CHEP'09 Trip Report

Prague, March 23-27, 2009

Simone Campana, Jose Benito Gonzalez Lopez, Andrew Maier,
Ricardo Salgueiro Domingues Da Silva, Alan Silverman (Editor), Juraj Sucik

Introduction

There were 615 attendees from 41 countries who throughout the week presented a total of some 560 papers and posters. The Symposium was preceded by a WLCG Workshop with 240 participants. Despite these impressive numbers, many plenaries were badly attended and in general the number of people queuing at breaks seemed to vary significantly from day to day. Regarding plenaries, the question must be asked (and I intend to ask it during a future IAC discussion in preparation for the next CHEP) whether we should not reduce the number of these and devote more time to parallel talks where we are clearly short of capacity. This time for example, 500 paper offers were received to be assigned to 200 oral slots with the rest being shown as posters which many authors regard as second-class. Yet the main room was never busy for the morning plenaries, a number of which should most definitely have been scheduled in one or other parallel session, aimed a more specialized audience. Already in discussions with other IAC members and others I sense a fair amount of support for scheduling fewer plenaries.

This report, compiled from contributions by the various authors listed above attempts to cover most streams of interest to CERN/IT Department but we cannot claim to cover all sessions, all talks. The editor would like thank all contributors and apologises in advance if he has misunderstood some point or made some other error in transposing the material[1]. Most overheads from the meeting are available on the web

[1] For one thing you will find a mixture of UK and US English because some of us have different default settings and for this report there is limited effort to search out zzzz's.

(http://indico.cern.ch/sessionDisplay.py?sessionId=1&slotId=5&confId=35523#2009-03-24) and the reader is invited to consult those of sessions of particular interest.

Plenaries

The Symposium was opened by the Managing Director of CESNET who introduced in turn the chief executives of the various host organisers of the Symposium including the Czech Academy of Science and the Charles and the Prague Technical Universities each of whom presented his institute or university. Charles University in particular dates back to 1348 and has a rich history, not only in science research.

The **opening technical talk of the conference was given by Sergio Bertolucci, CERN Director for Research and Computing**. Sergio reviewed the startup and initial running of LHC, the accident of September 19th, the steps being taken for the repairs and to avoid any repetition, and the plans for the restart. Quench protection has been increased by a factor of 3000, with the sensitivity rising from 1 volt to 0.3 volts. Not having a winter shutdown next year will cost some €8M in electricity costs. He explained in some detail the reasoning for limiting the energy to 5 TeV until the end of 2010, basically the number of quench cycles that will be needed to re-train the magnets to 7 TeV. He compared the work being done currently at Fermilab and how CERN will learn from this in the Higgs search. Computing for LHC was summarized in one optimistic slide.

Neil Geddes than posed the question "Can we deliver", referring to the computing for LHC. He noted that 33 countries have now signed the MoU, which is a measure of the resources being provided at the different Tier centres, but availability is mixed and total resources, although lower than experiments' requests, are within the error bars of what is required, at least in the short term. He showed impressive plots of steadily-increasing grid usage and data transfer rates so he concluded that the grid can and does perform, although further improving reliability is a continuous focus. Although the take-up of common software in the wLCG Application Area has been a success there is some frustration about the multiplicity of grid stacks and some confusion among the user communities, who are building their own high-level frameworks. He summarised that the wLCG has built and operates a successful grid infrastructure for the LHC experiments but what will happen when there is massive usage of the resources when production data becomes available for analysis? So, yes, wLCG can deliver, but there are challenges left to face.

Kors Bos described the status of LHC Experiment Computing, starting by comparing the various computing models, all of which started from the original MONARC tier structure but some of which have diverged in how data and jobs flow between the different tiers, in particular ALICE. All benefit from the excellent networking available and he does not consider networking to be a particular concern. The SAM tool has proved particularly valuable in monitoring grid metrics and he put up some SAM graphs which showed, among other things, that ASGC is gradually recovering from their recent fire. Finishing on the challenges to come, the quantum jump from cosmic ray runs and the successive computing challenges to a solid 12 months of production running is a real concern in terms of practical problems such as site availability, especially Tier 1 sites. Also the combined load on tape services is an open question according to Kors. He was happy that the weekend's wLCG workshop had agreed to a combined test of the tape system in May and June. The third practical problem concerned how the overall service will cope with the large number of analysis users expected once production running begins.

The last plenary of Monday featured **Les Robertson on how we got here and what's next**. Like Kors, he started from the MONARC distributed model and how that evolved into a very simple grid, first presented at CHEP in

Padova in 2000. He described the 2000s as the decade of the grid. The "grid", in fact a relatively small number of broadly similar grids thanks to the development and adoption of standards, has developed and matured and an increasing number of sciences and industrial applications have made use of it. But is the model of a general science grid sustainable? In particular, is EGI the correct model and will it be ready in time to take over from EGEE next April when EGEE is due to end? But MONARC was 10 years ago and the environment has changed radically since then as shown in a slide comparing computing technology then and now. While back then HEP led the need for massive computing, now HEP is a relatively small player compared to many commercial firms. Les thinks we should be looking at

- energy usage, locating grid centres where energy is cheap
- virtualization, sharing processing power better and permitting more flexible hardware installation and software upgrade planning; virtualization also offers portability of applications
- clouds, what are they in comparison to grids? They both need complex middleware but whereas grids are transparent, clouds are opaque and these are early days for clouds so we need to be aware of them, how they may develop and how they may be usable for our science
- mobility, where bandwidth to the home is now as good as it once was to computer centres 10 years ago and notebooks are powerful enough to run analysis jobs. But extending grids to notebooks is too complicated so perhaps we need to look how they could be integrated in some other way, using virtualization perhaps.

Tuesday was opened by **Ruth Pordes talking about grids, clouds and "collaboratories**". Many small groups of researchers would make more use of grids (and Ruth includes clouds in her definition of grids in this general context) if we could make them easier to use, especially with an easier way to get involved quicker. The US embodies this in an initiative for a national cyber-infrastructure which should establish an environment to share campus clusters and small local (e.g. campus) grids. She distinguishes "collaboratories" from VOs because the former include not only the researchers but also the computer scientists providing the infrastructures and the former provide feedback to the latter to improve the system. She sees more need for interworking between collaboratories and between user communities. So bridges are needed between grids, not simple gateways but gateways with additional capabilities such as filters, data validation, traffic management, job control. A local example is a Fermigrid OSG to TeraGrid gateway for dispatching work into TeraGrid. Another is the US NSF-funded Nanohub (nanohub.org) which uses a portal model to enable fast and easy access to resources and claims more than 90,000 users so far.

Erik Gottschalk from Fermilab discussed operations centres, especially from the viewpoint of remote operations. He described the original LHC@FNAL concept of 2005 and how it evolved into a joint LHC and CMS remote control centre. Innovations were needed to ensure secure and authenticated access and authorized control (e.g. role-based access) and others for collaborative working (e.g. a screen snapshot service). He then expanded on collaborative tools with examples not only from other physics experiments but also from other fields, a number of which he had adopted at FNAL. He believes that such tools, and he gave examples, will encourage increasing remote collaborations as a trend for the future. The gloss was slightly dimmed by a questioner who said that Fermilab was the hardest lab in the world to work at largely because of network protection measures and asked what would be available at Fermilab in 4 years that was not available at other labs now. Another questioner was concerned about the proliferation of tools and whether some top-down structure should be created.

Next came a review of the **experience of running Monte Carlo production for Belle on the Amazon EC2 cloud**. The task is MC production for the proposed SuperBelle at SuperKEK (a proposed major upgrade for KEK). The experiment estimate needing 10 to 80 times more computing than today. MC production seemed suitable for cloud computing. The speaker had devised a so-called "value weighted output" which measures the value of work produced compared to the purchase price. As technology improves, obviously this degrades over time as clusters age, cost more to operate and depreciate not only in value but also in what can be produced per dollar compared to more recent equipment. This measure added more justification to investigate a cloud service. He listed the characteristics and costs of EC2 and how to interface to it via a browser. First step was to build an Amazon Machine Image (AMI) and in fact they created the first Scientific Linux AMI which is now freely available. He described the build process for various configurations of AMIs and the first tests. In the most efficient AMI configuration they estimate that the cost of generating MC data for a 3 months Belle run (10^9 events) was around $47K and they estimate the cost of buying and running the required hardware to do this in-house would cost twice as much. So they went ahead and built a full-scale cloud MC production chain and he described this in some detail. In preparation for this talk, they have produced some 750K events in EC2 at a cost of some $80 in CPU power and a total cost of $86.80 including 20 cents storage and $6.60 data transfer costs. They are working on tools to keep their pool of AMIs always active and to minimize the time the results data is stored inside EC2 before retrieval. And a faster data transfer link to and from Amazon would be very welcome (although apparently accessing Amazon from the US is notably faster than from Australia).

IBM, main sponsor of the event, presented ideas on "cooler, denser and more efficient computing power", otherwise known as iDataPlex, for which the speaker was CTO. By and large, there then followed a sales talk on iDataPlex with perhaps a bit more technology than the average sales presentation, but not much.

Another sponsor, **Intel, also discussed "more computing for less energy**". Intel are making a serious effort beyond the Petascale, targeting the top 10 of the Top 500 processors. He started by noting work done under the openlab partnership and he showed some slides from Sverre showing our computing growth plan. Intel hope to partially address such plans by increasing computing energy efficiency (denser packaging, more cores, more parallelism) because they realize that power is constraining growth in every part of computing. Higher and more energy-efficient memory bandwidth is also needed which probably means new memory technologies, including the possibility of advances in solid state drives.

The third sponsor talk was from **Sun on Datacentre evolution**. The challenge is to cool and power a large datacenter where the load is heterogeneous. He listed some recent newly-commissioned Sun datacenters which are considered state of the art and which are attracting lots of customer visits. Another issue is to build "green" centres and he quoted solar farms in Abu Dhabi and a scheme to use free ocean cooling for floating ship-based computing centres. He claims that raised floors are dead - use containers or a similar but container-less pod architecture which has built-in cooling to handle up to 25Kw per rack. Power is also modular, connecting to an overhead hot-pluggable busway. For a $250M investment, Sun has reduced its global datacenter space by 41% and operating expenses by 30%. Their philosophy is to share and they are documenting their experiences at sun.com/blueprints and they are co-founders of Datacenter Pulse (datacenterpulse.org), a community sharing large datacenter experience. Going beyond this, they are testing 40-50-60Kw racks based on refrigerated cold plate technology.

GEANT and Optical Networking was presented by Hans Doebbeling from DANTE. He explained the history of DANTE and GEANT as well as giving a brief tutorial in optical networking. He presented the various services offered by GEANT and how the €40M cost is shared by the EU and the subscribing networks; most income is from the

GEANT IP users who therefore subsidise the GEANT circuit offerings. He described the part played by DANTE in the LHC-OPN network, including how they are installing monitoring based on the perfSONAR tool, with 2 measurement devices per site. Turning to trends, he noted that 40Gb optical transmission is already available and 100Gb is in the labs but he does not expect them to be available for use for some time. He ran out of time just as he got to the GN3 proposal.

Grid Security and Identity Management was addressed by the OSG Security Officer, Mine Altunay, focusing mainly on Id management. Although she said she would not do so, she first spent 5 minutes describing how Shibboleth works. Does this belong in a plenary? Anyway, having got through that, she then expanded the scope to grids and how Shibboleth can be made to work with grid certificates and how OSG and EGEE are experimenting with this in a limited way with federation-based CAs. She thinks diversity in Id management will continue and we must cope with this because interoperability is required. Her unsurprising conclusion is that operational security is hard to teach and to execute.

Computing for the 4 RHIC experiments: two of these, PHENIX and STAR, have data rates comparable to the LHC experiments. The BNL Tier 0 centre is RACF, the subject of many HEPiX talks. The speaker spent a little time on how these two experiments handled distributed discs and he warned that home-built cheap solutions were not always the cheapest in the long run and that moving data is far from a simple problem. Only STAR uses grids he has doubts on grids for production running (complex, too dynamic, hard to troubleshoot, little opportunistic use) so STAR has investigated clouds, Amazon EC2 again, for Monte Carlo production. His impression is very positive, although, like the Belle speaker the previous day, he confirmed that networking is a problem. But the virtual machine aspect of the image they create to run on the cloud, plus the truly opportunistic nature of running jobs is very reassuring. So far they have not attempted to negotiate a contract with Amazon but in summary, he believes clouds will change the game.

Software development tools to improve core performance by Paolo Calafiura, an ATHENA developer: to the question is optimization needed, he noted that LHC reconstruction codes, especially in ATLAS are pushing against the 2GB memory wall and he described them as “hungrier than Vista”. The first optimization step is to integrate performance monitoring into the code to measure not only CPU usage but also lots of memory metrics. The next step is to extract useful info from the mass of data produced by most performance management tools and some experiments have decided to store this data in a database. He believes the analysis of this data is best done by physicists not computer people, only they are trained to spot significant events he claims. There are three areas which could be optimized, CPU, memory and I/O and although different applications (MC, analysis, reconstruction) may benefit more from the optimisation of a particular one of these, they cannot be optimised in isolation. For CPU, gaining 10% is usually fairly easy but going beyond this is usually much harder. But he found optimising memory much more difficult; knowing how Linux memory management works and measuring the correct metrics are very important. Although they could be considered bugs, discovering memory leaks is a good place to start and he described the most common detection tool for this – valgrind – and then continued to give a high-level tutorial on memory optimization.

The final talk of Wednesday morning was given by **Dirk Duellmann on Distributed Data and Meta Data Management**. Databases are needed to provide consistent, concurrent, shared access to data in a consistent state, independent of client or storage problems. Thus databases play a key role in LCG. Dirk highlighted some current open issues and future work in the pure database area. For file management, more and more tools appear as use cases arise since there has never been an upfront design. Perhaps it’s time to consolidate these and he gave some examples in tape and disc file management. In general the impact of file management on analysis is largely

unknown as is the appropriateness of the model of Storage Elements; Dirk recommends that some experienced users should be working on recipes for analysis jobs in preparation for the large number of new users who will require to run such jobs when data analysis begins. Another area where there is a profusion of choice is file access protocols, ranging from RFIO and xroot to Lustre and other commercial or semi-commercial file systems. Other file system issues include security and global name space. Dirk continued with some words on using tapes more efficiently and closed with a plea that when LHC starts up and some failures inevitably occur, we must all work together to share the risks and maintain cooperating collaborations. In summary, he recommended more effort on physics analysis use cases, risk analysis and be prepared to integrate new technologies as they occur.

Vicenzo Innocenti opened Thursday by posing the challenge to adapt physics applications for many-core CPUs. Chip power has steadily increased following Moore's Law but is now meeting three limitations – memory access speed, power consumption and architecture difficulties, for example how to make effective use of longer and fatter instruction pipelines. Hence the recourse to multi-cores and the need to increase parallelism, something HEP codes have not historically been good at, despite the implicit parallelism of analysing independent events. We can schedule as many jobs as cores but we hit memory and chip cache contention. The first step is code optimisation, as presented in an earlier plenary. Beyond that Vicenzo is leading an R&D project in the context of openlab. We need to allow better sharing of common data by different processes but experience does not show this to be easy with existing legacy code. There have been some successes, for example GEANT, but 10,000 lines of code had to be modified. There has been more success adapting code for simple multi-process running and he showed some examples. One of these is a move towards a threads-parallel GEANT. However, ultimately, we need to devise parallel algorithms and work has started on a parallel MINUIT. He believes the outlook is positive and that future hardware and software technology may indeed help us.

Next was **Pere Mato on tools for distributed analysis**. He started by stating that he expects physicists not to want to distribute their analysis work but the scale of the data may force some degree of distribution so there is a need for some special common middleware for analysis jobs which should be run on the grid. Examples of such middleware exist but for experiment specific cases – DIRAC, PanDA, AliEn – plus user-friendly interfaces to hide complexity – Crap and Ganga (which is in fact shared by 2 experiments). This is not optimal for small VOs but most of the talk was evidence that the large LHC experiments are working on experiment-specific tools with only a little sharing. In short, the grid has had little effect on how analysis is done, each LHC collaboration has developed its own tools to handle the large amounts of distributed data and, although the architectures of these tools are similar, they are not common.

Summaries

The wLCG workshop which preceded the conference was summarised by Harry Renshall. There was a good mix of Tier 0, T1 and T2 representatives in the total of the 228 people present. It started with a review of each experiment's plans, all of which include more stress testing in some form or other over the summer, some wanting this to be in parallel by all the experiments. EGI to EGEE transition is clearly an issue (e.g. support for GGUS and ROCs and the timing of the transition) as is the lack of a winter shutdown in the LHC plans. Looking at site reports, it must be accepted that Tier 1 sites will go down sometimes and the experiments must be able to cope with this. As a result of some database service incidents, it has been decided to create a wLCG ORACLE Operational Review meeting, initially quarterly. There were reports from earlier dCache and CASTOR workshops where plans were

made to improve the reliability and efficiency of both. The User Analysis working group reported a large number of issues to be investigated and there was much discussion and lots and lots of recipes on how to report, work-around or respond to operational problems. Finally there was a discussion on whether there should be a CCRC'09 or not. The eventual decision was yes, but to rename it STEP'09 (Scale Testing for the Experimental Programme), schedule it for May or June and concentrate on tape recall and event processing. In summary, ongoing emphasis is put on stability, preparing for a 44 week run, and continuing the good work now started on data analysis.

Collaborative Tools track summary: 7 oral and 9 poster presentations. First comment is that their assigned room was full, more than 45 people at times. There was a lot of discussion, for example on Indico which is now used by 40+ institutes and stores material for over 60,000 conferences. Other subjects covered included the CMS Control Centre, lecture archiving, the Dirac interface to make grid access more user friendly and EVO. More details on these talks can be found later in this report.

Hardware and Computing Fabrics track summary: 17 oral and 24 poster presentations, audiences over 100 on first day (many more than the room capacity), 60-80 on the second day. Most of the talks are summarised later in this report but some highlights include:-

- FNAL have investigated Lustre and are pretty happy with it in most areas, but it is not suitable today in large-scale tape I/O applications.
- The ALICE online data storage team feel they are ready for LHC startup
- The Tier 0 will move to Linux SLC5 well ahead of data taking
- DESY is now the home of a German National Analysis Facility
- Solid State Discs are coming, mentioned in several talks.

S**oftware Components, Tools and Databases summary**: 35 talks, some in parallel, and 72 posters covering a heterogeneous range of subjects. Databases seem to have converged to Oracle by design and SQlite for distribution. Several talks on Python which seems to be used by everyone everywhere; this seems to have been a community movement rather than a top-down edict. The various experiments presented their respective frameworks. Of course root was discussed several times although Boot, discussed at the 2 previous CHEPs, appears to be dead. Other topics covered include monitoring and development environments but perhaps the most exciting topic this week has been virtualisation; there were some interesting technical talks in this stream but it was also mentioned by a number of plenary speakers. Prompted by Tony Cass and Michel Jouvin, the next HEPiX meeting, in Umea, Sweden in May, will attempt to explore this subject in more detail. Most of the talks in this stream are described in more detail later in this report.

Grid Middleware summary: 44 oral and 76 poster presentations, so many that they needed 2 parallel tracks on Thursday. They started on operations talks from all the major grids, all of whom state that the building phase has ended and they are concentrating on operational aspects such as monitoring, stability, etc. The stream convener has found Ian Bird's LCG middleware talk quite negative, lots of problem areas with no apparent solutions at the moment, and he eagerly awaits the full paper. ITIL only came up once, from the GridKa people. There were various talks from experiments using the grid, not only LHC experiments, on their experience. Security was also covered, as was software distribution, monitoring, job management and data management. In this last area, reviews of both dCache and CASTOR were presented and both were declared by their author as being ready for production although the former had areas which could be yet improved. The summary is that production grids are there, grid middleware is usable and is used (despite the previous negative comment on Ian's talk), standards are evolving

but have a long way to go and network bandwidth use seems to keep pace with technology [Ed: or it the other way round?]. Most of the talks in this stream are described in more detail below.

Distributed Processing and Analysis summary: 35 oral and 63 poster presentations. The talks were split into those on general tools and others more specific to one or sometimes more than one experiment, but the speaker went so quickly through her summary that she gave not much more than the talk titles and speakers. Luckily most of these talks are described in more detail towards the end of this report. Her summary is that clearly a lot of work has been done on user analysis tools since the last CHEP and there are some commonalities between the LHC experiments. Data management and access protocols for analysis are a major concern and the storage fabric will be stressed when LHC running starts. She noted in passing that BaBar said that this may be the last CHEP where they present anything.

Conference Summary (Dario Barberis): he noted some 100 ATLAS attendees, about 50% of those currently active in ATLAS computing. The most common word in the 500 abstracts was “data”, sometimes linked with “access”, “management” or ”analysis”. He noted that users want simple access to data so we need to provide easy-to-use tools to hide the complexity of the grid. Of course “grid” is another of the most common words in the abstracts. The word “cloud” did not appear in the top 100 abstract keywords but clouds were much discussed in plenary and parallel talks. And for Dario, the last major theme was “performance”, at all levels from individual software codes to global grid performance. Dario felt that networking is a neglected but important topic (for example the famous digital divide and end-to-end access times); another is collaborative tools (although more work is starting hee), His conclusion is that performance will be a major area of work in the future and a major topic at the next CHEP in 18 months time in Taipei, October 17 to 22, 2010.

Grid Middleware

Jeremy Coles described the status of the UK particle physics grid, GridPP. He showed quite positive graphs of Tier 2 reliability, a significant improvement from the past, although there is some degree of underuse of installed resources. The new Tier 1 centre at RAL is nearly ready (problems with distribution of cold air I heard). CASTOR performance has improved after a long period of instability and a migration plan to version 2.1.8 is being planned (important for analysis). ATLAS has been running analysis tests everywhere, including the UK and this was found to be very helpful and crucial to spot problems especially at the level of storage and network. In the last months, some problems coming from resource contention (initiated by activities of non-HEP VOs) have been observed). GRIDPP sites find it difficult to identify when experiments’ production systems blacklist a site and a Nagios alarm would be very helpful. There has been more emphasis recently in improving general operations and service management of the grid infrastructure, including a review of disaster recovery and service resilience, and some effort has been invested in experiment-driven issues.

The next talk was a **presentation of ITIL use at GridKa**. The computer centre has implemented ITIL V2 service support processes covering configuration management, incident management and helpdesk, and problem and change management. The Configuration Items, related to service components, are stored on a central database; it is duty of a Configuration Manager to make sure this happens. Changes are handled by a specific “officer”, the Change Manager. They have a nice idea of a change calendar, a bit more organized/fancy than CERN/IT’s Service Status Board. There are shift procedures (also relying on personnel on call) dealing with the incident management,

while dedicated task forces take care of the problem management (understanding of the reasons which caused the incident). Generally, incidents can be identified externally (SAM test, GGUS tickets) or internally (Nagios alarms). Those two workflows interoperate. Each intervention or change request has an owner or person responsible. In answer to a question on how long it took to implement and install the new processes, he said it took at least 6 months to just get everybody to know ITIL. They met some resistance especially from the technical staff while implementing ITIL and it is an ongoing process and some people find it hard to adapt to a new style of working – too much paperwork for example.

Laura Perini presented the business case for EGI. She listed the main actors, National Grid Initiatives covering Operations, Middleware Consortia for maintenance and R&D, EGI.org (coordination) and Users, stating that CERN was implicitly treated as an NGI. She listed the various functions of EGI and how they will be funded, what EGI will offer in terms of secured shared resources, a unified middleware distribution, etc. and what added value EGI will bring to its users, to the resource providers and to the funding agencies. The relationship with the customers is based on a NGI to User relationship. The NGI to Resource Provider relationship is handled via SLAs. She showed how the different tasks will be shared between the NGIs and EGI. Finally she showed the various bodies of the EGI.

Continuing the tour of grids, **Ruth Pordes presented an OSG update**. She presented statistics showing a significant increase in installed resources and in use, and various technology improvements such as growing use of different techniques for job scheduling (overlay frameworks, pilot jobs) and opportunistic use of storage resources by D0. Although there are risks in doing so, OSG has decided to write down areas of value of and benefits from OSG. There were 5 areas covered :-

- collaborative research support
- sustainable US cyber infrastructure for researchers
- contributor to computer science
- building US expertise
- promoting opportunistic computing.

Ruth listed some of the analysis of these areas, including measures of benefit in many cases. In preparation for LHC production, OSG is concentrating on establishing and promoting US Tier 3 sites, currently numbering 20 but targeted to rise to 70 by next year. She listed some future plans such as creating a structure for short-lived VOs, improving the software stack and so on. Turning to clouds, OSG sees clouds as another interface to deliver well-defined storage and processing and demonstrations exist linking Condor pools to the Amazon cloud and running STAR applications in clouds via virtualisation.

CDF way to Grid - Donatella Lucchesi (University and INFN Padova): The CDF computing model was designed when Grid was only an idea and therefore is based on having dedicated resources. There is one major dedicated farm (CAF) at Fermilab and distributed CAFs all over the world. A CAF head node is the interface between the User and the Batch System. Tools for job submission in batch mode and monitoring (CLI and web-based) are available. Users can get email notification about job completion, together with a summary. Condor has been adopted for the farms. The CAF head node includes CDF specific daemons (monitoring, submission and mailing) and the Condor Negotiator and Collector. The monitoring is interactive, without dedicated access of the user to the worker node. At the end of 2003, CDF started its transition to Grid to benefit from more resources. Grid resources are accessed via Grid Access Points (5 worldwide). The initial requirement in moving to the Grid environment was to shield the user from this move, so most of the tools/interfaces were kept (Kerberos authentication, submission clients). Data Analysis happens in a few dedicated centres (again, to minimize impact on users), while Monte Carlo productions

runs worldwide. For OSG, the CAF code has been modified and extended to the Condor Glide-In mechanism to uniform the access to all resources at different sites, keeping the condor submission mechanism. CDF in the last year used 23% of resources in OSG. The best performance achieved was in the order of 5K jobs running concurrently, at which point they met a scalability problem, so that a new system (the glideinWMS) had to be developed. The new system achieved the rate of 50K concurrent jobs. The code of the CAF was also adopted for submission into glite (LCGCAF). In this case the GlideinWMS in the CAF Portal is substituted by the gliteWMS. In Italy, CDF used in the last year 10% of the available Grid resources. For the CDF code distribution, CDF relies among other things on the Parrot mechanism. In conclusion, CDF adapted very successfully to the Grid system "on-the-fly", in a completely transparent way for the users.

GOCDB, A Topology Repository For A Worldwide Grid Infrastructure - Gilles Mathieu (RAL): GOCDB stores information about Sites, Regions, Countries and Users (Operations, Site Managers , etc …, but not end users). It consists of a backend database and a web portal. Data handled by GOCDB consists of administrative information (contact names for example), Resources and Services and maintenance plans. Historically, GOCDB started as a static list of sites and contacts. It evolved with the introduction of a MySQL database and access scripts first, and a web service later. Subsequently the database backend has been migrated to Oracle running on an 11G cluster, with three schema: production, test and development. The web interface, coded in PHP, includes X509 authentication (not all information is available to everyone). The last improvements in GODDB includes a programmatic interface (based on Oracle XML DB), while a SOAP web service for the frontend is being considered. As a central tool, GOCDB could be a single point of failure, so it must be highly reliable: there are backup databases in the UK and in Italy. There is also a replica of the web portal of the web server in Germany. In case of DB problems, the UK backup can be used, alternatively, in case of major problem in the UK (e.g. network) the IT+DE instance can be used. Switchover is fully automated. The future evolution is focused on decentralization, based on regional instances and lookup methods (logical view) from the central instance to regional instances (when they exist). This follows naturally the distributed EGEE/EGI model.

DISCUSSION: regarding the scalability of regional databases, the point is not how many instances there are of regional databases but which information will be provided.

Bringing the CMS Distributed Computing System into Scalable Operations - Jose Hernandez (CIEMAT): the CMS model is based on multi-tier organization. The definition of tiers is based on the function of the site in the computing model. CMS has undertaken in the last five years a variety of computing challenges, for both workload management and data management. 18 months ago a task force for site commissioning (data transfers, site services, workload management) was set up and this has become a permanent activity. CMS needs to transfer data from T0 and T1s, among T1s, and between T1s and T2s, and, given the importance of such activity, a dedicated effort within the task force has been established. A load generator continuously injects load on the system to provide the required amount of traffic to test all the links. The availability monitors show a clear improvement in data transfer activities. In terms of site monitoring, a SAM-based framework has been developed to measure readiness of sites. The site availability takes into account specific SE and CE tests, as well as results for CMS-specific services at the site. Besides the SAM test, a Job Load Generator (JoBRobot) is in place to generate load on the workload management system. Strict metrics are defined for site readiness: they include the number of active links from/to the site and quality of commissioned links. All results from site commissioning are aggregated in the Site Status Board. This combines all metrics into a single daily readiness status. For the commissioning of production and analysis various scaling tests for Production and Analysis agents have been carried out. Since the number of Production Agents and Analysis Servers can be increased without limitation, there is no scalability issue foreseen for those activities. CMS in fact handles routinely 70K jobs per day and more

than 100K jobs/day have been run during challenges. The overall job efficiency for production is approximately 80%, while for User Analysis this reduces to only 60%, where the biggest problem is the remote stage-out of files. Pilot based mechanisms show higher reliability since they reduce grid failures. In conclusion, commissioning exercises are crucial for computing activities. Data and Job management seem to scale well.

A Dynamic System for ATLAS Software Installation on OSG Grid site - Xin Zhao (BNL): The ATLAS production system relies on experiment software being pre-installed on grid sites. The previous ATLAS installation system from OSG was based on gridftp to transfer install scripts and globus-job-run to run the installation. The main problem of this approach is that the job is run in the gatekeeper, which is now identical to the WN environment (and in addition this causes load on the gatekeeper). The new installation system is based on DQ2 (ATLAS data management) to deliver the software Pacman Tarball to the Storage Element of the site and on Panda (the ATLAS production system) to perform the installation. The Panda system is pilot based, and the installation process is just another payload for the Panda server, like an analysis and production job. In addition, the installation job can be given priority over the other jobs at the same site. The installable module is a Pacman image wrapped in a script: it is self-contained, self-installing and non-updatable. The installation script downloads the pacball from the site SE and runs the real installation. The installation scripts are the same as the ones used in EGEE and, in addition, since Panda and DQ2 are used in ATLAS worldwide, this mechanism can interoperate between OSG and EGEE. Panda has now a new type of pilot (the installation pilot) which is continuously run at sites (just 1 job per site) under dedicated Software Group Manager credentials. The installation jobs can be monitored via the Panda monitoring system but also via a curl (http) CLI. For the future, the system will need further automation. One possibility is to develop a new job payload script scanning for missing releases and installing what is missing. One of the problems is the latency of distribution of the pacball to the T1s: a new release process (and possibly distribution technique) will be put in place. Lastly, further modifications are needed to better support opportunistic T3s.

Migration of ATLAS PanDA to CERN - Graeme Stewart (University of Glasgow): Panda, the ATLAS production and Analysis system, is a central ATLAS component and it was decided recently to have it hosted at the CERN computer centre, so a migration from the previous existing instance in BNL was needed. The Panda service at BNL runs against MySQL databases for various components (task request interface, Panda Server, Monitoring), while at CERN this was going to be accommodated in Oracle. The first part of the migration was the monitoring interface, which was not problematic since the system was already architected to have multiple monitoring instances. The migration of the task definition was more problematic since a full database migration was needed. In addition, while the old database was running on MySQL, at CERN it is now run on Oracle. The migration of the Panda server was the real challenge since archive information also had to be migrated. One good point is that Oracle allows for database partitioning which could not be done in MySQL and the ArchiveDB is quite large. The migration was done gradually (no big bang) on a cloud-by-cloud basis, to minimize downtime. The production instance at CERN was installed on a integration database (was not considered tested enough for the production database) but with backup in place (not usually the case for the integration DB). Quite a lot of tuning and tweaking has been done in the Panda server to optimize the access to Oracle rather than MySQL. In general, despite the successful migration so far (not all clouds have been migrated), there is still quite some work to be done to improve the stability of monitoring and support procedures.

Critical services in the LHC computing - Andrea Sciaba' (CERN): All LHC experiments run on the wLCG infrastructure. wLCG enforces a certain level of reliability and availability, which periodically must be evaluated. wLCG defined the readiness taking into account software readiness, service readiness and site readiness. Service reliability is not included but rather measured *a posteriori*. Before CCRC'08 every experiment provided a list of critical services, rated from 1 to 10, a rank of 10 being tagged as "critical" so that the maximum downtime allowed

is 2 hours. ALICE and CMS defined a quite fine-grained criticality of services, while other VOs, including ATLAS defined only few levels of criticality (high, medium, low) depending on the impact on data taking and other main activities. Almost all CERN central services are fully ready, but there are a few concerns about WMS monitoring and several services are only "best effort" outside working hours. The documentation of some Grid services is not fully satisfactory. ALICE looks ready; ATLAS is ok but relies on experts for problem solving and configuration; the analysis impact is largely unknown. CMS services are basically ready but for the Tier0 services (data processing) the development/deployment process is in continuous evolution. LHCb has shortcomings in procedures and documentation but they note that DIRAC3 is still relatively new. For the support at T0 and T1s, there is general 24/7 support and the VOBOXs are covered by the SLA. Again, reliability is not included here, so several critical services which are considered ready could in fact be "fragile". Nevertheless, no key showstoppers have been identified and experiments depend on reliable and mature services from the point of view of documentation, deployment and support.

Where is the Internet heading to? - Olivier Martin (Ictconsulting[2]): The Internet consists of two main branches: the academic & research Internet, which is bandwidth-rich, and the commercial Internet. The latter is affected by a series of issues which could undermine its existence in the next years, particularly the fact that IPv4 address space exhaustion is predicted to occur within the next 2 years.[3] There are also issues in routing, security, inter-domain Quality of Service. IPv6 looks "almost" unavoidable but is by no means "guaranteed" to happen. On the other side "clean-slate" solutions are unlikely to be viable before 7-15 years (need a gradual step-wise evolution). The instability of the Internet routing system is preoccupying as well as the increasing lack of "network neutrality", copyright infringements, etc.

SVOPME: A Scalable Virtual Organization Privileges Management Environment - Gabriele Garzoglio (FNAL): The SVOPME system is built to ensure uniform access to resources, providing an infrastructure to propagate, verify and enforce VO policies at Grid sites. The last decision about which policies to support is always in the hands of the site. A VO Policy Editor (integrated with VOMS) will allow the VO Admin to specify policies, starting from templates (with possible extensions via plug-ins). A policy advisor compares VO policies with site policies and informs the site admin about the VO requests. Similarly a VO/Grid policy comparer helps the VO admin to understand what the VO can do at the site, given the site policy.

DISCUSSION: it was clarified that in SVOPME it is not necessary for the site to advertise the policies (in the gLite authorization service this is achieved via private and public policies). Also, there is an effort between OSG and EGEE to interoperate the gLite authz and SVOPME, at least at the level of XACML and protocol.

VOMRS/VOMS Utilization Patterns And Convergence Plan - Tanya Levshina (FNAL); Andrea Ceccanti (CNAF): VOMS is the source for Grid authorization, based on extensions of X509 proxies. Currently two tools exist for VO registration: VOMS Admin, not fully compliant with the Joint Security Policy Group requirements; and VOMRS (a web service extension of VOMS Admin) providing a more sophisticate registration workflow, which is compliant with JSPG. The intention is to provide a unique administrative interface which implies implementing in VOMS Admin missing features existing in VOMRS. The migration process consists in: implementing JSPG compliance for VOMS Admin, the migration of essential VOMRS features to VOMS Admin, implementation of an interface with a third party directory service (such the CERN Human Resources Database) and the validation and testing of the all above (including client compatibility).

[2] and ex-CERN CS group

[3] Excuse the editor's sceptism but where have I heard that before?

The new gLite Authorization System - John White (HIP); Andrea Ceccanti (CNAF): AuthZ was designed to provide policy decision making to VOs and Sites, in an uniform way. The service consists of four components: Administration Point (to formulate rules), the Decision Point (to evaluate the request), the Enforcement Point (where the enforcement of the decision takes place) and the Runtime Execution Environment Point (to define under which environment the application should run). AuthZ should enter certification in April 2009 while the deployment will take place in self-contained steps: support for glexec on Worker Nodes, implementation of Banning Lists (both global and local to sites, with the first taking precedence), integration with the CREAM CE and the gLite WMS (in parallel), support in the matchmaking. Integration of AuthZ with the Data Management systems should be considered.

DISCUSSION: the AuthZ is a replacement/evolution of SCAS, but it was clarified that SCAS should be rolled out immediately since it will be used in the next 40 weeks or so, before AuthZ is fully deployed. DISCUSSION: since data management is not included initially in the picture, the banning of users can not be enforced at the global level

On the role of integrated distributions in grid computing - Oliver Keeble (CERN): The middleware deployment currently relies on integrated distributions, but some of the advantages of this approach are no longer relevant (e.g. coexistence of services and clients), due to virtualization and multi-core architectures. Currently the release timetable is dictated by the slowest component, but going to independent distributions (specialized for each service), would allow to relax this constraint. On the other side, going to independent distributions would need multiple versions of libraries on the infrastructure (but this is already the case) and assurance of protocol compatibility. One step forward would come from virtualization, especially in the area of client distributions, where VOs would define OS+MW+application boundless.

DISCUSSION: the VM approach opens security concerns about the operative system running in the virtual machine.

CDF software distribution on Grid using Parrot - Simone Pagan Griso (University and INFN Padova): In CDF it is complicated to pre-place all software packages on Worker Nodes (distribution size, upgrade frequency etc). The adopted solution consists in extending HTTP-based mechanisms making it act like a file system (showing files in a Code Server), emulating structured directory listing. At this point Parrot can be used to attach this file system to the job. Parrot is an interposition agent, intercepting the I/O calls and therefore can virtually mount a remote file system as local. Parrot runs as client on the WN (wrapping the job), resolves files using directory listing and downloads the file using Squid caches. In terms of performance, Parrot introduces some overhead in syscall latency, in the order of 1% on a 12h job.

Grid Middleware for wLCG - where are we now, and where do we go from here? - Ian Bird (CERN): The role of WLCG is to provide computing resources for LHC experiments (not creating a Grid). The key point is simplification for more reliability. Grid development did not fully delivered the requirements but in the other side these were probably overstated, In 2005 it was decided to focus on baseline services rather than a full all-inclusive solution (streamlining of the middleware stack, focusing on key components). There are multiple areas where no issues are observed: single sign on, data T\transfers, a simple file catalogues (LFC), the network, databases, batch systems (including the old fashion LCG CE, which still scales to today' s needs). In terms of workload management, pilot jobs removed most of the complexities and make centralized matchmaking not necessary any longer. Data management is complex and SRM lacks uniformity. In general, complex functionalities are always very application-specific. For the future, wLCG could become a cloud-like object where details are hidden by virtualization. At the site level, there is still a need for a CE, but this could be something very simple, just capable to interact with the

batch system and launch a pilot factory. This would also simplify the information system where fully updated information would not be needed any longer. In terms of data management, the overall picture needs to be simplified. gridFTP is probably still needed, but the management interface (SRM) is too complex and having a single access protocol (xrootd?) could be envisaged. On the other side, mountable file systems now look more scalable. Web services did not really deliver what was promised (GSOAP did not really mature): in the rest of the computing world messaging systems are used to decouple distributed services. All this should be seen as a long term plan (no big changes in the near future) since we have a working system, but the risk is the maintainability.

DISCUSSION: an interesting idea would be trying to apply the pull model not only to jobs but also to data.

DISCUSSION: about the storage, for T2s, mounted file systems with a simple interface (like StoRM) seem the future. For T1s, usage of dCache and CASTOR is complicated by SRM and this is what should possibly be simplified.

Modern methods of application code distributions on the Grid" Pablo Saiz (CERN): ALICE built a software installation system suitable for multiple platforms. This system is based on small, self-contained packages where grid software updates can be triggered by a VO admin and the installation can be automated (on demand) on every worker node. The system does not rely on any shared area: a job lands on a virgin worker node and triggers the installation of the software in a scratch area. The software is then transferred using the torrent technology. Such technology has already been tested by millions of users, reduces inter-site transfers and does not require any peculiarity on the worker node. The tests at CERN have shown no scalability issues for 600 concurrent jobs downloading 180GB of software, but an obvious advantage would be to re-use the same software on the worker node after job completion. The security impact has been fully evaluated by security experts at CERN and no risk has been identified in the technology. The same algorithm could be used by other VOs.

DISCUSSION: the P2P technology requires port opening within the site, which many sites do not like for the proxy renewal mechanism. Open point.

DISCUSSION: in multi-core machines, the software is downloaded once per core, not per node.

PhEDEx Data Service - Ricky Egeland (Minnesota): PhEDEx is the CMS Data Management system. The PhEDEx database contains 4.3M files and information about dataset location and file transfers. Datasets are organized in a hierarchical model, i.e. divided into units of replication (data blocks). The Data Services communication is based on HTTP and is not read-only (you can inject, for instance, subscription requests). Data Services relies on a security module for authentication/authorization. The authentication is done at the web front-end, while the authorization relies on information in a policy database (SiteDB). There is a CLI available for interaction with Data Services. The Data Location Service in CMS is the main customer (Grid analysis) but similar amounts of requests come from the Data Discovery system and the Tier0 workflow.

Data Management Evolution and Strategy at CERN - Alberto Pace (CERN): The monitoring of CASTOR improved with the addition of several new indicators and key performance is now calculated in real time (good for alarming). In terms of security, there is now strong authentication: there are no problems identified for xroot access, but there are limitations using RFIO for both Kerberos and X509 and any further improvement would mean massive manpower investment. The SRM CASTOR interface reached the required robustness and there are foreseen improvements (more logging, protection against overload, better integration between SRM and stager DBs). Tape access now benefits from new queue management (supporting policies, ACLs and protections) and current work focuses on a new tape format supporting data aggregation (better support for small files). The CASTOR file access latency has never been a problem for the T0 operation, but is a concern for the analysis access:

therefore LSF has been removed from xroot client read access (giving 2 orders of magnitude improvement) and it is planned to do the same for write operations. As a conclusion, the usage of the xroot protocol becomes of strategic benefit in terms of various improvements in performance (this could be extended to mounted file system support via FUSE).

Data Management in EGEE - Ákos Frohner (CERN): The EGEE Data Management software stack consists of Client Tools, High Level Services and the Storage Elements. The DiskPoolManager, managing disk-only storage, supports many access protocols, among which xroot, but only for the Alice use case. The next release of DPM will include support for checksum and srmcopy. Before LHC startup, support for generic xroot is foreseen. For the LCG File Catalog, future developments will focus on performance, providing bulk compound methods. The File Transfer Service will soon support channels between sites and clouds of sites. The main new feature of the currently-tested release consists of the split between the SRM negotiation and gridftp transfer phase. In the next release, support for checksums verification will be added. The GFAL and lcg_util clients (hiding data management complexities from users) currently offer more detailed timeouts and will support checksums in the next release, together with more protection against SRM overload. Hydra is the offer for an Encrypted Storage solution. In conclusion, the priorities for the future directions are stability, reliability and maintainability, together with providing more administrative tools and better integration.

DISCUSSION: future developments would not focus on providing a common RFIO library for CASTOR and DPM since the uniform direction should be xrootd. The xrootd protocol for DPM is not more performant than RFIO, but will be there for compatibility.

dCache ready for LHC production and analysis - Patrick Fuhrmann (DESY): dCache.org is independent from LCG and OSG for structure and founding. Various components are provided by different labs, but there is a virtual central place where dCache is built as a single product. dCache is a single component talking different protocols: SRM v1 and v2, gridftp, dcap and xrootd. dCache offers support for tap access. It has the capability to move data internally for optimizing data access and data transfers transparently from the users. It offers advanced administrative functionalities for pool-draining and system partitioning (even on the WAN). The SRM, supposed to solve the problem of globalization, was a partial success: it was supposed to serve requests as fast as possible and protect the backend but this goal has not been achieved and, in addition, the abstraction layer is too complex. Effort was therefore invested in making the backend faster: one example is the implementation of Chimera for the namespace instead of PNFS. In addition, several improvements have been implemented at the SRM per se (such as asynchronous srmls). In terms of standardization, work is being done in providing SE namespace dumps and supporting extensions of Glue. For the future, support is foreseen for NFS4.1 (adopted by major storage providers). About analysis, the situation is not clear: there is no evidence of problems, but metrics are not clear.

On StoRM performance and scalability - Luca Magnoni (CNAF): StoRM offers SRM service for disk-based storage systems. It has a multi-layer architecture: a front-end exposes the web server interface, while the backend executes all SRM requests. The front-end and backend communicate via XML-RPC for synchronous requests. For asynchronous requests they are stored in a central database. A performance analysis has been carried out to understand the limits of critical components and the response under heavy load (not throughput tests). The test took into account the different nature of synchronous and asynchronous SRM requests. The first test consisted of submitting SRM requests via both SRM clients and direct interaction with StoRM DB and tuning front-end(s) and backend configurations to check how the system reacts. The test has shown that the test layout is able to sustain a load of around 15 Hz of FTS-lite; a single backend is able to manage around 1800 srmPreparetoPut/minute (this is

the most expensive request); processing time at the server side is a small fraction of client side response time; the front-end replication works as expected; GSI authentication on each request is really CPU intensive.

Will Clouds Replace Grids? Can Clouds Replace Grids? - Jamie Shiers (CERN): We should build a set of criteria which cloud computing should satisfy to be considered for LHC computing. The key point of Grid computing which must be demonstrated in clouds is whether a minimal level of service can be achieved. In wLCG we have metrics: number of tickets submitted to sites, number of scheduled and unscheduled interventions, site availability as measured by experiments. In addition, the scalability and reliability of wLCG services must be tested at the "petascale" for CPUs, data management and databases. In all cases, it has been shown that you need to have very detailed control down to the lowest levels to get the required performance and scalability. Moreover, data volumes, rates and access patterns representative of LHC data acquisition, processing and analysis must be considered. Obviously, the cost must be entered in the equation. How can this be achieved through today's Cloud interfaces is the question. So we cannot afford to ignore major trends in the computing industry and we have to evaluate cloud computing based on well-proven mechanisms for determining whether the computing service(s) satisfies an agreed set of requirements. Not considering Cloud Computing for at least some HEP use cases would be a mistake.

The CREAM-CE: First experiences, results and requirements of the 4 LHC experiments - Patricia Mendez Lorenzo (CERN); The CREAM CE allows submission via the WMS but also offers a native CLI for direct job submission. In wLCG, sites have been encouraged to provide a CREAM CE in parallel to the LCG CE (even if there are some aspects which represent a show stopper for at least 2 LHC experiments using the CREAM CE). For a full migration to CREAM, the submission via Condor-G should be implemented, the ICE component in the WMS should be in place, a robust proxy renewal mechanism should be in place. Also, several scalability metrics have been provided. ALICE would like to deprecate the usage of WMS and use direct CREAM submission at all sites as soon as possible (no proxy renewal required). ALICE tested submission to CREAM in the summer of 2008 (55K jobs) and, as of today, is using CREAM at 12 sites. ATLAS would also be interested in direct submission but for the moment the main need is to have Condor-G being able to submit to CREAM. For this, there is a Condor prototype which CMS is testing. At the moment there is a 25% error rate coming mostly from Proxy Renewal. For LHCb, testing CREAM is not high in the priority list (glexec and porting to SLC5 come first). LHCb intends to submit jobs directly to CREAM (without WMS or Condor-G), but requires support for CEMON in the direct submission (at the moment only via ICE).

Use of the gLite-WMS in CMS for production and analysis - Giuseppe Codispoti (Dipartimento di Fisica): CMS built a job framework (BossLite) to shield users from the complexities of different services on various Grid flavours. The architecture is made of an interface part and a scheduling part. In the Boss interface to gLite, bulk submission, query and match-making are used via the WMProxy python API and the LB API. CMS uses its own data location system and therefore the match-making is done only for sites holding the selected data. BossLite covers the use cases of Monte Carlo production, basic analysis tasks and intensive (centralized) analysis tasks. The most interesting use case is the one of the CRAB analysis server (multi-threaded job submission and status query). Currently production and analysis are balanced over 11 WMSs, where a single WMS activity can reach 15K jobs/day (tests have shown no scalability issue at 30K jobs/day, sustained for several days). The WMS architecture in fact is such that many WMSs can be used in parallel and added to the system in case of scalability problems. The overall CMS activity using gLite WMSs achieves approximately 75K jobs/day, with a success rate which varies from 58% (analysis) to 87% (fake production). The Grid-specific failures are always 10% of the total number of jobs. The CMS experience has brought important feedback to gLite developers.

DISCUSSION: using a very light match-making (among few sites) simplifies the activity of the WMS and increases stability/reliability.

Using CREAM and CEMON for job submission and management in the gLite middleware - Massimo Sgaravatto (Padova): CREAM is a service for job management at computer centres, while CEMON is a general purpose event notification framework. CREAM provides functionalities for job submission (can be also disabled automatically), suspension and resume, purge, proxy renewal. Submission to CREAM can be achieved via the gLite WMS (via the ICE daemon running on the WMS) or directly via the CREAM interface. The interaction with the underlying batch system is done via the BLAH component, while the credential mapping is implemented via glexec (via LCAS, LCMAPS). CREAM has been released for production in October 2008, but the submission via ICE still has scalability issues. Criteria have been defined for replacing the LCG-CE with CREAM. Submission to CREAM has been tested in dedicated setups: the submission rate for CREAM has been measured as 100 jobs/minute (to be compared with 35 jobs/minute for the LCG CE). In terms of standardization, CREAM exposes a BES-compliant interface, while submission from Condor has been implemented and is currently being tested. For build, installation and configuration, it is fully synchronized with the gLite procedures. For the future, CREAM will support bulk submission, together with the possibility to deploy it as load balanced service.

CDF GlideinWMS usage in Grid computing of High Energy Physics - Marian Zvada (Fermilab): The CDF CAF allows users to develop, debug and submit jobs to local batch resources in a secure way and with a pseudo-interactive monitoring. The GLIDECAF was implemented in order to move the framework to GRID resources. GLIDECAF benefits from the pilot jobs technology for both submission and monitoring. CDF has been successfully using Grid resources through glide-ins for the past 4 years but the scalability limits of the home-grown software have been reached. At the same time, CMS has developed a more scalable glide-in solution, and this being general purpose, CDF could adopt it immediately. Thus CDF is migrating to glideinWMS and the experience up to now very positive.

DIRAC3 - the new generation of the LHCb grid software - Andrei Tsaregorodtsev (CNRS-IN2P3-CPPM): DIRAC is a distributed data production and analysis system for the LHCb experiment, initially focused on MC production and later extended to data analysis. In 2006 the project was reviewed. As an outcome of this, the base technologies were considered adequate and the general architecture (service based, modularly designed, pilot based WMS) was approved. Following the review recommendations, DIRAC3 now covers all needs in the LHCb distributed data processing: this includes all the data export and handling framework and automatic distribution of the analysis data. DIRAC3 is built on a new secure framework based on X509 and fine-grained authorization rules, and includes a fully-featured proxy management system for supporting multi-user pilot jobs. The DIRAC WMS applies VO policies (production and analysis jobs) handling job entries in the same task queue. In addition, the pilot job technology shields from the heterogeneity of different grid resource and infrastructures. In terms of performance, DIRAC has been stress-tested in the recent FEST'09 exercise, where up to 15K jobs have been running concurrently across 120 sites. An advanced failover system protects jobs from grid-specific failures. Several systems have been then built on top of the DIRAC system, including a job definition service, a data management system (for data distribution operations, based on a request management system). Finally, LHCb developed a new web portal to monitor DIRAC activities and for the accounting.

The ALICE Workload Management System: Status before the real data taking by Patricia Mendez Lorenzo (CERN): The ALICE WMS is based on a central Task Queue and Optimizer Agents (splitting tasks, sorting jobs in terms of priority, applying VO policies) and Site VO-BOXes (submitting jobs to the local computing element and taking care of software installation). Jobs are currently delivered to the site via the gLite WMS. What runs in the worker node is a Job Agent, pulling in the real job. The ALICE experience with the gLite WMS is quite negative: jobs

remained backlogged in internal queues (the reason is not really clear) and moreover there is external access to WMS internal status. Those zombie jobs have been a real problem for the ALICE workflow. The CREAM CE on the other hand allows for direct job submission and the first tests in FZK have shown very promising results. From the site point of view, the installation of CREAM is not straightforward, but there is good support from the CREAM team and, once installed, it shows good stability; the missing feature is the possibility to deploy the CE in load-balanced mode. In terms of performance, direct submission to CREAM is a factor of 2 faster than submission to the LCG CE via the gLite WMS. For all those reasons, ALICE will follow the direction of direct submission to CREAM and abandon the submission via gLite WMS.

Hardware and Computing Fabrics

Michele Michelotto presented the results of the HEPiX Benchmarking Working Group. It was noticed in 2005 that HEP codes were no longer linear with respect to SI2K and indeed which SI2K definition to use, there appeared to be several. Thus was the working group born, using resources provided by CERN and some tier 1 sites and using applications from the LHC experiments? After some tests, they agreed to use SPEC 2006 as the basis for future benchmarks because it was the one that showed the better correlation with real physics jobs, and they created a unit, HEP-SPEC06, with a conversion factor of 5 from SPECI2K. From Q&A: Using a custom SPEC measure which is not known outside the community could be a problem in terms of outreach, a possible solution is to report number of cores.

Sverre Jarp presented the Intel's ATOM processor and asked if it is ready for HEP. He started with a description of the processor and its performance measurements. He showed the raw benchmarks of an ATOM compared to a Harpertown system and deduced a factor of 16 in cost ratio (based on pre-release prices of ATOM) and 13 in throughput advantage, but the test setup was 5 times less performant in thermal production. (power per watt). So, until Intel increases the memory capacity, improves the thermal efficiency of the chipset and adds more cores, HEP is probably not that interested in ATOM (although the audience in that room almost doubled before the talk and almost halved afterwards).

Tony Cass presented the "reality" around air conditioning and computer centre efficiency. He listed various cooling options, illuminating each with examples from other sites. He also showed the current CERN scheme of enclosures with hot and cold isles and running servers somewhat higher than we used to. One problem is to control air speed to optimize cooling but not make the room inaccessible for maintenance. Highlighted the need for (a lot) more monitoring of the cooling in computer centres to enable the search for the optimal working points for each case. Next challenge is to reduce the cost of creating chilled water for the units, much easier if starting from an empty room. Chiller selection is important and should not be rushed. Ideally have a range of sizes of chillers and the combination which satisfied the current needs.

A High Performance Hierarchical Storage Management System For the Canadian Tier-1 Centre at TRIUMF: they wrote their own HSM system, "tapeguy". Implemented in Perl with MySQL DB back end, supports prioritization, reordering, file and tape grouping. I/O balancing on reading and writing. It provides an estimated wait time for all files. From Q&A: They are happy to share it.

Fair-share scheduling algorithm for a tertiary storage system: Very interesting talk, they plan to use fair share scheduling for tape reading. Pre-staging is done automatically, each job publishes the list of files to be read and a meta-scheduler does the pre-staging. Created an MC simulator of the tape system to study different scheduling solutions for tape reading, FIFO, WFQ (queues per user), WFSG (fair share looking at 3 parameters, number of files per tape, usage and size of the file). WFSG provided the best user experience (good throughput, maximal QoS and lowest delay).

Lustre File System Evaluation at FNAL: Evaluation of different file systems started by defining a list of weighted criteria for the evaluation of the different candidates. A very thorough study about how Lustre matches (or not) their requirements. From Q&A: FNAL stayed out of the HEPiX storage systems working group and are doing similar studies on their own. Would be interesting to see them join the working group and share the results.

The ALICE Online Data Storage System: Presented a detailed view of the HW and SW components of the system. Showed the importance of continuous testing and validation of all the components to achieve the desired stability and performance of the system.

Integration of Virtualized Worker Nodes into Batch Systems: The main motivation to use virtualization was to be able to share the resources between the different user groups and the grid while keeping the different environments isolated. VMs are deployed "on demand" depending on the environment requested by the job. One VM is deployed (or resumed) per job, the images for the different types of VMs are in each WN and are started (or halted) via wrapper scripts around the actual job. DESY uses Xen, KIT uses KVM with a similar approach (one job per machine) but the VM images are on a cluster file system (Lustre). In both cases the VMs have no knowledge of the batch system, for the batch system the VMs are a regular job. From Q&A: Each VM takes ~20 seconds to boot. The performance overhead of using a VM is at most 3%, there was a poster with the results of this study; There are no user images, all images used are provided by the site. They mount the software area in each VM to provide access to the experiments software. The memory overhead of each VM is of around 200MB, but their nodes have 32GB of RAM so it's not a problem.

SL(C)5 for HEP - a status report (CERN): From Q&A: There were a couple of questions: Is CERN considering changing to XFS as default file system for SLC5? Is CERN considering providing the "distribution" as simple add-ons based on one of the common distributions (CentOS for example)?

ScotGrid: Providing an Effective Distributed Tier-2 in the LHC Era: Cfengine for box management, Ganglia and Nagios for monitoring; Skype is used to keep a contact channel open between the different sites. Virtualization is used to host all the front end service nodes, for resources consolidation and improved availability with fewer machines. They improved the performance of their DPM by simply spreading the service over existing boxes and doing some basic tuning of the back end database.

Study of Solid State Drives performance in PROOF distributed analysis system: SSD is up to 10x faster than a single HDD with 8 parallel workers, even with 3 HDDs in RAID 0; there is only some scaling up to 6 parallel workers. SSD RAID brings performance increases on some specific cases, but only 20% (over single SSD) in the best case. They are planning to investigate a tiered solution with HDD + SSD.

Monitoring Individual Traffic Flows in the Atlas TDAQ Network: Using statistical sampling and a standard H/W-based solution (sFlow) they can achieve over 90% accuracy in determining individual traffic flows at any data rate, with minimal impact on the rate. They can collect source and destination for the Ethernet, IP and TCP/UDP layers,

and even the first bits of the data section for each section. All the information is collected into one central collector; they are currently looking into a distributed solution to address future scalability issues.

A Service-Based SLA for the RACF at BNL: They implemented an SLA "system" which automatically creates support tickets (using RT) and sends notifications based on the rules defined by their SLA for different services and the input from their monitoring system (Nagios). The SLA system is implemented in Python with a MySQL DB backend. They plan to merge SLA and RT because they are closely related.

The Integration of Virtualization into the U.S. ATLAS Tier 1 Facility at BNL: Yet another talk on virtualization. They are using the stock Xen shipped with RHEL/SL5. Main motivation was sandboxing of the different environments. Typically each HV has 2-3 static VMs (1 interactive + 1 or more WNs) with pinned CPUs to minimize any performance impact. They also use virtualization for some main services. A custom solution was developed for Xen management (xenconf.py). Overall the solution is similar to the "prototype" we (IT/FIO) currently have at CERN, a VM is treated as a normal machine in terms of installation and overall management.

Collaborative tools

CMS Centres WorldWide - Lucas Taylor: CMS Centre proved to be very useful for face-to-face communication for the last year. It consists in HD video links from CERN (P5 and CMS Control Centre) to other institutes (DESY, FNAL, etc). Infrastructure is provided by CERN IT for the main links. Because of the escalation of the project to small centres, EVO is also going to be used. Basic hardware needed: PCs, screens, a video system. Software: a web browser. Everything is web-based, no need of standalone applications. Examples of applications: data quality monitoring tool, ci2i (see eye to eye), CMS-TV (define programs and channels that everybody can follow using a web browser). How much for a small centre? < 15 kCHF, rapid spread of CMS centres, standard design.

DIRAC Secure Web Interface – Adrian Casajus: DIRAC is a distributed data production and analysis system for LHCb. Usually, interaction with DIRAC is via a command-line or desktop applications. They showed a new web-based solution but emulating a desktop application. Main feature: users are organized in groups and they provide a secure interface based on grid certificates. A couple of slides with screenshots were shown. Site: http://lhcbweb.pic.es

Lecture archiving on a large scale – Jeremy Herr: Web lecture is a low-bandwidth media-rich presentation viewable with a web browser. It includes synchronized audio and video lecturer streams and high-resolution slide images. Why archiving? Large scale dissemination, overcoming time/schedule constraints. Jeremy talked about CARMA (campus automated rich media archiving), a service that they provide to University of Michigan and has more than 26 University customers. They use a portable box for recording High Quality lecture objects. CERN/IT UM partnership: CERN/IT is very committed to deploying a large scale archiving technology. The partnership has three main goals: to perform a market survey of recording technologies, to implement one solution and to integrate it with Indico.

EVO – Philippe Galvez: Description of the main features of the product and statistics of usage for LHC. Usage in 2008:

- CMS: More than 6.000 meetings, 1597 phone connections, 2119 users participated in those meetings.
- ATLAS: 1977 meetings, 2997 phone connections, 1394 users participated in those meetings. ATLAS participation is very relevant using the phone bridge.

High Definition Videoconferencing for High Energy Physics – Erik Gottschalk: Description of Fermilab's perspective for the CMS centre worldwide project (see summary above of Lucas Taylor presentation): very important for a high quality and robust collaboration in real-time. Description of the biggest event during 2008: LHC Beam Day was the biggest media event ever organized and without a single technical failure thanks to CERN IT-UDS-AVC. Emphasis given to the fact that this kind of events can only be organized with very reliable technology as the one that was used.

Indico Central – Jose Benito Gonzalez: Main messages: new release will fix many usability problems and strong emphasis in the quality of the service. Usability studies and usability problems were shown. Also we had a demo showing the main improvements including conference customization, timetable, in-place editing, etc. An attendee asked about the improvement of the installation process for future collaboration with other institutes: Jose said that there is a plan for a project this summer in order to re-write and re-design the installations process.

Collaborative tools: some success, some plans – Steven Goldfarb: Description of the progress in the collaborative tools area for the LHC since last RTAG12. Big praise to the IT-UDS-AVC service (VC room installation, VC support, CERN collaborative environment, Indico service and its integration with collaborative tools, etc). Description of the next challenges for the collaborative tools in the LHC, in particular the need of a fast decision for the standard videoconferencing systems in the LHC.

Software Components, Tools and Databases

Advanced Technologies for Scalable ATLAS Conditions Database Access on the Grid - R. Walker: ATLAS jobs for reconstruction and calibration access the condition data (DCS or calibration and alignment data) from the database and T1 RAC cannot cope with the peak loads. Access from T2 is limited by the WAN. Solutions:

1. stagger jobs starts - add sleep delay as function of number of job sent
2. pilot query approach - T1 RAC configured to respond to query from WN - OK or not for job start based on DB load. Version 2 - track requests, tested at Lyon T1 - it works
3. sqlite access - DB extracted to sqlite files and accessed as a flat file - avoid DB overload and scale cost - cannot create files for every use case, used for bulk access to 2GB sqlite file - bandwidth restrictions lead to timeout – possibility to use pcache.
4. final workaround – Frontier access to oracle through web proxy cache; Frontier web server close to DB decodes and runs query and returns results same query = same URL

LCG Persistency Framework (POOL, CORAL, COOL) - Status and Outlook - A. Valassi:

- Coral - abstraction of access to relational databases, used directly or via cool/pool**;**
- pool - technologically neutral hybrid data storage**;**

- cool - conditions data management - metadata like IOV + payload used by ATLAS, CMS, LHCb experiments and other projects; new in 2008

Changes:

- dropped SEAL dependency
- new platforms – now 20 platforms supported
- gcc43, slc5, vc9 - visual studio 2008, MacOS X 10.5 with Oracle

Plans for 2009

- coral server development - see next presentation by Vallasi
- support for partitioning of schema
- ATLAS expects 300 GB /year cool data from DCS
- PF essential for LHC data taking - event data, conditions data

Distributed Database Services - a Fundamental Component of the WLCG Service for the LHC Experiments - Experience and Outlook - M. Girone: Nowadays DBs play key role in the experiments' production chain operations - Oracle streams replications and Frontier/Squid; partnership between DBAs and developers in 2008. Main points -

- took over online DBs from experiments, standard setup
- backup performances improvements from 30 MB/s to 70 MB/s
- security - isolate online DB from external access, account policies, - enforcing read and write accounts

2009 priorities - tool to detect malicious access, oracle 11g evaluations.
Conclusions - complex setup, requirements met, monitoring implemented.

CORAL server: a middle tier for accessing relational database servers from CORAL applications - A. Vallasi: Motivation: secure DB access - currently no support for X509 proxy certificates in Oracle; VOMS authorization; hide DB ports; efficient, scalable use - multiplex connections, caching tier 1 coral client using custom protocol - no need to install whatever DB client.

The JANA Calibrations and Conditions Database API - David Lawrence: GlueX experiment, high data rates 300MB/s**.** API designed for reconstruction code authors simple C++ API using C++ templates. SOAP web service is good option for external access built-in mechanism for dumping constants to local disk regardless of the source.

A lightweight high availability strategy for Atlas LCG File Catalogues - Barbara Martelli (INFN): Problem - ATLAS LFC is a single point of failure for regional cloud level purposes – need to facilitate disaster recovery, load balancing of read requests, transparency for jobs started after each failure. Solutions: exploit Oracle Dataguard in order to create a standby instance at the remote site, load-balance the read request between master and standby DB. Application failover is done with a Nagios script which monitors its local DB back-end, changes the setting of LFC and changes the DNS alias**.** Feasibility tests were done with a Python script using the LFC API, showing a maximum delay of 23 seconds**.** Power cut test - 10 sec DB failover, LFC restarted 5 sec after failover. Conclusion - Dataguard can be used as a disaster recovery strategy for Oracle. LFC back ends lag is < 30sec (required < 10min)**.**

A RESTful web service interface to the ATLAS COOL database - Shaun Roe (CERN): aim: universal web service, where each resource is identified by a URL**.** Uses XML as data transport format**;** implementation in CherryPy – a

Python application server**;** allows manual mapping of Python classes and methods to URLS**.** You can call the web service with CURL CLI or library, which is in the standard Unix OS installation plans. Conclusions: cherryPy allowed rapid development; used during 2008 as the engine for the run list query page; DCS ret future - JSON output, zipped format, sqlite file.

Ajax, XSLT and SVG: Displaying ATLAS conditions data with new web technologies - Shaun Roe (CERN): aim: graph with hyperlinks in the web page without blinking (page reloading. Plan: XML from DB -> style sheet to format XML -> inject grap. Variations - XSLT transformation can be applied to Oracle, web server, browser**.** Conclusion: code-efficient way to produce graphics from database queries

Usage of the Python programming language in the CMS Experiment - Benedikt Hegner (CERN): groups decided to use python on their own; one of the reasons for doing this is to have scripting and programming in one step**.** CMS jobs are defined by configuration files whereas previously CMS used custom declarative language which was not enough for users, so that they had to use full programming language Python. The boost is that Python is used to translate configuration in Python to C++ class

User-friendly Parallelization of GAUDI Applications with Python - Pere Mato (CERN): goal - get quicker response on multi-cores. The speaker presented a relatively simple way to improve performance of analysis by using processing or Python modules for parallel processing (both modules use a fork() system call in the end) and this is more memory-efficient than running separate processes in parallel (shared memory between child and parent processes)

Flexible Session Management in a Distributed System - Zachary Miller (University of Wisconsin): strong authentication can be expensive (CPU, network) with single threaded client blocking on networks become a problem for scalability. The solution: security session cache - semi permanent information exchange which is setup and torn down. Secret key (remembered by client and central manager) is associated with session id. The problem - network latency adds significantly to cost of authentication in CONDOR even when using cache. Solutions - central manager authenticates both the submit point and the execution node (no need to authenticate between them again). The results are that before were 4K jobs/day (500 simultaneously running jobs) and after are 200K jobs/day (22K simultaneously running jobs). The conclusion - efficient delegation and caching of trust can be an important optimization in distributed systems

Global Overview of the current ROOT system - Rene Brun (CERN): the CINT7 (C interpreter considered for ROOT) is too slow for production; maybe upgrade CINT with LLVM (Apple-driven open source project). LLVM is a gcc-compliant compiler with a parser and a jit compiler. For 2-D graphics - move to openGL? All detectors are modelled with TGEO package in ROOT. 3D GL views are highly optimized. There is a new root website: http://root.cern.ch - managed with CMS drupal. ROOT usage is rapidly expanding outside HEP. After 15years of development there is a good balance between consolidation and new developments. The main packages are entering a consolidation, optimization phase. A curiosity - some collaborations had a class name with 1024 characters.

CernVM - a Virtual Software Appliance for LHC applications - Predrag Buncic (CERN): goals – to provide a complete, portable user environment for LHC data analysis, independent of platform and to reduce cost of maintenance of the experiment software. Implementation - small linux distribution (100MB) with web interface and XML RPC API for OS management and CVMFS, a read-only caching file system based on FUSE and http which can work offline if data is present in cache. CVMFS is used to access the latest published release of physics software; it shows better performance then AFS from outside CERN. Planned improvements - http proxy; P2P

mechanism for discovery of nearby CernVMs; and cache sharing between them. CernVM uses existing content delivery networks to remove single point of failure. It could be used to run data analysis on BOINC-like infrastructure.

A comparison between xen and kvm - Andrea CHIERICI (CNAF): asymmetric network performance in kvm**;** good stability and reliability, disk i/o not so good, small effort for sysadmins to adapt quattor profiles**. R**esult of test shows that xen is still the most performing solution

Virtual Machine Logbook (VML) - Enabling Virtualization for ATLAS - Paolo Calafiura: CernVM performance with files cached has the same performance as an LXPLUS node**.** ATLAS extensions allow running CernVM with nightly builds of s/w**.** CernVM can be used to watch live events during the next LHC start**.** VML . The idea is to provide a space-efficient snapshot mechanism (several version of CernVM). **The** solution is save the differences by using tar to save differences to file, then automatically rebuild CernVM or CVMFS from base image**. The** same idea can be used to share your work (VM) with others. CVMFS could be used for software distribution at T3 sites**.** CVMFS enables easy sharing of the work environment.

GPU's for event reconstruction in the FairRoot Framework - Mohammad Al-Turany (GSI): Nvidia's Compute Uniform Device Architecture (CUDA) development tools work alongside the conventional C compiler, provided for free. It is documented and supported for Windows , Linux, Mac. It is easy to learn – an extension to C (in contrast to gpgpu). It automatically manages threads, ~1024 threads on an 8 core gpu; it uses SIMT (single instruction, multiple threads) architecture. It access external shared memory among threads but this is very expensive. It supports heterogeneous programming (e.g. a mix of gpu and cpu code). A test of data analysis with a tesla card called "monster" – 240 cores, 1.3GHz, 4GB, ~1TFLOPS, 160W – showed a factor of 60 times improvement (with 25% utilizations of gpu cores).

An update on perfmon and the struggle to get into the Linux kernel - Andrzej Nowak (CERN): perfmon is a hardware based performance monitoring tool in the OS. There is no need to recompile applications, it has a negligible 2-3% performance impact. A large group of hardware vendors supports perfmon which consists of a userspace and a kernel part (patch). It is used for counting and sampling profiling restrictions. There is a lack of robust symbol resolution (rewritten 3 times); a lack of compatibility with large frameworks; and a lack of general interoperability, although this has improved at CERN. Plans: monitoring of batch nodes, deployment across standard CERN configurations. It would need years of effort by the author and influential people to get included in the kernel. The good news is that there is serious discussion about perfmon and linux which is nearing a solution; the bad news is that the new solution is not perfmon, and lacks many features. Gpfmon is a gui frontend created at CERN. Meanwhile, there is continued perfmon support at CERN.

Distributed Processing and Analysis

Towards the 5th LHC VO: The LHC beam studies in the WLCG environment Patricia Mendez Lorenzo (CERN): Report on 2 applications from the beam group at CERN. Tracking studies for tracking the transfer beam from the PS to the SPS for the Grand Sasso Neutrino beam using beamlets. When in production they do 800 Jobs/day in 12h jobs. Simulate collimation of beam and simulation of beam halo from collimator (1200 jobs/day) 12h jobs. Both

applications are CPU limited, not I/O limited. Uses standard gridification structure: Ganga and/or Diane inside the GEAR VO. Now In contact with ITER project to use the same infrastructure.

CMS FileMover: One Click Data - Valentin Kuznetsov (Cornell University): Requesting file transfer is possible, but complicated in CMS. FileMover aims to simplify this. Web interface to allow download of data to a local disk. Implements policies such as restrict to N files per day and M GB per day**.** Users make requests via the web, get feedback via email or download link from a local disk cache. In addition, it is possible get almost immediate access to the file via CmsFS a FUSE file system that allows POSIX I/O while the file is still being transferred to the file cache. CmsFuse is at the stage of proof of concept. The tool needs more testing, in particular on scalability because so far only small files have been tested. Their main concern is abuse, that is, a user would try to download large amounts of data. Since this tool should not replace the official data transport tool PhEDex, it attempts to limit what users can do. So far it has 200 users, 300 files in local cache. It solves only interactive access. The amount of data which can be downloaded is not useful for a full analysis.

Ganga: User-friendly Grid job submission and management tool for LHC and beyond - Daniel van der Ster (CERN): General talk on Ganga. Introduces the Ganga concept, explains how to use Ganga, its main users being in ATLAS and LHCb. Emphasis is made on the testing and building framework with over 500 test cases. He explains the rotating release management. Ganga has over 1000 users with about 50% ATLAS, 25% LHCb, 25% others. Ganga is installed in over 100 sites. For ATLAS, Ganga is one of 2 entries for user into the Grid and has the ability of reaching all ATLAS clouds. Ganga can also connect to PanDA and use the PanDA monitoring. Ganga can also send personal pilots to sites which do not run the PanDA pilots. These personal pilots can then pull the jobs from the PanDA system. For LHCb, Ganga is the only entry into the Grid using the DIRAC backend. Ganga can be used to create Grid applications, e.g. the GangaRobot to do automated end-to-end tests of user analysis. In addition ATLAS uses HammerCloud built on top of the GangaRobot to do stress tests of individual sites.

Babar Task Manager II - Douglas Smith (SLAC): Distributed production system using multiple sites and local batch systems, currently SLAC, GridKA and soon IN2P3. It is used in skim production (event filtering). Each round of a full skimming production is called a cycle**.** So far 10 cycles have been made using the Task Manager II (since 2007). The Task Manager creates tasks which create actual jobs based on task and input data set. Each job creates various output streams. Designed in perl with a database backend (MySQL or Oracle). The DB contains the task and job states and information. Output is merged and transferred back to SLAC for archiving and added to the bookkeeping.

Scalla/xrootd WAN globalization tools: where we are - Fabrizio Furano (CERN): Discusses 2 use cases for accessing all data through the WAN for a physicist wanting to use ROOT on his laptop and access data. Accesses the conditions of ALICE from the central ALICE conditions database. In both cases access is possible even with high latency (180ms).

End-to-end monitoring for data management - **Sophie Lemaitre (CERN)** - Debugging file transfer is expensive. Developed tools to aid debugging. Currently the only way to debug is to grep over all log files. Problem turns out to be more of an integration problem. Our previous attempts focused on recording all events – you collect all events from all sources, all the time, parse them, and put them in an index database. On-demand extraction from data sources (request/response). Integrate (join) the data from the various sources for that specific debug request. Instead of doing a grep over all logs, make a request to retrieve on demand debug information using MSG.

Scalla As A Full-Fledged LHC Grid SE - Andrew Hanushevsky (SLAC): Development of a plug-in for the Scalla/xrootd system to create a full SE. For SRM access a FUSE module was developed to attach an SRM (BEstMan) onto Scalla. In addition a gluelib has be developed to interface to gridFTP. Since SRM assumes a file system like access, a global namespace has to be created for the entire cluster. That is achieved by having all file servers notify a redirector of any POSIX like operation on its files.

CMS data quality monitoring: systems and experiences - Lassi Tuura (Northeastern University): For online, 10% of recorded data goes to online Data Quality Monitoring (DQM). Histograms (300k produced, 50k visible from GUI) available within a few minutes from a web server. Archive of 6 months accessible from same server. Output from DQM plus input from shifters goes into initial conditions database. For shifters the expert has selected quality histograms for sub-detectors in a high level view. These histos can show alarm states including documentation. For offline DQM histograms saved in normal data are harvested periodically and merged. Creates conditions. Once certified/signed off goes into final state with flags for analysis. In general, a lot of effort is made towards high level views to help shifters to discover problems early and easily.

Performance of Combined Production and Analysis WMS in DIRAC - Stuart Paterson (CERN): 1M jobs since beginning of the year. 62% production, 28% user jobs, 10% SAM. A max of 15k concurrent jobs, 40k jobs per day.

A generic Job Submission Tool (JST) - Guido Cuscela: A talk from within the Bioinformatics field**.** Jobs created from a Task. Status of tasks and jobs is stored in a DB. Typically jobs are short and small. Tasks are normally tolerant towards failures of individual jobs which are resubmitted on failure. Main advantage is that it can run in pull mode. If nothing is to be done the job exits.

PROOF-Lite: Exploiting the Power of Many-Core Machines - Fons Rademakers: Make full use of your multi-core machine when using ROOT. This is a zero configuration version of PROOF. The only additional call a user has to make is to do a Proof::Open("") call at the beginning of the ROOT macro. In addition, the analysis should use a TSelector and let ROOT do the event loop. Bottlenecks are typically I/O. The use of multiple disks (optimally number of disks = number of cores) and the use of SSD disks helps. Once the setup works it can be easily extended to a full PROOF cluster.

Building a Reliable High Performance PanDA Facility - Dantong Yu (BNL): A report of the hardware setup of the PanDA facility at BNL. Main focus is made on redundancy and failover. Most of the setup will migrate to CERN at some point.

Distributed Analysis in ATLAS using GANGA - Johannes Elmsheuser (Ludwig-Maximilians-Universität München): Ganga is one of two approaches for ATLAS users to submit jobs to the grid. A new PanDA plugin now allows to submit jobs to all available ATLAS grids and clouds. ~1600 users using Ganga, more are expected. A new mailing list based support structure with shifters has been established

Performance of an ARC-enabled computing grid for ATLAS/LHC physics analysis and Monte Carlo production under realistic conditions -Bjoern Hallvard Samset (University of Oslo): The NorduGrid cloud in ATLAS performs well for MC production. They are now preparing for user analysis. To test for this they used HammerCloud to test for likely problems. An initial test with HammarCloud brought the NorduGrid cloud down, mainly due to large file transfers going on prior to starting jobs (This is a NorduGrid feature, that data is fetched before the jobs started.) This was solved by limiting the amount of data a user can access per job to 5GB. A user therefore might have to split his jobs into smaller sub-jobs. A second test with reduced amount of data per jobs then worked fine. Currently the NorduGrid team is trying to understand which data access pattern (remote site read, read from HSM, or copy data to worker node) is the most efficient.

A Comparison of Data-Access Platforms for BaBar and ALICE analysis computing model at the Italian Tier1 - Fabrizio Furano (CERN): Compares the production setup for ALICE and BaBar. Main difference is the use of optimisation and read-ahead settings of the ROOT and xrootd set-ups for both experiments at CNAF. ALICE runs with all optimisations and monitoring switched on, whereas BaBar has turned them off. In addition, the report discussed the influence of using xrootd over GPFS (IBM's distributed file system) or using xrootd directly. Conclusions are that turning read-ahead and optimisation on tends to be more efficient. Also, using GPFS tends to be more CPU-efficient.

-

User analysis of LHCb data with Ganga - Andrew Maier (CERN): Report on the status and usage of Ganga in LHCb. There are about 60 LHCb users of Ganga per week. In total, close to 200 user are using Ganga. This means that Ganga reaches all LHCb users as foreseen in the computing TDR, which assume 140 users. The developed application and backend plug-ins for LHCb in Ganga considerably simplify the usage of Grid and batch resources for LHCb applications. The application plug-in takes care of configuring the application for submission, automatically extracts user libraries to be sent, as well as extracts and automatically retrieves outputs from the job without further user intervention. The LHCb-specific DIRAC plug-in interfaces jobs to the DIRAC WMS. The DIRAC backend is the only supported grid backend for LHCb users.

Functional and Large-Scale Testing of the ATLAS Distributed Analysis Facilities with Ganga - Daniel Van Der Ster (CERN): Both functional and large-scale tests are performed using Ganga in order to increase the reliability of jobs for ATLAS users. Functional tests are performed using the GangaRobot. These tests include end-to-end tests of short user jobs. Failure of these tests can trigger actions such as blacklisting of sites. Large-scale stress tests are done using HammerCloud which is based on the GangaRobot but here the emphasis is on sending a large number of larger user jobs to a specific site or to a cloud. HammerCloud can detect bottlenecks in the tested site, e.g. I/O or bandwidth limitations, and can help to address these problems.

Monitoring the efficiency of user jobs - Ulrich Schwickerath (CERN): The LXBATCH farm at CERN on average uses 70% of the available CPU. This is often due to a bad choice of the jobs running together on one node, e.g. two jobs that are I/O-bound. Improvements have been made by trying to schedule jobs where one job is I/O-bound and the other one is CPU-bound. This guessing does not always work, but has improved efficiency somewhat. But it could be improved if the system is told what type of job has been submitted. Therefore tools have been developed which enable the monitoring of user jobs. The tools rely on the MSG system and allow to send tokens and information to an Oracle DB where they can be analysed later. The end goal would probably for experiments and users to be able to give hints to the system for better resource usage.

Distributed Monte Carlo Production for DZero - Joel Snow (Langston University): Usage of a variety of resources for DZero MC production. With the reduction of manpower due to people moving to LHC experiments, a lot of emphasis is made on automation and fail-over.

PROOF on Demand - Anar Manafov (GSI): Application to create a private PROOF cluster on demand using worker nodes on either LSF or LCG. A GUI simplifies the set-up of this cluster which, once created, can be used like a normal PROOF cluster. Looks very simple and nice. Certainly a nice thing to create a small ad-hoc cluster, but of course wastes resources as the PROOF slaves are mostly idle. Very nice talk.

Posters

Each of Monday, Tuesday and Thursday, they organised a poster session consisting of the display of some 100 posters each day. The morning coffee break was extended to a full hour and both then and partly during the lunch hour, attendees were encouraged to review the material and discuss with the author(s). There was the usual mix of highly graphical and colourful displays and one or two consisting of individual pages of the Proceedings contribution cut up into sheets. Unfortunately, none of the summary speakers spent any time discussing posters, although during one summary, a member of the audience reminded us that Indico is perfectly capable of accepting electronic versions of posters and he encouraged all poster authors to upload their material. This message should be passed to all CERN poster authors.