework includes technologies (models and inventories), members, talks, papers, appointments, alarms, etc. The system interacts with current repositories based in Oracle/MS SQL MySQL and allows to import/export of different formats.

EVO (Enabling Virtual Organizations) by **Philippe Galvez, Caltech and CMS:** Update about the EVO videoconferencing system usage and deployment for the LHC. EVO holds currently 2500 meetings per month, where ATLAS (with almost 1100 meetings) and CMS (more than 1000 meetings) are the biggest users. EVO infrastructure consists basically in 60 servers deployed in the LHC network worldwide, to support around 850 sites simultaneously connected. The usage is estimated to grow around 60% per year.

University of Michigan & CERN Lecture Archiving System by Jeremy Herr, University of Michigan and CERN: The project is based on the CERN and UofM agreement with the goal of having a Lecture Archiving system available for CERN users integrated in the CERN environment. It has developed into a comprehensive lecture archiving based on the UofM system that provides: recording, processing, archiving, publishing monitoring and analytics. A recording manager has also been developed in Indico. The system also uses Micala (http://micala.sourceforge.net) that contains all the system's monitoring info.

Towards a New PDG Computing System by **Juerg Beringer, Lawrence Berkeley National Laboratory:** PDG is an international collaboration charged with summarizing Particle Physics, as well as related areas of Cosmology and Astrophysics. It currently gathers material from 176 authors from 21 countries and 108 institutions and 700 consultants in the particle physics community, coordinated by the PDG group at LBNL An update about the new computing model was presented to face new challenges

AbiWord and AbiCollab – Real Time Collaborative Document Creation by **Martin SEVIOR, University of Melbourne and Abisource B.V.:** AbiCollab allows real time collaboration between arbitrary numbers of AbiWord sessions and allows real-time document creation. It operates in a decentralized peer-to-peer network. It also provides abicollab.net webservice – central document repository (a company has been created to commercialize abicollab).

BOF sessions

WLCG Operations: This was a special lunch time session convened by Jamie Shiers to discuss perceived operational problems with WLCG. It was lightly attended and largely animated by Shiers asking questions to the audience which presumably means there are no major outstanding operational issues, at least among those present in Taipei. Shiers asked in particular if there had been a noticeable difference since the transition from EGEE to EGI. Once again, little response from the audience. One issue raised by CMS is how to handle support grid sites where there is no "home" NGI and this is one area where coordination between WLCG and EGI could be useful since both have the same issue. Is there a need for another WLCG workshop? Are things going well enough to forgo this until the next CHEP (spring 2012 in New York)? Little reaction other than a suggestion of one per year, even if there are no particular issues at the time.

WLCG Service Coordination: chaired by Maria Girone, attended by some 60 people and following on from her talk on WLCG Operations in the Data Processing and Analysis Track. The main points of discussion were on prolonged site downtimes, better categories for the GGUS ticketing system (most are classified as "other") and desirable enhancements such as escalation from TEAM (shared by "shifters") to ALARM tickets. As far as service incidents are concerned, these can be broken down as follows: Infrastructure Services: rather constant number of problems, at least some of which are probably unavoidable (human error, power and cooling etc.), Network Services: although typically degradations some actions are required (and underway) to improve expert involvement and problem resolution, Middleware Services: for which there are very few incidents resulting in a report (at least partly as the experiments have learned to work around any major issues) and finally Storage & Database Services: typically complex problems that sometimes cannot be resolved within a day or so and which dominate. Other topics that were discussed but not concluded on included the evolving strategies of ATLAS and others for handling detector conditions data and shared software areas instability problems – the source of regular discussions in the WLCG operations meetings in recent weeks.

GPUs: this BOF on Thursday was very popular, including some prepared presentations, lots of interaction and the suggestion that openlab should provide a mail list on the subject.

Poster Sessions

As usual, although some traditional plenary time had been scheduled for additional parallel sessions, there were far too many submissions to be given as oral presentations so 197 of these were scheduled as posters[10], displayed in two batches of 100 each for 2 days each. Between the two sessions, some 15-20% of the poster slots were empty, usually attributed to authors unwilling or unable to attend "only" to show a poster. Posters were displayed on corridors between the main room and the coffee area, so very accessible, and the morning coffee breaks were extended to one hour to permit attendees to view them and interact with the authors.

Exhibition

Each of the 7 sponsors was given a booth but only a few were known by most participants from outside Taiwan.

Summary Talks

Online Computing: 35 talks, 24 posters. Dominated by LHC DAQ of course. All experiments' DAQ worked well, leading to fast production of physics results. What will happen during heavy ion run with the other experiments also writing to Tier 0? Seeking more automation, less shifters needed, more efficient, faster recovery from problems. Studies into data quality monitoring.

Event Processing: 45 talks, 31 posters; 49% LHC but 51% non-LHC. GEANT4 is mature and ALICE is now getting on board as well. Simulation has been able to provide an amazingly accurate description of LHC data. Analysis hot topics are data tiers and sampling of data for analysis; and the organisation of analysis tools and code. Prompt processing at tier 0 (ATLAS and CMS) works. New experiments (e.g. at FAIR) should take advantage of current experiment frameworks but the analysis challenges of FAIR experiments are beyond those of LHC. Reconstruction works today, how will we cope with increasing pile-up? Many reports on using GPUs and multi-core systems and a few concerning vectorisation of code.

Software Engineering, Data Storage and Databases: 33 talks, 24 posters. A very heterogeneous range of subjects covered – quality assurance, performance monitoring new technologies such as svn, CVMFS, SciDB, etc), databases, software re-cycling and data preservation. Summary – software frameworks for LHC are in good shape and other experiments should be able to benefit from this.

Distributed Processing and Analysis: 48 talks, 20 posters, most popular stream in terms of abstract submissions (130) so two parallel streams at times. A main theme was the successful processing and analysis in a distributed environment,

[10] Compared with 256 scheduled oral presentations and only 12 plenary talks

dominated of course by LHC. The message is positive, the computing models are mostly performing as expected. The success of the experiments relies on the success of the grid services and the sites. But the hardest problems take far longer to solve than target service levels. The other two main themes were architecture for future facilities such as FAIR, Belle 2 and Super B; and improvements in infrastructure and services for distributed computing. The new projects are using a tier structure but with apparently one layer fewer than LCG. Two new non-HEP projects, Fermi and JDEM, seem not to use grid-like schemes.

Computing Fabrics and Networking Technology (summarised by Tony Cass): 42 talks, 24 posters. Surprises – no talks on Lustre, GPFS or IPv6. Hardware is not reliable, commodity or otherwise – this statement from Bird's opening plenary was illustrated in several talks. Lots of talks on fabric management but are we not re-inventing the wheel? Quattor is no longer hot, Puppet seems to be taking its place. Deployments of upgrades, patches, new services are slow – another quote from Bird: several talks showed we have the mechanism so perhaps the problem is in communications and not in the technology. Yes, storage is an issue and there is lots of work going on in this area as shown in several talks and posters, but the various tools available today have proved that they work and we have stored and made accessible the first 6 months of LHC data. Lots of talks and posters on different aspects and uses of virtualisation. It was shown that 40Gb and 100Gb networks are a reality. Network bandwidth is there but we need to expect to have to pay for it. Still an ongoing discussion on the trust needed to run VMs in clouds.

Grid and Cloud Middleware (summarised by Markus Schulz): 38 talks, lots of active discussions but little of previous controversy. Shift from previous CHEPs - pilot jobs are fully established, virtualisation is entering serious large scale production use and there are more cloud production models than before. Schulz recommended Stephen Burke's presentation on lessons learned on grids although he admitted he did not agree with all of them. He also quoted Rob Quick's final lesson learned working on OSG – "there is no substitute for experience". Various monitoring and information system tools were presented as well as work on data management (for example an interesting talk on trying to scale SRM dCache with Terracotta which has not yet been successful). Various aspects of security were covered. On clouds, although STAR reported impressive production experience and there were various successful uses of Amazon EC2 clouds, other cloud initiatives are still at the starting gate and some may not get further. There was a particularly interesting example linking CernVM and Boinc via a CernVM Co-Pilot.

Collaborative Tools (summarised by Joao Fernandes): 11 talks and should have been 6 posters but only 3 were displayed. Their first session was dedicated to outreach (Web 2.0, ATLAS Live and CMS Worldwide) and new initiatives (Inspire). The second to tools (ATLAS Glance information system, EVO, Lecture archival scheme). The posters covered a couple of e-learning schemes and one on visualisation via HD videoconferencing.

Closing ceremony: no conference summary but a slide show of photos, some conference statistics and a presentation of the next CHEP (see below). And of course a vote of thanks for the organisers.

Social Programme

For those who know the conference chair, Dr. Simon Lin, the social programme was well and truly up to par. The Welcome Reception on Monday was held in the well-named Grand Hotel, a huge grand palace highly visible on a hill above the main road from the airport into town. Delegates passed through an arcade in front of the banquet hall showing off the work of traditional Taiwanese artisans before sitting down to a chinese banquet of 11 courses, interspersed with a traditional puppet show by a famous local company.

The next event, restricted to the IAC[11] and PC[12] was a dinner hosted by Simon at a top Japanese/Taiwanese fusion restaurant in town. Another 11 courses (or was it 12?), this time with a Japanese theme.

As is now traditional in CHEP, attendees were given Wednesday afternoon off and a choice of 4 half-day tours were organised. Although the weather continued to be as showery as it had been all week, this was an excellent opportunity for attendees to see a little of the local countryside and/or get some flavour of Taiwanese culture.

Finally came the Conference Banquet, held on the Thursday evening, again in a downtown hotel. Twelve courses this time, interspersed with a Taiwanese aboriginal troupe performing various dances.

Weather

Not usually part of such reports until HEPiX Lisbon (the ash cloud). This time the talking point was tropical hurricane Megi. It battered the Philippines at the start of CHEP week before moving north towards us. Taipei was rained on for the entire week and as Friday approached, there was rising concern on the effect of air travel between Taipei and Hong Kong in particular for the journey home. However on Friday, the storm lessened and the threat receded.

Next CHEP

The next instance of this series will be from May 21st to 25th, 2012, hosted by BNL and situated at the NYU campus in Greenwich Village, New York. See the preliminary web site at http://www.chep2012.org/. And bids were opened for CHEP 2013, scheduled to be held somewhere in Europe.

[11] International Advisory Committee

[12] Programme Committee