CHEP07 Trip Report

Contributors – James Casey, Eric Grancher, Manuel Guijarro, Michal Kwiatek, Giuseppe Lo Presti, Andrea Sciaba, Ulrich Schwickerath, Alan Silverman

Editor – Alan Silverman

Contents

1. Introduction

CHEP is a series of conferences devoted to Computing in High Energy and nuclear Physics. It is held approximately every 18 months in the cycle Europe – North America – Elsewhere. CHEP07 was held in Victoria, British Columbia from September 2nd to the 7th, the first time it has visited Canada. It was hosted by TRIUMF and the University of Victor and held in the Victoria Conference Centre next door to the elegant Empress Hotel, one of the major landmarks of Victoria. Some 460 people attended from all across the globe. The conference was sponsored by a number of commercial firms, notably IBM and SGI, and some academic and national organizations.

As in the previous CHEP in Mumbai in February 2006, the conference was preceded by a 2 day workshop on WLCG, during which the organisers were able to debug some last organizational glitches, in particular in the catering. Unfortunately the lack of power connections in the main room, used for plenaries in the mornings and 3 parallel sessions in the afternoons remained an unsolvable problem (there were some power connections in other rooms).

2. Plenaries

After the welcoming addresses by the main organiser, Randy Sobie of Victoria University, and directors of the two hosting organizations, the first talk of the conference was given by **Jim Virdee, the CMS spokesman, on the LHC** itself, both the machine and experiments. The slides of this and all talks, and videos of plenaries (check) are available on the CHEP07 web site via Indico. There were a total of around 450 papers and posters, split 50-50.

Virdee started with a review of the physics expected. Installation of the accelerator us complete and commissioning is ongoing. After a test at operating temperature, the warm-up of sector 7-8 has thrown up some problems but after an initial scare, these are currently not thought to be major and should not affect the current schedule although there is no slack if some major problem does arrive. Beam pipe should be closed by end March 2008, first beam injection end May and first collisions in July. In relation to beam tuning, he noted that LHC beam energy is the same as that of a Boeing 747. All the experiments are completing installation and starting commissioning using cosmics. CMS has lowered 10 of its 15 large pieces into the cavern and will lower the rest by the end of the year. He noted that the experiments were preparing dress rehearsal of their computing chains and he ended by relating expected ramp-up of LHC luminosity to the expected and hoped-for physics noting that quite a bit of luminosity will be needed to confirm the Higgs. In this respect he surprised some in the audience by noting how quickly the software would have to deal with event pile up.

The next speaker was **Les Robertson on WLCG progress** and its use by the experiments. WLCG now runs some 2.5 million jobs per month of which a significant number are not HEP but the target for beam conditions is 9 million. Among the baseline services, progress on stabilizing and introducing SRM 2.2 may be the largest remaining challenge. Overall, reliability and stability of the storage management system needs to continue to improve. Another positive step is that experiments are developing their own tests, often based on SAM, which are starting to replace the standard LCG tests in reporting, at least for the Tier 1 sites' performance; also more and more Tier 2 sites are appearing in the statistics. This performance data shows that site reliability is still an issue, especially many Tier 2 sites; that operational complexity is the weakest link; and that ramp-up and reliability of data transfer is perhaps the most worrying issue. Nevertheless, Les concluded that "we are getting there …. slowly".

Ian Fisk of Fermilab continued by reviewing of the higher layer of LHC experiment computing. He noted that the layered computing model of the experiments is rather complicated with a multiplicity of inter-site interconnections. The speaker claimed that the need for common interfaces and protocols in grid software also increases software complexity compared for example to LEP computing. All this complexity obviously affects reliability and overall robustness of the software. Happily, a review over time of performance as reported by the experiments' dashboard reporting schemes shows steady average improvement. However, as noted previously by Les, data transmission and access remains a challenge. ALICE and LHCb have pull-based job scheduling and ATLAS uses this mode also in one analysis mode but most ATLAS modes and all of CMS use resource matching to schedule analysis jobs. Using Monte-Carlo for production processing tests is ideal as Monte-Carlo time can be predicted. One of the worries is that more processes are needed to utilize the CPU capacity, it means that the scheduling facility will be used 4x more. Tape systems are used in HEP as write once / read many, while the tape companies have written it for write once / read ~never. He showed the required ramp-up of resources per tier needed over the coming year and stated that disk ramp-up in particular may be problematic on many sites.

Sylvain Chapland then surveyed DAQ and triggering systems for the LHC experiments, comparing the different requirements, the DAQ design challenges for each and the expected data flow rates. He covered all the LHC experiments but with a certain bias towards ALICE. He noted that by and large the experiments used common components for the triggering with a lot of custom electronics, common hardware and operating systems for the DAQ farms with customized layered software on top. Qin answer to the question that the DAQ system being so complex, how to avoid data corruption, he answered that data is checked in various phases of the DAQ.

The first morning of plenaries concluded with an intriguing talk from **SGI**, one of the two gold sponsors of the conference on **computer facility limitations**, the fact that more and more data centres are running out of capacity in power, space and weight for expansion. The US is even proposing a law which may penalize inefficiently powered centres in the future. He showed graphs of the expected evolution of components used in computing fabrics from the raw cost of copper, packaging of components, energy efficiency of power supplies and. In relation to the first, he advises customers with large amounts of cabling to lock their centres for fear of copper thieves, which has indeed occurred in places. Talking of available computer centre space, he noted some things to watch. Adopting blades may be vendor locking as there is no standard blade system. One tight packaging of a new technology from SGI was so heavy that a standard 19 inch rack would have been too heavy for most computer room floors and they had to use a wider rack to accommodate the weight per floor area. They have designed a more efficient power supply without a 48 Volt step-down but the result is a 1000 amp backplane! [In answer to a question about the possibility of doing away with this last step down, he noted that you still need to do one conversion, from A/C to D/C. Finally he noted that HEP is not alone in facing the limitations of computer centre capacity, the movie film industry faces the same challenge to digitize their archives. Hopefully Hollywood has the funds to enable firms such as SGI to resolve the challenges and we can reap the benefits.

Tuesday was opened by **Miron Livny from on the future on Grid Computing**. He started by declaring that, as a movement, grid computing has run its course – no more easy funding, no more an easy sell, etc. But distributed computing is here to stay and will continue to evolve but we must stay away from hype and the technologies of the moment in order to advance the state of the art; concentrate on fundamentals to deliver stable and scalable distributed computing. Specifically with OSG in mind, Livny believes the major challenges ahead are “soft”, to develop organizational and management structures while maintaining and evolving a software stack and operate and evolve a dependable and well-managed distributed computing facility. In summary, software quality must improve, especially security software; restructure organizations to separate software development from soft deployment; and add capabilities, especially for storage management and file management. In answer to a question on how to improve software quality, he suggested “stick to have you have, take small steps”.

From IBM, the second Gold Sponsor, **James Sexton gave a presentation on HPC at the Petascale and Beyond**. Although he believes Moore’s Law still holds, we are approaching fundamental limits, every time we want to make a step forward, we need to make workarounds and much more performance gain is coming from parallelism in the cores. IBM’s Blue Gene is an example of this and he showed some examples of Blue Gene in use. Currently we add cores to increase the power but writing programs for such computers is difficult. The fundamental limit is due to the fact that we reach the size of the atoms. Blue Gene relies on embedded chips (card with SPU, Ethernet and memory controllers), low frequency CPUs and good connections; 4000 cores in a single rack. Ratio between Blue Gene / Intel based racks (42 1U computers per rack) is 56/210. The largest Blue Gene has 64 racks, 130’000 cores. In the future he believes the memory footprint per processor will reduce because it cannot be affordable at current levels as the processor count rises. Also, I/O is not keeping up with CPU performance increases so getting the data in and out will be a problem. Going to exascale, power consumption must be reduced by decreasing frequency and/or complexity and other fundamental limits will be approached and very significant innovation will be required.

The first plenary on Wednesday was given by **Alberto Pace on Identity Management**. Closely linked to this is access management. Covering both of these there are three components - authentication, authorization and accounting – the AAA rules. Finally there is the concept of role-based access control which brings in the concept of groups having access rights. To implement these rights, an Identity Management Database is required with associated access tools. One account per individual or multiple? The latter tends to be less secure because often multiple accounts are intended to enable sharing which is inherently insecure. But if only one account, there should be a method for default working at low privilege but requesting higher level privileges for certain activities for limited periods. Finally Alberto presented plans to move from CERN's current multiple databases, multiple accounts to the model he had described and to extend that to a CERN Certification Authority which could be used in the grid context. In summary, Identity Management reduces complexity, increases security and, as a side benefit, should reduce administration costs.

Lawrence Field then discussed Grid Operability. What is the problem to be solved? For a start, users are in different organisations and administrative domains so create Virtual Organisations and mappings of these to local organizations or domains. But these created their own problems so people devised the notion of interoperability and interoperation, the former covering how to pass information, data, etc. between domains. In the short term one can do this by building gateways or translators and he showed examples of both but in the long term there should be standard interfaces. On the interoperations side, a study of the differences, for example between OSG and EGEE, can show how and where adopting a common solution can resolve particular interface problems. In the case of OSG and EGEE a task force as setup in Feb 2006 with the target of demonstrating interoperability by Nov. The GLUE schema for interoperability came out of this but here also Lawrence noted that standards should make interoperability issues easier to solve in the longer term and work continues on GLUE in the OGF[1] environment.

Frank Wuerthwein of San Diego described the benefits of grid technology for currently-running experiments. He gave examples showing that the both grid middleware and grid infrastructure had benefits. Examples of the first were the movement of data, for example data from PHENIX and STAR at BNL to collaborating sites; the development of the Fermigrid campus grid as used by CDF and D0 as well as CMS; the inclusion of the SQUID web interface to databases in the middleware has been adopted by CDF and which resulted in a 60% performance gain in some cases. Turning to infrastructure benefits, some experiments have integrated their legacy operations into the transparent operations schemes offered by grid technologies; this for example helps in retrying failed jobs. Examples here include H1 and ZEUS Monte Carlo production. Also, as further explained in a parallel session earlier in the week (see below), D0 focused more than 50% of their re-processing on the OSG.

Harvey Newman was next up, Networking for HEP and the Digital Divide. He presented graphs showing the recent growth of networking in general and by HEP in particular, including some saturation situations caused by massive transfers by certain experiments. He showed how links which the LHC experiments will rely on will increase in number and capacity over the coming years according to the agreed roadmap. The major backbones in both Europe (GEANT2) and the US (Internet 2) are moving or have plans to move from 10Gb/s to N times that; CANARIE is the first one to use 40Gb. There has also been a quiet revolution in making networks easier to use, demystifying the use of TCP, NFS 4 targeting larger-scale distributed file bases, etc. Tools are being developed and implemented to offer managed transfers – monitoring all aspects of the transfer to guarantee delivery delays, respect experiment priorities and so on. The leading example of this is the Mona

[1] Open Grid Forum, the new name of the Global Grid Forum

Lisa tool of which Harvey has spoken before. But he again pointed again that such advances risk leaving further behind less favoured areas of the world so he noted in particular efforts by the EU and somewhat less so by US and world agencies to address this imbalance. In passing he acknowledged successful efforts by David Foster in dramatically reducing the costs of a particular plan to link India to other areas of the world.

From Intel, **Steve Pawlowski talked about multi-core processors**. He started with a quote from Sverre Jarp's April 2007 CERN Courier article on the use of these in HEP under the slide title "Intel is listening". The target is performance via parallelism because improving raw chip performance is offering diminishing returns for the investment involved. But adding more and more cores means the need for more and more memory bandwidth must not be forgotten. One method to do this is to bring the memory closer physically and logically closer to the CPU, creating packages including both CPU and memory. In developing such technologies, Intel prefer to concentrate on small but relatively high-return segments of the market such as high performance computing, in which he includes HEP[2]. Inevitably also power consumption is a major issue and must be taken into consideration at every stage, for example simpler and less power-hungry cores but many more of them. Also more common are features to improve reliability as low level problems become more difficult to debug. Intel's target for 2011 is 10 Petaflops but they already have drawing board plans to go well beyond that. But he admitted in answer to Sverre that finding mass market applications for such HPC-specific technologies was one of Intel's greatest challenges.

Sverre Jarp continued the technology theme by comparing current and future technologies with the needs of LHC software. Starting with some history, he noted a presentation at CHEP95 in Rio where X86 PCs were first proposed as price-performant HEP compute systems. Among the recently-introduced chip features from which HEP has particularly benefited are frequency scaling, out of order execution and multi-core systems. As previous speakers, he noted that multi-core systems imply the need for more and more memory and memory bandwidth. Despite misconceptions, HEP codes do of course use lots of floating point calculations, he estimates up to 50%, but wrapped in if-then-else programming so our codes scale with SPECint. Recent core architectures offer the opportunity to execute multiple operations per instruction cycle but most HEP codes are much less efficient and typically only execute one operation today. Also, we today make little use of multi-threading, at least partly because the codes are compiled at a high level to run anywhere on the grid. These lead to some suggestions for improvement – improve use of multi-threading; write codes to improve instruction-level parallelism; simplify frameworks. During the questions, Rene Brun and others emphasised the size of the challenge of re-writing HEP codes to take advantage of parallelism and different speakers expressed their optimistic but more usually their pessimistic view of how quickly this can happen in reality. It was agreed to come back to this at the next CHEP in Prague in 2009.

Dietrich Liko opened the fourth day with a review of the analysis tools being used by the LHC experiments.CMS offers a range of analysis strategies, leaving the choice to the users. Their data model supports direct ROOT access with no separation between persistent and transient. ATLAS prefer to work with smaller data sets so they try to identify interesting events and work with derived data in thr AOD and ESD

[2] If we are offering high (financial) returns to Intel then neither FI nor the editor of this report are doing their jobs correctly! Let's hope in CERN's case he meant technological feedback in the openlab context.

which can be analysed interactively. LHCb is similar where event stripping is run centrally and the resulting data can be stored on disk and replicated to the Tier 1 sites. ALICE is entirely based on ROOT. He then moved on to the analysis tools. CMS use CRAB – CMS Remote Analysis Builder – for the grid submission of analysis jobs. It interfaces to Condor and WMS and it is a command line tool. LHCb use Ganga for job submission and DIRAC as the workload management system. DIRAC is a pull model, layered on top of EGEE. ATLAS also GANGA on EGEE (and Nordugrid) but with direct job submission; on OSG they use PANDA. For ALICE, Alien provides a unique entry point to the grid for batch and PROOF is used for interactive jobs. For all of these tools, Dietriech provided a couple of slides of features and explanations. Overall, although work is getting done on the grid, many users feel they are helping to debug the grid and he feels that strong user support is important.

For **Don Petravick** who could not be present, **Amber Boehnlein** presented **the power challenge facing HEP computing facilities today**. Currently running at 10kW/rack at Fermilab, it is possible to go much higher still using air cooling (20kW/rack). It requires some 330 watts to provision 100 watts of electronics of which about 40% is for the building itself. Searching for more efficient power supplies is becoming more important. Don believes that commodity processors will remain air-cooled but noted that some sites are starting to look at water-cooled racks. The presentation concentrated almost exclusively on experience at Fermilab and it ended by declaring that facilities will have to continue to expand. This may not come as a surprise and I don't think we learned anything new in this talk.

Peter Tenanbaum of SLAC spoke about computing for the ILC. He briefly explained the ILC, what and why (but of course not where). The greatest challenges are the accelerating gradient, the luminosity and beam availability in a very complicated machine. The talk was how advanced computing can help in each of these, for example advanced CAD tools are heavily used in the design of components for the acceleration scheme. He went into some detail in the beam optics simulations. Although it was an interesting talk on ILC technology, we learned more about what the computing results had shown in the understanding of the required ILC components and how to build the eventual machine than about the actual computing, except in the area of beam optics simulation. In summary, advanced computing is being used to design the ILC. Obviously a good thing. Also, their computing needs will increase over time. In summary, it was an interesting talk but at the wrong conference.

Richard Mount then spoke about moving SLAC from HEP to non-HEP computing. SLAC has a long history in HEP, since its foundation in 1962, but next year will see the end of BaBar data taking and the operation of the accelerator for HEP. Photon science at LCLS will begin operations in 2009, its DAQ will have LHC-like rates to storage, its offline computing has many opportunities (will coherent crystal imaging need 40000 cores?). Power and cooling future needs are very problematic. It is expected that SLAC will spend $6M/year in computing in the foreseeable future. Data compression is impossible for photon science. Non-HEP projects have already started to replace the HEP activities – cosmology, astrophysics, basic energy science. Richard gave examples of the new projects (KIPAC[3], LCLS[4] and photon physics). He also noted the ILC study going on at the lab as described in the previous talk.

[3] Kavli Institute for Particle Astrophysics and Cosmology

[4] Linac Coherent Light Source

Richard described the computing challenge of photon physics and compared this to LHC computing. The offline computing for this will offer many “opportunities”, Richard’s optimistic euphemism for open questions or challenges. These include estimating the scale of computing needed and where to perform this. Summarising the cultural differences in HEP and non-HEP, he noted that the former form large collaborations and plan in great detail in advance what is needed whereas members of KIPAC are in much smaller groups and although they also consider computing vital they shun a “professional” approach to computing services. The light source astronomers (LSST[5]) are more like HEP - large collaborations, detailed planning. The photon people are a different breed again: no history of large collaborations, see little need for advance planning, believing it not at all necessary. Technologically, the cosmologists need much more visualisation than HEP ever did and dedicated file systems like Lustre are important to them. Also they tend to be Mac users as are the light source people. The photon people need large bandwidths but not extraordinarily so and broadly they do not yet k now what they want. Turning briefly to power and cooling issues, the near future will be difficult, becoming very difficult in the medium term and “totally scary” in the longer term. Experience with Sun “blackbox”: it has actually its own cooling system as another box about the same size as the computer container, it was made available in 8 months after the decision while typical other major computer center addition take 2 years. In summary, Richard was enthusiastic about the scientific computing future at SLAC.

[5] Large Synoptic Survey Telescope

3. Summaries

Jamie Shiers summarized the WLCG workshop held the previous weekend. It was attended by nearly 200 people. To the question, are we ready, the answer depends on the component and the viewpoint – released, deployed, in use by the experiments – and Jamie presented a table. We are currently not ready for data taking in summer 2008. It was found that the services explicitly required by the experiment (CMS for example has a web page where it lists the services that it relies on) differ from the definitions in the MoU. Storage services are one of the biggest issues affecting many of the sites. A recent Tier 0 service review must be extended to critical services as seen by the experiments and also to Tier 1 and the larger Tier 2s. In the meantime, it identified varying degrees of robustness but it did establish a baseline on which we can build. Looking at site readiness, the picture is slightly worse than for bare services and there is still much work to improve throughput and stability. And there are grave risks of some centres ramping up their resources to promised levels "just in time" but which could be in effect too late for production use. Overall, service, site and exp readiness has shown great improvements since previous reviews but much remains to be done to achieve production levels. He then summarized the various BOF sessions which had been held. On data mgmt, the single most important point is the production deployment of SRM 2.2 services for WLCG. In Operations, he was pleased to report the acceptance of the use of SAM, including by the experiments. ATLAS and CMS have proposed a Common Computing Readiness Challenge (CCRC) in which the other two experiments should also participate and work will start now preparing these for early next year with a pre-challenge in February and the real challenge in May. Jamie reminded the audience that the timetable of activities next year and for the life of the LHC, we are not in charge and we need to agree our plans with the people commissioning both the accelerator and the experiments. The next workshop should be in the week of April 21 next year in CERN. Jamie judged the workshop to have been successful and the session chairman agreed and publically thanked Jamie for organizing it.

Peter Clark of Edinburgh Uni then summarized the Collaborative Tools talks, 2 sessions, 11 talks and 8 posters. On video and audio meetings, progress is such that the CHEP after Prague could be by video conference according to Steve Goldfarb. EVO, the follow-on for VRVS was presented and the improvements confirmed by experience. Another interesting tool is Hypernews, a merge of e-mail and web forums. There were a couple of presentations of information archives, one from Uni Michigan for audio and video meetings and CDS Invenio from CERN for documents but also multi-media objects. There were some medical applications making use of HEP-developed tools, one using GEANT, one using a typical HEP DAQ in an Antarctic experiment and the Health e-Child EU related project. There were 3 presentations by IT/IS on single sign-on, the RSS alerter and a poster on the new printer service.

Nick Neufeld summarized the Online track, consisting of 27 oral presentations and 21 posters, all talks except for 3 on non-LHC and one on future and 2 non-LHC posters. Most of the papers were published by the LHC experiments. The online systems for the LHC experiments are now designed. All use commercial components wherever possible, all based on powerful LAN technology. The installations will be ramped up when luminosity increases. Transient storage is made of cheap SATA disks, all permanent storage is in CASTOR, online capacity buffer is available for 2-3 days. Outside of the LHC related projects, reports were given of mature and new frameworks. Experience with D0 shows that online computing can be run with few (3) experts (with the computing having being taken over by central IT team). There were only few papers on future DAQ systems or upgrade of current ones, everybody wants their systems working first.

Roger Jones reported on the Distributed Data Analysis track, 31 talks and a number of posters. Much talk of frameworks. Even the LHC experiments are now working with near-live Data Acquisition frameworks. Experience can be gained from running systems.Status of GANGA, CRAB and ASAP was given. GANGA (LHCb and ATLAS) is used outside HEP and even commercial / non-academic applications. GANGA provides a transparent transmission to batch system. It is an extensible framework with possibility to plug-in. CRAB (CMS) is now supporting submission through a server (with retransmission), it is a global trend. An especially interesting option is Xgrid (Apple product). CDF is moving to Grid, especially for its MC production. It actually improves the reliability (thanks to the Grid middleware). Data access is not as efficient as data transfer (disk, LAN and WAN latencies). ROOT and xrootd can use certain techniques to avoid the limitations imposed by latency, it allows efficient analysis of remote data, even if the bandwidth is not very high. ATLAS and LHCb have decided to use 3D Oracle/Streams for the conditions database replication. Frontier is used by CMS and has been shown to match requirements. Encouraging tests have been made using Oracle database for TAG querying, challenges are ahead. Babar has completed the migration out of Objectivity. Lot of monitoring tools have been developed, BaBar have developed an unobtrusive data access monitoring system which can be decoupled from the underlying data server and is general enough to be reused by other experiments. CMS has described the operational structure intended to be used. Distributed analysis in the Grid era is a big challenge, each link in the chain matters, the required tools are maturing, even running experiments are starting to benefit, there is still scope for re-use and common projects, various challenges are testing the system, lots of monitoring is available -and monitoring of the monitoring. One needs to have the human infrastructure ready to have everything working.

Federico Carminati summarized the Software Components talks, 33 in total. Metadata is a major concern of the LHC experiments. Regarding data storage and database, metadata is a main worry for the LHC experiments. It is usually an underestimated problem. There are large investments in manpower and hardware based on Oracle, more commonalities among experiments would help, the integration with the Grid world is not straightforward (especially security). ROOT is being actively developed, especially on graphics and I/O. Progress on BOOT is steady. There is a lot of complexity in the analysis part with schema evolution, experiment policies, metadata access, computing model, resource allocation and availability, data tagging… with an emphasis on the user interface. The role of the UI is unclear. Regarding Monte Carlo, geometry remains the sore point. An independent geometry representation is the general trend. Regarding code development / optimisation and lifecycle, some process control would help. Program performance assessment is much more difficult with C++, usage of multi-core CPUs adds to the complexity of the problem, the lack of good tool is an issue. It is important to maintain some forward looking; there will be opportunities to do better as we will have to rewrite most of the code during the lifecycle of the experiment.

Patricia McBride reviewed the Event Processing track, 63 contributions, half on LHC experiments, 35 papers, rest posters. Majority of the talks came from the LHC experiments, contributions from medicine, astrophysics and future experiments. Simulations: lots of work since last CHEP on validation, geometry and integration. The validation continues. Tracking: experiments have demonstrated the need to have more than one tracking algorithm. Alignment: procedures are under development for the LHC experiments. Commissioning LHC detectors: performance optimisation is well under way. Visualisation is not an easy issue, there was an interesting contribution from ALICE that will be made available to others through ROOT. Towards the ILC: 4 international detector concept studies for the ILC ongoing (with 4 frameworks).

Summary of computing facilities, production grids and networking track, Kors Bos (NIKHEF). NDGF is a collaborative work in the Nordic countries. Services are scattered and resources are shared. Also the storage is distributed (dCache) but looks as a single one. It works for ALICE and ATLAS. Memory is more expensive than disks or flash for storage systems but one should also consider the price performance ratio. An experience is being done at SLAC, it could provide revolutionary benefits for pointer-based data analysis. SAM provides sanity checks against selected grid and application services. It is used by all VOs. It can be used for other tools. Production experiences: the ATLAS (OSG PanDa) and CMS production system were shown to work (>50000 jobs per day for ATLAS, >20k jobs for CMS). Production is still manpower intensive. The LCG 3D project has setup a wold-wide distributed database infrastructure for LHC. For the networking, IPv6 will essentially provide mobility and a wider address range, in 2010 there will likely be no more IPv4 address range available. Pilots (somehow bypassing the Workload Management System) are becoming a reality but are introducing new problems including for security. gLExec can help solve the technical problems. It was noted that they were “too many” WLCG papers, that an enormous attention has been devoted to services.

Summary Grid middleware and tools, Ian Bird. Experiment experiences: a large amount of work has been put to identify job failures and the reasons. In many of the cases, the core reason is an error at the site. Data management: SRM has been discussed many times. SRM 2.2 is now being implemented for all of the storage back-ends. LCG data management tools: a lot of development has been performed. The FTS server infrastructure is scalable. CORAL provides a grid enabled database access, not as complete as OGSI but it is easier to use and faster. A pilot job is a job submitted on the normal job management system, it is an empty shell that calls for work. It has been implemented in several of the frameworks (DIRAC…). glExec needs to be used with pilot jobs to make them more secure. Monitoring is becoming a critical part. The dashboards take data from everywhere (including existing accounting systems) and add instrumentation. GridICE provides a different view of running jobs. We provide the data back to the sites so that they can react. The experience of using DPM over the WAN presents very encouraging results. The LSF scheduler can now talk to many clusters (batch systems), it raises the question of whether our work on Workload Management System is not superseded by industry. Solutions exist for most needs now. The service management issues start to be addressed by some services. A common requirement was identified: all have to debug individual problems to identify problems, it is helped by the dashboards.

Conference Summary, Mattias Kasemann CERN & DESY. Mattias first made a study on the conference itself: attendance is from many countries with 113 persons from Switzerland, 104 from USA. There were 429 abstracts with 1208 authors. He then pointed out some ideas which were expressed in some of the session. Computing at laboratories: in order to provide 100 Watts to a CPU, one typically needs 330 Watts in total. Retirement after 3 years leads to 30% gain in compute power. There are other user communities that demand high computing capacity (=HD films). HEP applications are embarrassingly parallel _as long_ as there is enough memory. It was shown that the typical HEP code is not efficiently at all. Status of the machine and experiments: we expect that data taking starts next summer. By spring 2008, experiments will be ready to take data by next summer. Analysis tools for the LHC: we have quite good experience already on the Grid. We are in a situation where not all is perfect. An important aspect is user support. HENP other than LHC is significant fraction of total CPU utilization. Advanced computing is critical for preparing for future experiments (ILC). WLCG status: 45% Tier2 35% Tier1 20% CERN, we moved away from the Monarc ⅓, ⅓, ⅓ model but not too much. Reliability is a major concern, monitoring is key. We are “slowly” getting there. Network: HEP is among others driving network development. We have to learn how to use these N x 10G effectively. It is important to address the digital divide.

Finally, Mattias ended with a “thank you” to the organisers and to all whom attended.

4. Computer Facilities, Production Grids and Networking

Jamie Shiers opened this parallel stream with some **lessons learned from the successive WLCG Service Challenges**. From the WLCG Memorandum of Understanding, service availability targets of the different sites range from 95% to 99% for compound services, for example: acceptance of raw data from Tier0 or data-intensive analysis services, including networking to Tier0. Jamie described these requirements as “somewhat aggressive”, especially that such “services” involve many sub-services, for example storage services.

He found it interesting that scheduled interventions appeared to cause greater downtime that unscheduled ones and the worst have been extended downtimes at numerous sites cause by cooling or power work. Jamie rhetorically asked if the schedule interventions can be done in a way transparent to the users. The answer is yes, but with some provisos. Especially the common interventions, such as adding resources, replacing hardware, operating system or database upgrades can be done transparently if a few well known tricks, such as clustering, load-balanced servers and rolling upgrades are used. Not all software supports it, but the advantages to both users and service providers are huge. Among other things, the use of these technologies makes the life of service providers much easier and the relationships with the users much easier.

Jamie also mentioned things that should be done whenever a schedule intervention takes place:

- Always have intervention procedures, and stick to them
- Plan your interventions: prepare an intervention procedure, understand the decision points, for example what if the database does not start – can you roll back?
- Record what happens during an intervention, keep it for future reference and distribute to others, so that they can profit from your experience.

Jamie finished his talk with the conclusion that after a very long period of hardening the main grid services are finally reliable enough, but the single points of failure, that can still be found, are “the enemy of the grid” and should be found and removed. After a question from the audience, Jamie confirmed that this work is currently in progress and gave Castor as an example.

An interesting comment from the audience was that it is important to understand the impact that the downtime of a service has on its users.

Markus Schulz then presented on their behalf **grid security incident management** as implemented by the EGEE grid security team. He described the various possible incident cases and why grids in particular offer a large exposure to attack, for example high bandwidth to the Internet and the need for transparent access once authorised. In fact, thus far the grid has not been successfully used as an attack vector but he believed it was just a matter of time. A formal response procedure has been defined and Markus highlighted some of these, for example an incident should be analysed and the report summarising lessons learnt should be distributed within a month of the incident, so that people are still interested to read it.. There is a feeling that small sites are more exposed than large ones because there are more of them; they tend to have less security-experienced or security-aware administrators and are generally new to being open to external access.

Ruth Pordes of Fermilab and OSG Executive Director then gave a **high level view of the OSG** (Open Science Grid). In order to fulfill their overall vision, a lot of effort is spent in creating a broad community of contributing partners. It is currently a 5-year DoE and NSF funded project covering 34 FTEs in 18 institutes, now in its second year. Both physics and computer science is represented in the management. There are 75 processing sites in the US plus a few others processing about 75,000 jobs per day on ~30,000 cores and moving some 10TB per day and it covers some projects beyond HEP. OSG does not own resources or develop software but they have prepared collections of down-loadable modules from the VDT for different target nodes such as compute elements, storage elements, worker nodes, etc. OSG version 0.8.0 is in use now and version 1.0, targeted for LHC production use is planned for February 2008. As well as Linux, they have kits for AIX and MacOS and the first Windows sites have appeared on OSG and there is a question of what level of software support they should receive.

Following the OSG talk was one by **Jeremy Coles on GridPP in the UK**. It is a quite broad project, but LHC community is their larger user community. Despite the LHC resource promises being met and more about to be added, the bad news are that these are not much used yet although usage is growing. They will migrate from dCache to Castor (migration still ongoing). A new machine room is needed at RAL and not just because they need more space, but also for more power and better cooling). GridPP 3 starts mid 2008 and they are founded until 2011.

Continuing the grid overview talks **Michael Gronager described the Nordic Data Grid Facility**. Nordic countries constitute together 25 Mio people but no country is bigger than 10 Mio people. The organization includes a CERN coordinator. The 7 biggest sites offer naturally high-availability. Of the services is networking. 3 D setup is currently made of a single server, a RAC is scheduled for installation. A specific sensor has been developed ARC-CE. Sites report using SGAS, the report is then translated to be included in WLCG.

Next, **PetaCache: Data Access Unleashed was presented by Richard Mount**. Motivation: CPU per $ has fallen faster than disk space per $. CPU is doubling every 1.9 year, storage every 1.37 year. Three orders in magnitude since LEP area. In-balance is increasing. Every application will is random access must be RAM based or de-randomized so that it can be made on disk. Cost effectiveness: DRAM $100/GB, Disks $1/GB, Flash based storage $10/GB. PetaCache started as Babar had data access problem. It was devised rapidly that new strategies and software for data access would also be needed. Prototype (operational since 2005), flash based prototype (operational soon) with 5TB of flash memory fine grained, high bandwidth access. Violin has (or is about to have) commercial DRAM and flash based systems. Latency is caused by the high number of stacks (from client application stacks to server-based storage), disk access is the biggest part of the latency. ATLAS AOD analysis disk versus memory, 2.8 time ratio. We have made our applications serial mode, are we will be able to make TAG database driven analysis. The way we do physics analysis has to change in order to be able to make use of the hardware to come.

CASTOR 2 was presented by Giuseppe Lo Presti. He started with an architecture overview and the various workflows. He noted the advantages of being database centric – for example easy replications of components, state persistency, simple scalability, stateless daemons, etc. Tuning at the database level code has been made so that the database is no more the bottleneck. The entity relationship model is mapped to classes. The code generation system is based on Umbrello, extended by the Castor team. Architecture has been able to scale up to LHC data challenges. The infrastructure is mature, the focus is on tuning. He showed the development framework based on the “Factory” design pattern and including a code generation facility. He claimed the scalability had been proved although this statement was later (gently) challenged by the session chairman (Kors Bos).

Experience with gStore, Horst Goeringer (GSI). gStore is based on TSM combined with GSI software (interface to users, aPI to TSM, management read/write cache). One of the projects is to have 2nd level Data Mover, another to have a SRM interface. 250 TB stored, experiment data is stored twice as it is very valuable data.

Advances in integrated storage, transfer and network management, Rick Cavanaugh (univ of Florida). In-house applications are not aware of network topology, paths, available BW/path. Applications (blindly) retries and perhaps requires expert eyes to diagnose, they pre-suppose a passive network. Project UltraLight was started 3-4 years ago. It aims at developing and deploying a prototype. Example EVO can take actions in case of network problems. A distributed scheduler has been built based on distributed agents. One of the reasons for UltraLight is to maximize bandwidth, working on integradting FDT with dCache. Benefits for LHC: higher performance, less manpower, better end-to-end monitoring and alerts, pervasive monitoring and rapid reaction time.

Distributed Cluster dynamic storage, comparison xrootd/dcache… Alessandra Fort (univ of Manchester). Manchester has ~900 nodes with 2x250GB disk each. Disk1 = OS + scratch + data, disk2 = data. Slashgrid is a shared file system based on http. It was developed as a light weight alternative to AFS. Dcache is more complicated, difficult to manage but has 3 features difficult to beat (resilience, srm front end, 5 different protocols). Slashgrid could be useful for software distribution (easier to install than AFS). Speed within the cluster is comparable.

CMS Experiences with Computing Software and Analysis Challenges – Ian Fisk. He described challenges in terms of data rate, network transfers, job submission for T0, T1 and T2. Most challenges achieved, but not the one on number of jobs per day (targeted 12k jobs/day, being today 3k jobs/day). Showed work-flow for CSA to perform simulations. Tier-0 reconstruction test very successful in all aspects. T0 to T1 transfer performance test also met the challenge. T1 transfer quality, good results but still some sites had specific problems and down-times. T1 to T1 transfer: Only a small fraction if these were tested. Still far from targeted 50%, 90% successful execution of batch jobs.CSA07 will start shortly after CHEP, with new functionality and 50% scale attempt

Michel Jouvin from LAL presented how QUATTOR is used to manage a complex site, namely the distributed Tier 2 “site” in Paris, GRIF. QUATTOR came originally from an EDG development in 2001-4 and was (is) targeted at managing reproducible software con figurations in a service-oriented manner. Although he said he would not, he spent some time describing the tool and how it works, including the templates proposed/developed by the LCG QUATTOR Working Group. There is a single template for each high level configuration and one for each machine type. Templates can be layered in the sense that a high level template will include lower level ones. Coming to the main topic of the talk, he suggested that, apart from size, a site may be complex to manage if it has to offer a wide range of services spread across many nodes which is the case for GRIF where some constituent individual sites are not well-staffed with respect to the requirements for LCG service goals. By using a combination of standard templates and some site-specific ones, a lot of systems are covered. QUATTOR takes account of individual access rights which reassures local site admins. Overall some 40+ sites in Europe use QUATTOR, almost all inside LCG, from CERN of course with ~5000 nodes managed down to small Tier 2 sites. From these numbers and the work done by the QUATTOR Working Group, he claims that QUATTOR is now a community-backed product with the main developers and expertise being outside CERN.

Next came a talk on **EGEE user support by Torsten Antoni of FZK**, concentrating on day-to-day user support. EGEE has built a Global Grid User Support (GGUS), integrating existing user support efforts. There is a single point of access via a web portal; there are self-help options, search tools, etc. Each ROC has its own help tools so interfaces had to be built so that there is an integrated ticket system. FZK runs Remedy and has built interfaces to the ticket systems used by the other players. There are also interfaces to individual VO support teams. He showed the different ticket workflows, including how tickets can be rerouted to where they can be resolved. Support is split among different first, second or third level support units. Current efforts aim at improving ticket turnaround by involving more partners, their responsiveness and their expertise. Another challenge is better integration between GGUS and the VO user support teams; for the latter, GGUS is a small part of their load.

Antonio Retico described how SAM is used to monitor the EGEE/WLCG grid services. SAM has been in production for over a year; it is currently monitoring service availability on over 200 sites and SAM output is stored in an Oracle database and used as the basis for site availability. Various site metrics are defined for site availability. A site is available if at least one instance for each of its services is available – i.e. it passes a test run periodically. SAM is also used for middleware certification testing with a lightweight display interface. Antonio listed a number of use cases of SAM, Mainly for EGEE/WLCG operations but increasingly also for experiment monitoring. The framework can be extended with custom tests/use cases. LHC expts. are using it by plugging their specific test and they can benefit from the visualization. CMS given as example. ALICE, and LHCb asked and implemented a more sophisticated integration with custom visualization tools.

Jan van Eldik discussed streamlining and scaling CASTOR2 operations. The most important CASTOR2 feature for this talk is service classes where disc servers can be pooled, can be configured for specific activities and have their own policies for tape migration and garbage collection. Now used by all LHC experiments. Stability has improved after a number of fixes and updates and two non-LHC experiments are using it to write raw data to tape. BY the end of 2007 it should be controlled over 2PB of data on 400 servers and this will more than double next year. There is a three level alarm scheme – 24 hour operations service, the second level on-demand system admin team and finally the CASTOR operations team. There is a similar three level user support level – Helpdesk, service manager on duty and again the CASTOR operations team. Currently around 125 operational problems come in but only 5-6 get to third level. There around 15-20 user requests and one-third get to third level. They have implemented many automated tools, for example system installation, problem detection and repair and the management of grid host certificates. Also they take advantage of LEMON which is very useful for monitoring. Procedures have also been developed for some common operational tasks, for example the configuration of disk pools and disk cache repair schemes. With recent improvements in stability and the procedures for round the clock support, they believe they are in a good position to scale up the service.

Dirk Duellmann, Production experience with distributed deployment of databases for the LHC. The LCG 3D architecture is described. Oracle Clusters are used as the main building blocks of the database services at Tier0 and Tier1s. Data replication from Tier0 to Tier1s is implemented asynchronously with Oracle Streams on http caching (with Squid) or via cross-DB copy. This is important, because it blocks neither the source nor the destination databases. Additionally, TCP and Oracle protocol optimisations resulted in significant throughput improvements (factor 10). CMS uses successfully the FroNtier/Squid model. Monitoring and alerting is integrated with the WLCG operations. Scalability tests have been performed at tier-1's with realistic loads to validate the experiment resource requests. Oracle Streams is successfully used for online-offline replication. Demonstrated full local recovery and resynchronization procedures.

Xavier Espinal, Large scale ATLAS simulated production in EGEE. The ATLAS production system consists of a common database (ProdDB), a common supervisor (Eowyn), several executors (CondorG, Lexor and Cronus for EGEE, PanDA for OSG, Dulcinea for OSG) and the data management system (file catalogues + data transfer (DQ2)). The performance of the production system has been constantly increasing and now the wallclock time efficiency is about 90%, while it is ~60% in the number of jobs. The organization of the production operations is described. New developments are aimed at improving the monitoring and the level of automation.

Jose' Hernandez, CMS MC production in the WLCG computing Grid. After a brief description of the CMS computing model and the data processing workflow, the experience with the old production system is described. Understanding its limits (limited scalability, poor error handling and integration with the data management, inefficient usage of resources) was very valuable in designing the new production system (ProdAgent). Last summer CMS has produced 64M events/month, with 11K batch slots available to production. More than 20K jobs/day were submitted to 40+sites with an average job efficiency of 75% (including application+grid failures). The computing resources were occupied at

50%, due to operational inefficiencies. New components are being integrated to further improve automation and reduce the needed manpower.

Jim Shank on behalf of Yuri Smirnov, ATLAS MC Production on OSG with PanDA. More than 30% of the ATLAS MC production is done in OSG with PanDA. PanDA uses pilot jobs and provides a single task queue, and manages all production and analysis activities in US and Canada. Both the available resources and the efficiency of the production system are increasing. Panda simulated so far more than 30M events, producing 400 TB of data. Since April, the job efficiency is 85% and the wall clock time efficiency is 91%. The 44% of the failures are application errors, while most of the other failures are related to the data management system. Overall, PanDA is considered to be working well and the OSG production operations to be robust.

Dave Evans, CMS ProdAgent and Production Infrastructure. The new CMS production system was developed to make it easier to maintain, more scalable and to avoid single points of failure. The architecture is described. ProdRequest: is the interface to create and track new requests; ProdMgr splits a request in allocations which are then assigned to an array of ProdAgent instances; ProdAgent converts the assigned allocations into Grid jobs which submits to sites. The ProdAgent takes care of merging the produced files in larger sites and of injecting them into PhEDEx. A ProdMon monitoring component is being rolled out. Since 2007 the same system is used also for data processing (i.e. (re-)reconstruction and skimming).

Maxim Potekhin, Experience with on-demand physics simulations on Sun Grid at Network.com. The event simulation activities of the STAR simulation program (which takes care of simulating physics processes in many fields of particle physics) have been recently moved to the Grid. Although most of it is on OSG, a small part has been done on the Sun Grid as a means to evaluate a typical commercial Grid. A web interface gives access to Sun Grid and there is also an API interface. Users pay per cpu/hour. . It has two models: Infrastructure on Demand, where you run your own software on the grid, and Application on Demand, where you can run some of the supported applications. There is an application catalog listing all available applications. Sun grid provides access to any Internet node. Half a TB disk space was allocated for the experiment. A non-straightforward port to Solaris 10 was necessary. The main advantages were no custom software to install, an easy port to Solaris, a clear user interface, the availability of dedicated storage and a good support; the disadvantages were the high price point ($1/CPU hour but even this is a promotional offer and it may rise in the futuee) and the lack of a programming interface is a show-stopper, Sun is working on this issue. The data transfer to and from the Sun grid is the current biggest issue, it is being addressed by Sun.

Igor Sfiligoi, Addressing the pilot security problem with gLExec. This talk described how gLExec can be used with pilot jobs. Normally, pilot jobs create two problems: first, the site does not know who is the real owner of a job (thus denying the possibility to ban individual users), and second, jobs belonging to different users can interfere with each other because they are all run under the pilot job's credentials. gLExec can solve these problems by switching UID/GID for the running job depending on the real user's credentials. gLExec has been successfully deployed at FNAL on more than 500 nodes, using GUMS as a central authorization service. All authorization requests are logged by gLExec and the information is fed to Gratia (the OSG accounting system). No scalability issues were seen at this level.

Simone Campana, Experience with the gLite WMS in ATLAS Monte Carlo production on LCG. The talk starts describing the ATLAS MC production system and the new functionalities of the gLite WMS: bulk submission and bulk matchmaking, an internal task queue, automatic resubmission, and a limiter mechanism to prevent high loads. While in 2006 the WMS had lots of stability problems and was very difficult to use in a production context, during 2007 it managed to satisfy the CMS and ATLAS requirements after a long and intensive testing and debugging effort led by the EIS team of WLCG. Now the gLite WMS is used by ATLAS for the MC production, running up to 20000 jobs/day, with a

~66% job efficiency (~13% due to WMS but including also site-specific problems), and 87% wall clock efficiency (negligible contribution from the WMS). CMS is using the CMS for "fake" analysis jobs observing the same performances for the WMS.

S. Chechelnitskiy, Running CE and SE in a Xen virtualized environment. The talk described the experience at SFU (a WestGrid site) with running WNs and a dCache server on a virtualized environment. The motivation is being able to run ATLAS jobs and jobs from the local community (serial+MPI jobs) on the same hardware, but accommodating very different software requirements. Performances for serial jobs in the virtualized environment is more than 95% of the one in a standard environment. The tests, done on two WNs with dual core CPUs, were successful. The cluster will expand up to the full scale of 2000+ cores. A similar test was done to run the dCache core services on a virtualized environment, also with success. Virtualization is seen as an effective way to improve the stability and the flexibility of a site.

Zhenping Liu, BNL dCache status and plans. The talk describes in great detail the current dCache setup at BNL, serving, among others, ATLAS (BNL is an ATLAS tier-1). Some bottlenecks have been identified (on PNFS and SRM) and high loads on the GridFTP door nodes are seen. Upgrade plans include more nodes on better hardware, 64-bit OS, high availability for critical services and better monitoring.

ATLAS distributed data management and operations, experience and projection, Alexei Klimentov. The software used is the second generation of ATLAS DDM system (DQ2[6]) based on data sets and subscription. ATLAS has a hierarchical model of data replication, RAW data is replicated to Tier1s, ESDs are replicated to all Tier1s and Tier2s according to their request. Tier2 can even request RAW data replication. Web interfaces are provided for requesting data replication. An issue need to be discussed with the FTS experts: after FTS stop, transfers are still observed for the next 3 days. Replication only starts when all files are present in the dataset. 24x7 support will be provided for operation.

Daniele Bonacorsi, Computing Operations at CMS Facilities. CMS uses the PhEDEx agent. For the test reported, there were 7 T1s, 49 T2s, 13 T3s. T1s are used for Skimming and Re-reconstructions, T2s for Simulation and Analysis. Data processing is running at 20K jobs/day with recent data challenges as if for experiment running, with 1.2 GB/s sustained. CMS is approaching the target scale for data transfers, but the quality of transfers is not yet good: in many cases transfers don’t succeed at first try, participation of T2s is not at the target. A “Debugging Data Transfers” Task Force is in place since July 2007 to bring all T1-T1 and T1-T2 links at the required quality. He explained used of SAM in CMS. All SAM tests have run successfully and he described SAM as a good tool to focus people's attention on quality of delivered services. Looking forward at CSA07 and the joint data challenges w/ ATLAS in 2008.

Storage management solutions and performance tests at INFN Tier1, Luca dell’Agnello (INFN). CASTOR and GPFS/StoRM are used at CNAF. Tests were made to compare xrootd and dCache. Setup includes a SAN with fibre optic switches; CASTOR and dCache were setup with 23 diskservers; GPFS was setup with a unique large file system of 110 TB. Different tests at different levels were performed, from low level network throughput to I/O to the storage systems. GPFS performed quite better than the other systems. Final tests involved real analysis jobs from LHCb. CASTOR was excluded in this case due to errors observed with the previous tests. GPFS again outperformed xrootd and dCache. It was shown that the failure rate is constant for GPFS while the others have varying failure rate. The validation of GPFS setup was successful; the issues faced with CASTOR have been fixed with a more recent release. Conclusions: GPFS has shown to be performant and reliable. Fits LHC and non-LHC experiments hosted at CNAF.

[6] DQ = Don Quijote

Streamlining and scaling Castor2 operations, Jan van Eldik. The disk cache has been redesigned in Castor2. The stability has been greatly improved. The performance expectation has usually been met during the experiment data challenge, serving constantly 2GBytes/second total. By the end of 2007, the disk cache is made of 2PB over 400 servers, in 2012 it will be 21PB over 1000 servers; it is a major challenge to install and manage such a disk pool. The three level support model was described. Machine alarms arrive to the operators, user calls to the helpdesk or GGUS. Most of the calls are answered by the operator at first level by the operator or helpdesk, some get escalated to the sysadmin team (second level) and few needs to be taken by the CASTOR experts and developers (third level). There is 24x7 coverage for simple cases, and with a CASTOR Service expert involved for complex alarms and/or user errors. Current figures: the service expert gets 10% of the alarms and user tickets, and a very small faction is passed to the development team. A lot of automation has been created: machines are installed and managed through Quattor and LEAF, and, for example, if a box detects that it has a problem, it will itself take itself out of the system.

Implementing SRM v2.2 functionality in dCache, Timur Perelmutov (FNAL). Started by describing SRM V1.1. It was a solid protocol but lacking directory functions, file access permission management, and other features. To address those lacking features SRM 2 was created. The SRM 2.0 to dCache interface work started in 2003. Finally in May 2006 it was decided to drop SRM 2.0 and use v2.2, be released in October 2007. He listed SRM dCache services: Loginbroker (for protocol negotiation), PinManager (implicit pinning), Space Manager 9dynamic space reservation), Pool Repository and NameSpace, Transfer managers, and gPlazma for Grid access control services. Now supports several Access Control Mechanisms (dCache authorization, etc). A new functionality is the srmCopy, where SRM becomes client to another endpoint using user delegated credentials. This needed dedicated development, and it's currently implemented as a srmGet/Put file by file on the remote endpoint. dCache SRM implementation benefits from BNL (horizontal stress test) and OSG (file removal after space expiration) testing. dCache 1.8 supports both SRM 1.1 and 2.2 on the same port and with the same namespace. Plans: full WLCG functionality by early 2008, more stress tests, horizontal scalability (BNL contrib.): try to split internal SRM components (currently it's a single webservice on top of tomcat) to be able to distribute them on different load balanced nodes. The comment was made that increased work needs to be done on the understanding and tuning of long distance transfers. Conclusion: stronger collaboration involving e.g. BNL, OSG, will insure on time delivery.

5. Collaborative Tools

Emmanuel Ormancey presented CERN's solution providing a single authentication mechanism for CERN services. In the new scheme, the resources can be accessed with a single security token instead of many passwords, as was the case in the past. An additional motivation was to have a way of authenticating and authorising external people, who are not CERN users, but still need access to some of CERN's service.

The new scheme enables several authentication methods:

- Classic forms (login and password)
- Certificates (CERN CA Certificates, smartcards)
- Windows Integrated (reuse current Windows credentials)

Based on either of these methods, user information, group and mailing list membership can be retrieved. This can then be used to authorise access to a specific application. The authentication is external to any specific application, which makes it more secure, as the application owner cannot misuse the authentication token (for example, cannot store the username and password entered by the user).

There are two main architectural parts of the system: Service Provider (the application that requires authentication) and Identity Provider, which verifies the identity of the user and provides group membership information. At CERN, the identity provider is based on Microsoft ADSF product, which adheres to open standards and can therefore easily interact with applications from other vendors (for example the Apache Shibboleth open source project). For non-web clients, a home-made SOAP web service is provided.

The audience expressed their need to have AFS included in this SSO solution. The chairman also remarked, that instead of forcing universities to use CERN authentication, CERN could use universities' systems to authenticate CERN users coming from these universities. Emmanuel replied that this can well be the next step of the project: such federation of authentication services would be possible thanks to the choice of the product based on open standards.

Next, **Richard McClatchey described the Health e-Child** EU-funded project (16M Euros over 4 years ending in 2009). He acknowledged that a lot of the real work has been done at CERN and by Maat-G but the most interesting part of his talk was his comparison of the use and applications of grids to HEP and bio-medicine. The latter resemble more small Tier 2 sites, limited resources to install, configure and support a grid and perceived complexity to adapt the target application to the grid. Also confidentiality of data is much more important in this field than in HEP and doctors require to work interactively, not running batch jobs. He suggested that grids need to become more mature before being taken up more widely by other sciences than HEP. Also, biomedical communities are very different one from another and "no one size fits all".

Nick Robinson reported on using CDS Invenio to manage an institutional repository. The CERN CDS repository is already quite large – over a million entries, 10,000 searches per day – and includes multi-media entries as well as simple documents. They have designed their indices to be scalable to millions of records and have built both simple and complex full-text search interfaces. Documents are navigable by collection and there are 550 collections currently. Document submission can be done via the web or automatically via the web, by e-mail, using custom protocols and by synchronization to external databases. A lot of work has gone in recently to collaborative tools such as user-defined document baskets with automatic e-mail alerts and basket sharing; facilities for reviews and comments; authentications to provide security of information between groups and multi-language interfaces. For those interested, his slides contain many details on the internals, including the indexing methods used but he had twice as many overheads as could be seen in a 20 minute talk! Invenio is GNU GPL with 50 institutes using it. A question was asked if Indico conference data could be stored in CDS for easy access.

Douglas Smith described Hypernews feature by feature. It was developed by BaBar at SLAC. It was designed to merge the flexibility of web forums with the rapid (?) feedback of e-mail. Membership is configurable and can be configured that only members can post. It is highly configurable with some emphasis on security. Postings are saved in files and are searchable via a tool based on Swish-e. They have some 450K postings currently, they index every 15 minute and offer very fast searches. Discussions can be threaded. Anti-spam has been added to e-mail submissions and individual member posting can be controlled and e-m ail attachments can be submitted. It is also used by ATLAS and CMS and other experiments and outside HEP. Use and feedback is encouraged as it is an open source project.

University of Michigan Lecture Archiving – Jeremy Herr. Overview on collaborative tools used in Atlas and different activities in that field. Listed various projects implemented by Computer Science undergraduate students. Web Lecture (any web browser and RealPlayer plug-in). The lecture object is composed of standardized data object meta-data, timing, high-resolution media, etc. Listed all Lectures which are already archived in this system. MScribe: a larger scale, highly automated recording system to record and archive eight University courses in a wide array of subjects is also used by Atlas **.** The system is fully automatic to the extent that the cameraman has been replaced with a robotic camera system (4 robot carts are available; if none is available, the lecture can still be recorded by a human operating the camera). The lecturer wears an infrared LED necklace that allows to robot to follow the lecturer's movement. Currently, an ultrasound tracking system is being developed to get rid of infrared imperfections (which is less reliable in high sunlight). The lecture recording software is installed on a single portable PC. Several studies going on to improve their tracking system using also ultrasound. Work ongoing to make SMAC support lecture Objects. BlueStream used for advanced searches, which coverts speeches to text for facilitate searches. Long wish list for the future: integration with SMAC, desktop recording software.

Extra Dimensions 3D and time PDF Documentation – Norman Graf. Acrobat reader allows interactive 3D models as of version 7. Norman gave multiple examples of 3D objects embedded in PDF from simple geometric shapes to parts of parts of a detector. Users have the ability to interact with the scenes, customise the viewing experience and exchange commentary via PDF markup. Currently only commercial software (including CAD products) can create and embed 3D objects in PDF (U3D content), but there are efforts under way to provide a toolkit for some HEP applications. The speaker showed several examples. Some features are:

- Use scripting to control views
- Animation
- Detector Visualization

The IceCube Data Acquisition Software: Lessons Learned during Distributed, Collaborative, Multi-Disciplined Software Development – Chuck McParland. Overview on IceCube DAQ Development. It detects extraterrestrial neutrinos and it is deep in the ice in Antarctica. The "Eyes" are Digital Optical Modules. The speaker showed a schema of this DAQ system**.** This experiment requires multi-institutional collaboration. Its Project Management is not so easy. Different institutions use different project management strategies. Listed used communications tools (VoIP, IM, e-mail, Chat, other well known tools) as well as development tools (for Version control, issue tracking, etc). Showed wish list for Development Methodologies without mentioning any technical solution. In fact, this talk did not have any technical content. Conclusions were not too concise, just generic aims.

6. Grid Middleware and Tools

Pablo Saiz opened this stream with a talk on **grid reliability**. He and his colleagues have developed a dashboard which gets its data from different sources and publishes it as xml csv and html. The tool aids reliability by tracing job submission attempts; it produces reports and makes statistics; it can identify the most common errors such as waiting time, etc; and it can produce single or multi VO reports. It can be found at http://dashboard.cern.ch. The dashboard scheme is now used by all the experiments.

Oliver Gutsche reported on WLCG scaling tests from the 2006 CMS data challenge (CSA06). He started by explaining a "crab" job and the CMS computing model and then presented the results of 10^6 jobs. He set a constant job rate split between OSG and EGEE 90/10 in the beginning, later 40/60, at the end up to 45k jobs per day. There was a large fraction of unknown jobs, jobs which never arrived on a WN and finally aborted but at the end there was a 71% overall success rate, split 55% EGEE 83% OSG. The CSA07 run will be at 50% of needed capacity for 2008.

The next talk was from D0 on **data-intensive computing on OSG**, requesting 2000 CPU for 4 month. OSG consists of some 80 facilities ranging from 10s to 1000s of nodes and they have opportunistic usage model. D0's SAM Grid implements D0's processing needs and interfaces them to OSG. Adding new resources is an iterative process, including certification. For certification they run the some standard files and compare with a reference. Within the OSG sites, small differences are possible in site configurations and these created many problems in the setup phase but eventually they achieved a state where they were rather effectively using resources which would otherwise have not been available to them.

Tadashi Maeno presented PanDA, the ATLAS distributed production and distributed analysis scheme. It is based on pilot jobs and runs both production and analysis jobs. They have plans to expand it also to LCG sites. Using the pilot job, the data is already there when job arrives so the job does not need to wait for it. The question was asked what criteria are used to decide on the number of jobs to submit to a side. It depends on the number of jobs in the queue.

Pablo Saiz returned to speak about **AliEn2**, the ALICE grid environment. The Alien file catalog is organized like a UNIX file system with an LFN (the file itself) and GUID catalogues (Meta data). Jobs go to the side where the input data is. The Alien user interface hides different grid interfaces from the user. Pablo described the concept of a grid of grids (AliEn-AliEn interface), including support for multiple VOs. His TODO list includes better automated error recovery, bulk submission and better store management. For more detail see http://pcalimonitor.cern.ch

Following on from Les's comment in the morning about work remaining to be done on and around **SRM 2.2, Flavia Donno** reported on precisely that, concentrating on testing. She started by explaining the need for SRM2.2. After experience with SRM 1.1 and 2.1 (never deployed), experiments' requirements had changed and they needed new features. Specs for 2.2 were agreed in May 2006 at FNAL. She covered briefly the specification and the prototype implementation history. There are currently SRM 2.2 implementations for 5 storage systems to be tested (CASTOR, dCache, sTorm, etc). Tests must cover consistency, validation, verification and performance. The initial testing effort, up to Sept 2006, involved only manual tests with LBNL testing suite. Afterwards, CERN took over development of the S-2 test suite: automatic runs with results published to the web for developers to follow up. Flavia described the proposed testing framework including some techniques to limit the number of test cases; it is based on the S-2 testing language developed initially at RAL and available on sourceforge. She showed some test results and concluded that the evolved world-wide collaboration of developers, independent testers, the experiments and site admins had resulted in a much clearer specification of what an SRM should be and a well-established and agreed methodology to validate the protocol and any implementation of it.

After coffee, **Patrick Fuhrmann reported on dCache**, a collaboration between Desy and Fermilab, see dCache.org. As a WLCG storage element, dCache manages the largest fraction of LCG data. A key point o dCache is its strict name space and data storage separations. It allows for transparent replication on access hot spot detection and it has HSM integration. New features include SRM 2.2and the xroot protocol and in future they will add NSF 4.1 protocol which offers several advantages; the specifications are not fixed yet but it could get rid of clients. The eventual goal is to be an industry standard, like https for WAN file transfers.

Markus Schulz presented some tools for the management of stored data.

- gLite: open source; service oriented structure; access layer hides all the vendor protocols like rfio etc; soon to have a fully encrypted channels in FT, a prerequisite for people with sensitive data such as medical applications
- LFC: multithreaded C server, thread safe python and perl bindings; the largest LFC instance has more than 8Mio entries
- DPM: described as a poor man's CASTOR; it has GSI and VOMS based authentication; is http and https browsable; and offers as native protocol secured RFIO, but also GSIFTP, xrootd, https/http
- SRM - see Flavia's talk above, standard in gLite
- LCG-utils: command line and C-API, covers most common use cases, simple tools for accessing files
- Gfal: Posix-like C API; supports SRM V2
- FTS: "batch system for file transfers" between SRM-compliant elements; multi- VO service, VOMS aware; quite a bit of monitoring; for security it uses SRM/GridFTP credentials; it was used in data challenges, more than 10Mio files handled.

Mario Lassnig described experiences managing ATLAS data at the petabyte scale with DQ2. DQ2 is the entry point to all that data. Described basic concepts of DQ2 (for data subscription, etc), as well as its architecture (Common Framework, Client API, etc). Improvements of DQ2 since last CHEP: File Catalogue (common interface for various grid flavours). Authentication based on grid certificates, schema changes to accommodate partitioning, etc). The central catalogue has been moved to centrally hosted Oracle instances. There are various Site Services which use a parallel queue-based architecture. He listed challenges for the future. Among others Fail-over mechanism (retrial strategy, etc). The challenge is dealing with many small files. He was concerned with T0 exports during WLCG SC4 at the beginning of 2007 (T0->T1 tests). The throughput peak was above 1GB/s (a factor 5 improvement on previous tests) and they achieved a sustained 900MB/s over several hours, with ~ 600MB/s on average (factor 7 improvement). The plots in his talk are interesting, in particular that on slide 16. He believes that ATLAS is almost ready for data taking – his quote. Special thanks to all Tier1 site admins – and many teams at CERN, including the Castor team.

M. Bargiotti, Dirac data management: consistency, integrity and coherence of data. The DIRAC data management system has three components: the file catalogue (based on LFC) to know where files are, the bookkeeping metadata DB (BK) to know what is in the files and the storage elements (SE). These components must contain consistent information. Consistency checks are done by automatic Age nts or by a data manager. A Data Integrity Agent tries to fix any inconsistencies found.

M. Schulz, File Transfer Service. FTS is a reliable multi-VO data movement service, which balances the data transfers following VO and site policies. Built to handle many streams and large volumes. FTS is used by the experiment software frameworks (users do not directly interact with it, but production managers do). FTS uses delegation to act on behalf of the user. VOMS credentials are now used. FTS has a web service interface, a central DB, VO agents and channel agents. In WLCG, FTS servers are at Tier-0 and Tier-1's. Over 10 PB of data was moved in 2007. Perceived FTS errors are in fact often caused by other components such as SRM and several improvements are in the version 2.0 of FTS including support of SRM 2.2. A service type called FTM will provide better monitoring information.

I. Sfiligoi, glideinWMS - A generic pilot-based Workload Management System. glideinWMS is a Condor glidein-based WMS, developed in CMS. The glidein mechanism consists of starting a Condor worker on a Grid worker node as a Grid job, thus building a Condor pool of WNs spread over different sites. A glidein can also perform sanity checks, publish batch job characteristics, prepare the job environment. Glideins hide the Grid complexity from users, prevent obvious errors and simplify resource selection. The glideinWMS gets the list of jobs the users need to submit, and creates glideins on sites that claim to support at least one of the jobs waiting in the queue. In synthetic tests, glidein scaled well to ~4000 running jobs, then it was memory-limited on the submit node. glideinWMS worked well up to 100K queued jobs. glideins need outgoing connectivity to connect with the GCB (a Condor proxy server). A major drawback seems to be security: users could steal other users stuff or things from the glidein because all run under the same uid (the one of the glidein). In particular: a user can steal the credentials from the glidein submitter

- either run only jobs of the user who submitted the glidein itself
- or trust your users
- or proposed solutions: change the uid on th worker node (Q: does that break accounting ??? -> other talks later)

Work is in progress to improve scalability and resource consumption.

L. Zangrando, Job submission and management through Web Services: the experience with the CREAM[7] service. CREAM is a computing element implementation open to emerging standards, robust and scalable and with a lightweight architecture. It allows users to submit simple batch jobs, MPI jobs and soon bulk jobs. It exports a Web Service interface and also the OGSA-BES interface. All user commands are stored in the Journal Manager's backend, from where a pool of threads fetch the commands and execute them. Interaction with the LRMS is through BLAH, a generic interface to different LRMS. CREAM can be used with a client API, or via the gLite WMS through ICE (Interface to CREAM Environment). CREAM has recently passed the acceptance criteria defined by EGEE- a job submission rate of 10k/day passed, 8 days, 86k jobs submitted, no failure due to CREAM CE. Next steps are porting to SL4, to run more intensive tests, to finish the Condor-G interface.

Experience from a Pilot-based system for ATLAS - Paul Nielson. PANDA is the Atlas production and distributed analysis, uses pilot jobs with submission by Condor G, PBS or a new generic scheduler (AutoPilot)**.** The Pilot prepares/checks environment**,** pre-stages the data. It has automated error detection and kills user job if necessary and it cleans job directory. It sustains 10k jobs/day, increasing. They leave files in /tmp on the worker node, meant for job retry. This will not work at CERN because the scratch area is not /tmp !

DIRAC, Optimised Workload Management - S. Paterson. Dirac is the LHCB workload and data management system (WMS). It uses a pull scheduling, in principle a batch system on top of the GRID with one central task queue holding jobs waiting for scheduling, all of them can be very different**.** Pilot agents are used to create pool of resources, pre-staging *before* submission of the pilot jobs (!)

gLite WMS - Marco Cecchi. gLite is lightweight (?) stack of generic middleware which is useful for a variety of applications including MPI support even without a shared file system between the CE and the WN (remark: does MPI on the GRID make sense at all ? -> latencies are huge). The WMS interface uses web services**.** gLite now uses ETICS

[7] Computing Resource Execution and Management

automatic builds which makes migrations to new OS easier (eg SL3/32->SLC4/32->SL4/64)**. S**tress tests run 15k jobs/day over 11 days succeeded, peak 22k jobs/day. Questions:

- bulk submissions supported (but in condor already since several years, and condor can submit directly to a CREAM CE)
- synchronisation problem in the past: free resources on a site were picked up by several RBs, sending jobs to the same jobs and bringing it down this way. This is being addressed in the gLite WMS

The next talk was a so-called objective **comparison of workload management systems by Igor Sfiligoi from Fermilab**. He included Condor-G; ReSS, the OSG job matching interface which uses Condor-G; gLite and glideinWMS. He noted that submitting thousands of jobs to a Condor-G queue does not scale linearly when the queue already has more than 7K jobs in queue. He also noted that misconfigured sites can destroy all the jobs – is this surprising? gLite also uses Condor-G internally for job submission so had similar performance in many comparisons. But he claimed that gLite only allowed 5 job submissions per minute so was effectively unusable unless you switch to collection mode when it rose to 1000 jobs per minute but with frequent temporary overloads of the WMS. When asked which version he had tested, he did not know. The glidein WMS is pilot-based where the basic idea is that payload jobs are started on the worker node by a pilot job process which is submitted to the GRID. Glideins die off after 20min of inactivity (eg. if there is no work load). It is very fast, 200 jobs/minute once the pilots are there. Summary - a very useful overview over different systems

CRONUS - Sanjay Padhi. CRONUS uses condor glidein technology**. I**n case of firewalls it requires a GCB so that the call back works through fire walls. It is proven to work up to 5200 simultaneous jobs and it works also with other grid flavours, eg NorduGrid. But GCB is a bottle neck. Usage of virtual machines seems possible. In one incident a central system forgot about submitted glideinst, which piled up in the batch system and eating/blocking resources for nothing

Philippe Canal, The OSG accounting system: GRATIA. The OSG accounting system consists of a collector, which collects resource usage data from probes on the services and builds service usage records, a data store to save such information, an accountant to apply site policies and create accounting records, and a publisher to generate reports, graphs and plots. Gratia is installed on 2/3 of the OSG sites and collects information for Condor and dCache. For ATLAS and CMS, it also uploads usage information to APEL, the WLCG accounting system. The system is made robust by caching locally information in the probe and in the collector in case of communication failures. The backend database is regularly backed up. The database schema underwent several optimizations. The future plans include doing comparisons with the usage as measured by the VO software (e.g. PanDA).

Martin Flechl, Grid Interoperability: Joining Grid Information Systems. Interoperability aims at bridging different grids and enable resource sharing. Grid Interoperability Now (GIN) promotes interoperability among some grids: EGEE/OSG interoperability was achieved in 2005. The talk focuses on issues related to the information system. Achieving interoperability requires to have communication channels and translators from a schema to another. In the GIN architecture, information providers collect information from sites, or even whole Grids, and translate them to a common schema. Typical problems are no 1:1 mappings, which requires ad-hoc solutions. For the future, the goal is to converge to a common schema.

Julia Andreeva, Grid monitoring from the VO/user perspective. Dashboard for the LHC . The Experiment Dashboard project aims to provide a complete picture of the experiment activities in the Grid, combining Grid monitoring data and experiment application/activity-specific information. Some of the several uses of the Dashboard are described: job monitoring (including failure reasons), monitoring of specific activities (CMS analysis tasks, the ATLAS data management, the Monte Carlo production system of ATLAS and CMS etc.), I/O rate monitoring of analysis jobs, visualization of SAM test results, etc. Several new development are foreseen, including monitoring of Condor-G jobs and in general improving the effectiveness of the Dashboard for troubleshooting. The Dashboard is used in different ways by the four LHC experiments and even by other user communities.

There was a talk on **GridICE from Guido Cuscela**, described as a complete monitoring system for the grid. It started showing how it works. It collects hardware monitoring data provided by LEMON sensors and adds data from its own sensors on job monitoring and various other grid information. It has a web and XML interfaces for the results and can produce customized graphs and has a notification service. Using GIS it can automatically discover new resources. And it supports openPBS, Torque and LSF. There is different monitoring for users with different roles, respecting individual privacy rules, ranging from individual users to VO managers to site admins.

James Casey closed the day with a talk on **advances in monitoring in WLCG**. Monitoring can be considered in 3 layers, local fabric, grid services and the applications. The 2 previous talks in this session - dashboards and GridICE – covered layers 3 and 2 respecti vely and layer 1 is typically covered by NAGIOS, Ganglia, Lemon, etc. LCG established working groups to look into each of these layers. James is a member of the Grid Services WG which is concerned with level 2 and has the aim of improving the overall grid reliability by consolidating existing solutions, improving communications and reducing overlap. Initially they have been collecting information, what is needed, what is available and they are developing a prototype site monitoring system. They will define interaction points with existing systems, not aiming at standards (no time for this because they must get something out quickly) but rather specifications. They will not impose any solutions on a site but rather probes which should fit into existing frameworks. The initial prototype is with NAGIOS which many sites use but they have also done some work with GANGLIA. The prototype is being rolled out now and will be implemented as an option in a future gLite release. He invited the audience to join the early adopters, especially from smaller, Tier 2, sites.

Stephen Burke, Experiences with the GLUE[8] information schema in the LCG/EGEE production Grid. The GLUE information schema was born from the collaboration of the DataGrid and GLOBUS, predecessor of OSG, to promote interoperability between grids, for example WLCG needs to interoperate among EGEE, OSG and NDGF (at least). Its purpose is to describe the properties of Grid resources at a sufficient level of detail. The schema is defined in an abstract UML format, and the existing implementations of the schema are LDAP, R-GMA (Relational Globus), and XML. GLUE 1.2 introduced the GlueService concept, and gLite has a Service Discovery to query it. The schema works well enough, but there are issues in the mapping from real systems to the schema and some ambiguities in the way it is used. The most recent version of the schema contains several enhancements for SRM. Now GLUE is an OGF group, and a major redesign is foreseen for GLUE 2.0.

David Groep, gLExec, gluing grid jobs to the Unix world. gLExec can be thought of as a "Grid" verson of the Apache's suEXEC command. As suexec can be used to run an executable under somebody else's UID, gLExec can be used to run an executable under a given Grid identity (which is mapped to some UID). A typical application would be a CE running a gatekeeper with a non-root UID (like CREAM or GT4 WS-GRAM). A more interesting case is that of pilot jobs, which work only as placeholders of real jobs they download: gLExec allows the pilot to run the real jobs under the credentials of their real owners. The advantage is that the site knows who really ran a job in case of a security incident and that different users cannot interfere with each other. Future improvements include acquiring mapping information from a central service.

[8] Grid Laboratory for a Uniform Environment

Andrew Mc Nab, The GridSite security architecture. GridSite is a both a Grid security toolkit for C/C++ and an Apache module to add support for Grid credentials (X.509, GSI, VOMS, etc.). Andrew described some recent changes in the GACL format to represent all credential types in URI format. Other new feature new features include support for multiple forms of identity (extensible by URI naming of identity types - LoA and Shibboleth already added); and it allows VOMS roles to be included in websites (where certs have no VOMS roles) by downloading DN lists for VOMS roles from VOMS server. It also the ability to check an entire X.509 certificate chain. More documentation and cookbooks for lower-level code is now provided to allow better code re-use by others. Also logging is rewritten to allow output to syslog-ng, apache logging, A clearer API and C++/Perl/Python wrappers are on the roadmap. Support for Level of Assurance and Shibboleth have been added.

Riccardo Zappi, Enriched namespace to support content-aware authorization policies. The speaker describes a proposal to manage access control on files in an environment (like the WLCG one) with millions of files with many replicas. The idea is to use a role-based access control (RBAC) to map privileges to groups and roles, and user-defined labels to tag files (rather than just the file name and path). In addition to that, a language to declare access policies (XACML in the proposal) and a framework to manage, distribute and enforce the policies (G-PBox in the proposal) are needed. The user labels should follow a schema defined according to the Simple Knowledge Organisation System (SKOS). This kind of architecture could be implemented using only available technologies.

Alberto Forti, Cross middleware VO authorization. The Virtual Organisation Management System (VOMS) is used by several Grid middleware stacks (gLite, VDT, etc.) and in many grid infrastructures (EGEE, OSG, D-Grid, Naregi). It allows to assign groups, roles and attributes to users, and this information to be used by Grid services to implement access control to resources. The current implementation of VOMS does not expose a web service interface and uses proprietary XML messages. It works by attaching to a grid proxy the attribute certificate (AC), containing the information about group and role attributes. VOMS is being extended to support the Security Assertion Markup Language (SAML), which requires a web service interface and allows to decouple AC from proxies (to support some Grids which do not use proxy certificates, like UNICORE 6).

Dirk Duellmann, Distributed database access in the LHC computing Grid with CORAL. CORAL is a high level interface in C++ and Python to several database backends for physics applications. It also offers some high-level services, like authentication, authorization and DB service lookup. It is used by POOL and COOL, but also directly by ATLAS, CMS and LHCb in their software. As backends it supports Oracle, MySQL, SQLight and FroNtier/Squid. A particularly interesting feature is the ability to implement VOMS-based authorization even for backends that do not support it directly. Future improvements include investigating a multi-threaded CORAL server for better scalability and security.

Unified Storage Systems for Distributed Tier 2 Centres - Greig A Cowan. The idea is to use POSIX like protocols to access storage over WAN using RFIO . Measurements were done varying number of readers, TCP parameters, RFIO buffer sizes. Conclusion of the tests: it is possible, with a good utilization of the production network of 60%.

Ian Gamble from Uni Victoria discussed using xen with HEP applications. Xen's so-called "para-virtualisation" permits most instructions to run at almost native speed and has negligible performance penalties compared to VMware for example. After an abortive attempt to build their own middleware, they found that Globus virtual workspaces was a match for their needs. He showed the workflow to distribute the image of a worker node using the xen kernel. He admitted that work was an early stage and had some limitations – no authentication of images which are to boot (although grid certificates should solve this), security issues. Redhat V5 now contains a tool to create virtual machines and there are many other ways to build VMs. Tests are at an early stage but they are encouraged thus far and conclude that virtual machines can be used to access resources they would otherwise not have access to. But more research and testing is needed, especially security aspects.

Geographical Failover for EGEE/WLCG Collaborative Tools - Alessandro Cavall. Started with a definition of availability margins in different businesses, see table slide 4, and downtime causes on slide 5. Uses redundancy and SPOF removal (=Single Point of Failure). Common problems of DNS failover include latencies in ISP due to caching and caching inside browsers (eg. IE 30min). Tools are available at http://www.gridops.org**.** Used for CIC portal (IN2P3 and CNAF) and others. Currently, the switch is done by an operator by hand; later automatic switch. The tricky bit is the replication of the backends, eg Oracle databases. Currently done by hand with low frequency, later automatic

R. Stober, Multiple Heterogeneous Clusters with LSF. This talk by Platform shows LSF Universus, a tool which allows to submit jobs from an LSF cluster to a remote LSF cluster. Among requirements, use of Kerberos auth. and encryption is allowed, though communication among LSF daemons is not encrypted. A set of new commands are added to be able to monitor the remote job. Files are copied over by just scp. No shared filesystem assumed. The tool is also able to act as a “meta-scheduler” on top of a variety of workload management systems, provided that LSF is installed on the head node of the remote cluster.

7. Software Components

Giacomo Govi described CORAL , a Database Foundation Layer. He explained the goals and origins of CORAL. And new applications using it in ATLAS, CMS, etc as well as POOL and COOL and now also interest from outside HEP. It supports several DB back-end: Oracle 10gR2, MySQL 5 and 4, SQLlight v3.4 and FroNTier/SQUID. Interesting features include

- CORAL Python Binding...quickest way to prototype new applications.
- CORAL Copy database tool...

He also talked about new milestones which will not imply any changes to user packages already used. He noted that accessing DB's through CORAL will expose less DB servers in the Internet.

Andrea Valassi continued the theme by talking about COOL. He introduced COOL v2 and listed the experiments using it as well as collaboration with other software projects and the supported DB back-ends. He listed changes in the last months to improve performance, remove SEAL classes, 64 bit platform support, etc. There have been also relational schema changes, again to improve performance, functionality, etc. COOL provides support for new platforms: 64 bits linux, MacOS intel, etc. COOL is now more Grid aware since it has improved DB authentication and look-up. Several Server-side SQL optimisations and client-side profile optimisation now provide a better performance. Finally he covered distributed deployment on t0 and t1 sites (not yet fully finished).

Analysing CMS software performance, Lassi Tuura. We used to think that “memory is cheap”, it is clear from this conference that this myth is no more valid and 1GB per job is already a lot. We need to measure, analyse and identify performance problem before making changes in our code. We do not have many corruption problems, and that is mostly thanks to Valgrind. The analysis made on the CMS software shows that a lot of improvements can be done on the analysis software. 1.5 Millions lines of code need to be revisited. This can be eased by the fact that an important fraction of the code is being changed. A lot of the time is spent at memory management: 1 million times per second allocation and free memory (1/4 of the time). 66% of all memory allocations were by a matrix and vector package used in the reconstruction algorithm, it was “easy” to fix. Another problem is that the CPU is idle 60% of the time; it is interesting to note that some of the programs running IBM Blue Gene/L achieve 70% of the raw CPU capacity. One of the next hot topics is I/O performance. It is only the start of code performance analysis and improvement. Huge TLB could help, it is interesting to evaluate it.

Optimizations in Python-based HEP Analysis, Sebastien Binet (LBNL). Sebastien noted that flexibility in software is great, but only if it is needed. If not, then the performance penalty that it often brings about is just a burden. In Python, many temporary variables are created internally, types are considered dynamic and no block-level optimisations are applied. Optimisation of HEP Python code with some tools is difficult, because many handle only the native Python code, while the HEP code usually uses binding to other languages. Sebastien went through several tools and summarised their capabilities and limitations and presented a methodology to successfully optimise HEP Python code.

Searching for CMS data: Rapid web development using python and AJAX, Valentin Kuznetsov (Cornell Univesity). Valentin noted that there are many technologies for development of web applications and it is very important to choose the technology that suits your particular needs best. In the particular example that he mentioned, Java (under Tomcat), Python and AJAX have been chosen. Valentin underlined the importance of an intuitive and optimised user interface and showed how it can be easily implemented with Python and AJAX (with the help of RICO framework and the Yahoo Development Library). The important point of his talk was that AJAX, which stands for Asynchronous

Javascript And XML, and which is in fact a generic name for the new approach of developing web applications, can be easily used together with Python

The life cycle of HEP offline software – Peter Elmer. Talked on software cycle for three experiments (BaBar, CDF, CLEO) and one experiment about to start taking data (CMS).Figures on developers involved, CVS commits activity, etc. Showed graph on programming languages used and its importance (lines of Fortran, c++, Phyton, java, etc code) as well a quantity of code dedicated to Analisys, Trigger, Online, tools, ... in each experiment. Shows developing effort in terms of CVS commit operations. CMS is mainly c++ !!! Ended up with some conclusions just pieces of advice (mainly for CMS):

Perfmon2 - A leap forward in performance monitoring, Sverre Jarp. We need software monitoring. Though CPU power is not as expensive as people still we should get as much as we can out of it. Recall that expensive people sit in front of cheap hardware !!! New hardware comes with process monitor units (PMU). Something must be in the linux kernel to interface with that PMU. That bit is perfmon2, initially designed for use with Itanium chips. But it took years to have full support included in the Linux kernel as one had to be very careful to not slow down the kernel. The full implementation is about to be included in the Linux kernel 2.6.23. The user level commands will be the same on every Linux platforms, but the counters might be different. A GUI has been developed to make it easier to understand the results. We have established a test service and a number of machines at CERN can be used for it. The command interface pfmon can be used with a lot of sophistication (kernel, triggers, attaching to processes); it is the same in all supported linux platforms and allows finding ratios (cycles per instruction, floating point instructions, L2 cache misses, etc).. The overhead is below 1%. For Scientific Linux, we have to wait for RedHat to include it (which might take several years).

Explicit state representation and the ATLAS event data model: theory and practice – Marcin Nowak. Atlas offline software Athena currently uses POOL and ROOT for event data persistency. The problem is schema evolution and the solution is the persistent state representation. Schema evolution happens all the time since data model is sometimes occasionally redesigned. Ideally ROOT automatic schema evolution should cope with that but sometimes it does not manage. Solution is to create an independent persistent data model which includes schema version as typename. Showed persistent Data model structure (showing transient and persistent versions) and explained how Transient/Persistent converters work. Typeless persistent references were needed to implement this approach, but then ROOT could not follow and a container object needed to be created. Top Level Storage objects structure was shown. There are 3 types of schema evolution provided (picture shown). Ended showing the AthenaPool complete persistence Stack. Then performance was improved obtaining pretty impressive performance figures with 10-15MB/s read speed as compared to < 1MB/s even though performance improvement was not the initial aim. Described AthenaROOTAccess (automatic use of T/P converters without Athena). Atlas EDM is there in Athena release 13. Time will tell how important schema evolution support is. Questions/Comments: Rene Brun thinks this is a patch to solve a temporary problem. This will create huge reference problems in the future.

CMS packaging system or: how I learned stop worrying and love RPM spec files – Giulio Eulisse. Needs for Software packaging and distribution for reconstruction software for each person involved in that process...In conclusion, we all know it is very much needed. CMS has some 80 packages (gcc + externals, online software, etc). It is 1.7 GBytes of rpm packages (3.5GB after installation. Explained problems of central installations (mostly related with AFS) and how it was extended but was not good enough. Then went to building RPM from CVS source code. This solved all problems of the previous procedure.They use packaging tools which are better than plain rpmbuild. Use RPM and apt-get with a glue-script "cmsBuild". This was a very nice rationalization of CMS software distribution !!!! Future: New version of PKGTOOLS coming. Simplify integration and configuration management. Parallel build of non-depending packages. distcc support and more platforms (MacOS) supported**.** They use a private package database !!!.

8. Distributed Data Analysis

Andrew Maier described Ganga - a job management and optimisation tool. It is a tool to submit and manage jobs to different backends (batch, Grid, etc.) transparently. It has a Python command line and a graphical interface. It can support specific applications through dedicated plugins. In LHCb it is used to submit analysis jobs using DIRAC as a backend. It is used by several other applications, like lattice QCD calculations, Geant4, ITU, biomedical. More than 10% of the WLCG users use Ganga. There is an experimental Windows port.

Akram Khan described Distributed analysis with ASAP. This is a server to manage and monitor Grid jobs on behalf of the user. It is capable to resubmit failed jobs, avoiding sites which failed previous jobs. It is used by some people in CMS. It is interfaced with the CMS data management. The server is secure and uses the proxy delegation mechanism to operate with the user's credentials. The monitoring is done via the dashboard.

J. F. Grosse-Oetringhaus talked about a PROOF cluster for day-one physics analysis of LHC data. PROOF allows transparent interactive parallel analysis on a cluster, with a special focus on fast response time. It is used in the ALICE CERN Analysis Facility for prompt analysis, fast simulation and reconstruction, calibration and alignment. It consists of 500 CPU and 100 TB of local storage. The PROOF cluster is also an xrootd pool. Data to analysis is currently manually staged from CASTOR, but in future the staging should be automatic. PROOF supports disk quota and CPU fairshare based on user groups. The monitoring is done with MonALISA.

Next was **Stuart Paterson on Distributed data analysis in LHCb**. LHCb analysis includes stripping (a selection of interesting events from a larger data sample) and user analysis of stripped data. It is done only at Tier-1. The job submission is done through DIRAC, which implements a pull mechanism where pilot jobs get real jobs from a central queue. Policies are applied to this queue. DIRAC performs checks on the correctness of the job definition, resolves input files and pre-stages them when needed before job execution. There are more than 70 active users using about 100 Grid sites. In average, 70% jobs are successful, 20% fail with data access problems and 10% of jobs abort or remain stuck for any other reason.

Another Ganga talk, this time its use in ATLAS, given by Johannes Elmsheuser. ATLAS uses Ganga to run analysis jobs, customized to interface it to the Athena framework and the DQ2 data management system. In EGEE it submits to the gLite WMS, in OSG via Panda and in NorduGrid. The jobs always run near the data (which must be replicated where needed in advance). There are about 60 active users per week, at over 50 sites.

Adam Kocoloski talked of Integrating Xgrid into HENP distributed computing model. Xgrid is a distributed computing architecture from Apple, makes easy to build clusters. A controller accepts jobs from client and sends them to agents. Only one port needed to data transfer. Supports MPI jobs. Job scheduling is just FIFO. Used at MIT to usage CPU cycles from Macs in campus. Libraries accessed via NFS. STAR scheduler is used, supporting LSF, CondorG and now Xgrid. SUMS is running stably since march 2006, 50 machines, used by many research groups (also CMS). A preliminary support in Globus GRAM is available, but still work needed to allow the use of GSI certificates. The goal is to use it in the context of a Grid infrastructure.

Leandro Franco spoke on efficient access to remote data in HEP. Reading from a ROOT tree remotely can be very inefficient in case the network latency is very high. A modification to the ROOT client and server was developed in order to stack together a list of buffers to be read which are scattered over a large remote ROOT file, and send it in a single transaction to the server. The modification was ported also to http, xrootd and dCache.

Lassi Tuura presented scaling CMS data transfer system for LHC start-up. The talk described some recommendations on how to build a robust data transfer system based on the choices made in the PhEDEx architecture. CMS currently transfers ~100 TB/day.

- System architecture choices: to send the jobs to the data, to manage not single files but data blocks (5-10 TB), to use the storage itself as file catalogue, to protect the system from failures.
- Technical architecture: to have an agent-based system where communications go through a central database containing just the current and the desired state of objects, a very smart error handling and an efficient hierarchical monitoring;
- database engineering: lots of optimization work to minimize the need for updates and replace them with inserts to lower row lock contention
- validation: dedicated validation testbed to verify functionality and error handling at loads 100 higher than expected load.

PhEDEx is satisfying the needs of CMS for the next two years and most of the current work is to improve the web interface.

Development, Deployment and Operations of ATLAS Databases, Alexandre Vaniachine (ANL). Alexandre confirmed that Oracle RAC has been chosen as both the online and offline databases for ATLAS. He listed some of the ATLAS database applications that must be distributed wold-wide for scalability, as they are accessed by many computes on the Grid: Geometry DB, Conditions DB, Tags DB. Then he went on to enumerate the centralised database applications: Technical Coordination database, Online Database and various Computing Operations databases. He went on to describe the aims, requirements and challenges of these applications. Alexandre reported that all ten ATLAS Tier1 Sites are in production operation, leveraging the infrastructure of the 3D project.

Douglas Smith on Data management in BaBar. Douglas subtitled this talk – "life without a database"**.** Douglas recalled the historically first data storage model, which was based on Objectivity databases, which was put in use around 1998. This proved not to be scalable, and BaBar finally migrated out of Objectivity and now Root and MySql is used instead. As a result, the system has been simplified considerably and the conditions data, stored in root, is now distributed in the same way as other data. All BaBar data is contained in the event store, files are organized in collections, whose metadata are in the BaBar Bookkeeping system and collections are organized in datasets. Changes to datasets are recorded in a database. The talk describes how the system is able to quickly update datasets to reflect changes in collections and how datasets are tagged to provide a fixed set of data. Douglas ended his talk with the statement that transition out of Objectivity was painless, but only after years of painful development.

Andrew Smith on LHCb data production management. The talk describes how DIRAC works as a data management system for LHCb. It is used for:

- moving raw data from online storage to CASTOR
- performing centrally managed bulk data replication between tiers (requests are put in a TransferDB polled by a TransferAgent which submits transfers to FTS); a Stager Agent pre-stages files before transfer;
- performs automatic replication and processing of RAW data to tier-1 sites.

New Developments of ROOT Mathematical Software Libraries by Lorenzo Moneta. A complete new structure of ROOT mathematical libraries. Has 5 different blocks (MathCore, Linear Algebra, Statistical Libraries, Fitting and Minimization and Extra Libraries). The above implement new function Interfaces (with the new "Functor Classes").

Several examples of using Functors classes were given. He reviewed use of Numerical Algorithms (Numerical Derivation, Integration, Root Finders, etc).Also a new pseudo-random numbers generation. There is a new UNURAN package to generate non-uniform pseudo-random numbers. Performance figures of random number generation were showed. Physics and geometry vectors (Classes for 3D and 4D). SMatrix package (it now support symmetric matrices). Figures on Matrix Operation Performances were showed as well as on solving Kalman filter update equations. New Fitting in ROOT. ROOT Data classes can be fit directly with a new set of options. A new GUI (Fit Panel) to easily fit user defined functions. New coming features, Including advanced draw options and many others (particularly new Fit extensions). New Statistical Tools were mentioned very briefly.

CERN database services for the LHC Computing Grid, Maria Girone . The service is based on redundant hardware, 110 mid-range servers (Pentium 4 Xeons) and 110 disk arrays (~110 disks). The service relies on 5 DBAs and 24x7 service based on “best effort” monitoring and operation. Adding CPUs or storage is a transparent operation. The evolution will be likely based on multi-core CPUs and RHEL 64 bits. Tests with a quad core server have shown that upcoming CPUs will be very performing (tests on Phedex). On-disk backup are used with 2 days retention. Automatic test recoveries are in place (several times per month). 25% of the raw disk space is used for the database, the rest is used for the backups. The hardware will be late spring / beginning summer 2008.

The ALICE Offline Environment by Federico Carminati. According to Federico, the Alice computing problem is as big as that of Atlas and CMS (although it is less expensive). Describing AliRoot and computing model, he explained the software process in Alice (releases, nightly builds, etc). Moving to Subversion[9] in coming months. He explained the simulation model with an alignment framework which allows alignment-aware reconstruction and he explained data reduction process in ALICE as well as at the CAF[10], and gave a high level overview of the data-flow. Alice GRID middleware is AliEn (95% of its code was not developed directly by Alice). Detailed explanation of all job submission process in LCG for Alice. Alice never distributed software, it is automatically installed on demand. Xrootd is the only access to the data in Alice. Monitoring is based on ManALISA and they are embedded in the dashboard. Unfortunately there is no common application layer between all experiments (that was in the wish list back in 2003). Next challenge is exploiting virtually in the grid. Federico gave an interesting list of DONE items and things that could have been better (too many grid middleware in each experiment). The Grid problem is not yet fully solved.

Relational databases for conditions data and event selection in ATLAS – Florbela Viegas. Gave an overview on Atlas DBases.First ATLAS requirements for DBases (size, etc). Need for a RDBMS. Using CORAL. 3 different RDBMS can be used, production server is though running on Oracle. Described Conditions Data framework. 1 million jobs/day!!! Using Oracle Streams Replication technology. First test very successful and reliable. Also client stress test were performed in IN2P3 reaching 400-4000 jobs/hours. Event Metadata: Using Relational TAGS. Improved query to 1 PetaByte database (this was really needed, sometimes queries were very inefficient). Many Challenges ahead, particularly improving Backup and recovery, availability offline, etc. Strongly believes that everything will work (intensive testing has been done, right amount of resources dedicated, etc)

Assessment of Data Quality in ATLAS, Michael Wilson (CERN). Validation timescale is 1-2 days after data are recorded. Calibration will take ~24 hours. Histograms will be uploaded both at CERN and to the Grid. A configurable calculator makes automatic assessments. Detector commissioning with cosmic-ray data is ongoing. ATLAS will be ready to assess data quality as soon as the data are available. First implementation of tools are ready ; tuning and fixing is underway

[9] A replacement for CVS for code versioning

[10] Central Analysis Facility, at CERN

Manuel Guijarro gave the IT/DES experience in running high availability of databases on NAS[11]. NAS offers interesting features for database use – snapshots, RAID, failover, dynamic re-sizing, etc. Also it is both cheaper and simpler to configure than the fibre channel-based SAN[12]. A facility called Bonding or Trunking exists to aggregate multiple network connections to a single logical channel which is faster than any physical connection. Direct I/O over NFS is used to access all Oracle DB files. Stress tests reading 8KB blocks using multiple threads with direct I/O over NFS achieved around 150 I/O operations per second per disk (12 disks). The NAS is very stable, no downtime and easy to maintain. Manuel’s talk provided interesting technical offset to the commercial presence of the NAS vendor, NetApp, who had its booth at the conference.

Lucas Taylor on CMS Control and Monitoring Centres: CMS intend to co-locate the very many people involved with startup in different places. They foresee at least two remote control centres, one on the CERN main site and a second at Fermilab and maybe also one at DESY and/or elsewhere. These remote centres will have mirror displays to the main CMS control room and lots of data quality monitoring. They will also monitor grid operations for CMS and perform express analysis. Obviously good communications will be vital and must be secure to protect machine functions. Methods include web tools, web cams, shared desktops, etc. The Fermilab centre is open and the CERN one in the process of being built where the old PS control room was. Lots of the physical infrastructure in both remote centres will be the same as or based on the main LHC Control Room.

John Kennedy described the ATLAS monitoring system. Jobs are stored in a central database, grouped into “tasks”. A supervisor retrieves jobs and passes them to a grid via an Executor which adds wrapper for specific grids. The supervisor handles job retry. They have built a monitoring scheme to offer different views of the data using a stripped-down data extracted from the main monitoring database. In particular, a last 24 hour snapshot has been found very useful in problem solving because one can drill down to investigate failures. Also intelligent monitoring of the monitoring system to see which parts of the monitoring system use most (and check the DB isn't overloaded) - Gridview should take note! Will move into the ARDA dashboard - but keep the good stuff, like the rolling snapshot tables. Already an EGEE shift scheme has been implemented to monitor grid usage.

Ricardo Rocha - Monitoring the Atlas DDM System. Have developed and deployed a distributed system to gather monitoring status information for all Atlas data operations.2 million messages per week (bulk - 10-500 items per message). Usable by site admins, end users and production coordinators. As well as web interfaces, have python interfaces which allow for flexible queries e.g. operators using it to create alarms. They internally monit

[11] Network Attached Storage
[12] Storage Area Network

9. Globus BOF

Dan Fraser from Globus held a BOF on Tuesday morning describing changes in the structure of the Globus Project and some interesting technical improvements in the core Globus products. Globus has moved much more to a community mode of development, with a "Incubator" for external projects. One major contributed project is in fact European, the spanish meta-scheduler GridWay, which is now being integrated into the core Globus toolkit. Also mentioned was a move away from a toolkit approach to "Solutions for Science" - systems delivered to solve specific computing problems for particular science applications. On the technical side, one core improved is better support in Gridftp for UDT, a high-performance fair network protocol, based on UDP. This is said to give x5 speedup over parallel stream gridFTP on congested networks. The support will be made more transparent to the user, and soon will be configured by default as an option in the server.

10. Special IBM Seminar

Towards Petascale and Exascale computing, Jim Sexton IBM. 2-4 GHz will not change significantly. As we add cores to a chip, it is increasingly difficult to provide sufficient memory bandwidth. Blue Gene tops the top500 list of flops/power, 21 watts/chip, 5 watt/core (5x less than AMD). Memory will be the dominant cost for upcoming computers. Every vendor is using the same DDR memory chips. The programming paradigm will change to use less memory and move it less. Jim made a thought-experiment about the possible design of an Exaflop computer (for a single application). Scaling BG/L to an exaflop, it **would** lead to 32 MegaWatts. 50 kWatt is the limit of dissipation per rack, 83 with water cooling. The processor **would** need to be ~240W processor chip-stack with 6 stacked chips, each with 64 sets of 3.3 GHz core + 8-way Floating Multiply-add unit. The number of B/Flop ratio **would** drop from 1 to 0.06. The memory **would** be (density, speed) a major problem for which major innovation would be needed. One way to interpret this exercise is “Moore’s law is about to end?”. To get to an Exaflop, computers **would** need to be made of custom components (=end of commodity). To get and Exaflop computer

Blue Gene, Jim Sexton IBM. What OS on Blue Gene? The fact that Blue Gene should have a MTBF of 7 days, with 128k processors leads is means 64k OS instances. A dedicated lightweight kernel was built out of glibc with some limitations, for example it does not have fork system call. The service node is running a full Linux kernel (and a DB2 database). With Blue Gene, you have the choice to have 4 processes per chip or 4 threads per chip. OpenMP will become very important in the coming years. IBM XL compilers is used, it is the same compiler used on AIX. IBM engineers were worried that it would be difficult/impossible to run/manage a system with 72k systems, they created a very hierarchical management layers. 256 Blue Gene/L is the field (medium size is 8-16 racks), lot of interest within petrol industries. Financial industry has a problem with code management; one of the difficult areas is how to move data out of database. It is based on 32 bit architecture (floating point is 64bits), an emulator exist for Linux Intel. Blue Gene/Q will be a 64 bit architecture. Cell is interesting and fast for embarrassingly parallel problems like the HEP typical computing.

11. Posters

Integrated RT-AT-Nagios system at BNL USATLAS Tier1 computing center, Tomasz Wlodek (BNL). The poster described the use of three open source products for monitoring and book-keeping of incidents at BNL:

- NAGIOS, free configurable package for monitoring networks, machines and services

- RT, open source request tracking system, available from a commercial company (BNL is not paying for support of this software and there are no licensing fees)
- AT, a module of RT provided by a third-party, used to keep track of BNL hardware assets, their history and the list of their administrators.

Usage and extension of Maven to build, release and distribute Java and Native Code Projects, Mark Donszelmann (SLAC). This poster summarised Maven, the popular tool that handles all phases of the build lifecycle of a software module. Maven can be used to release modules and it takes care of any interaction with the source code management system. Maven is also capable of compiling documentation into webpages and PDF. Maven is extensively used by the FreeHEP projects including AIDA, JAS and WIRED. Mark particularly stressed the use of Native Archive Plug-in, which extends Maven to handle C++ or JAVA JNI projects. Dependencies for specific platforms are downloaded and unpacked in the local repository so that native compilers can access include files and link to native libraries.

Use of Fow Data for Traffic Analysis and Network Performance Characterization, Andery Bobyshev, (FNAL). The poster presents the methodology to troubleshoot network performance issues in the distributed environment of HEP experiments. It is based on:

- NDT (the Network Diagnostic Tool), used to collect various TCP network parameters and identify their configuration problems, or the problems of local network infrastructure such as faulted Ethernet connections, malfunctioning NICs, etc.
- OWAMP, the Network path diagnosis server.

These tools are based on the Linux kernel extensions which are provided for every new version of Linux kernel. To be able to diagnose a network problem, at least one machine with this extended Linux kernel has to be installed in each of the affected networks. However, the machine can be booted from an image containing this extended kernel when the problems actually occur. Andrey confirmed that machines running this extended kernel are present in many HEP networks, including the experiments at CERN.

12. Social

Apart from the short Welcome Reception on Sunday, the main social event was the Conference Dinner on Wednesday, held in the neighbouring British Columbia Museum where we could walk among the exhibits while eating, described in the programme as a “strolling feast”.

There were no sessions on Tuesday afternoon and many delegates participated in whale watching trips, with some success as shown.

Alan Silverman

19 Sep 2007