CHEP06, Mumbai, India

13th to 17th February 2006

This trip report has been compiled from contributions by Michal Kwiatek (various parallel sessions), Gavin McCance (DEPP sessions), Ioannis Papadopoulos (Software Components and Libraries track), Bill Tomlin (Software Tools and Information Systems) and Alan Silverman (plenaries, some parallel sessions including online processing) and edited by Alan Silverman. All errors, misquotes and omissions are the fault of the editor. The fact that sessions early in the week are reported more fully than those at the end is noticable, all correspondents were flagging over time and/or succumbing to the local cuisine! The overheads are almost available from the web site (via InDiCo) and the Proceedings will be published in due course. This report should serve to point you to talks of interest to you.

The Conference, the 15th in a series which takes place approximately every 18 months, was held in the Tata Institute for Fundamental Research (TIFR) in Mumbai. TIFR are celebrating their 60th anniversary this year. It's hard to know how to judge the organisation of the event because all agreed that is unfair to apply US or European criteria to a developing nation. The organising team certainly put in a huge effort but rather often they were busy with issues which might have been foreseen. The biggest headache for them was the Indian President's visit on the Friday. Obviously this is not something you slot into the programme if the "speaker" is late – as he was – and the Friday schedule was rearranged a number of times. It was a major coup to get him there and certainly worth it in the end but at times on Thursday and Friday they were doubting the wisdom of it.

There was no overall conference summary this time so everyone is left to make their own judgement. As usual, the plenary speakers were a mixed lot, Wednesday's selection rather good, Tuesday's not so good, with one exception. A lot of emphasis was put on the Digital Divide and network and systems experts from both OSG and LCG worked all week to set up very dramatic demos for the President's visit. Also during the week, the link between TIFR and "LCG" was significantly upgraded and jobs started to flow into TIFR.

The parallel technical sessions were also of mixed quality as usual but this is not surprising given the number, 8 parallel streams for 4 and a half afternoons (CHEP04 had 5 and a half streams for 3 afternoons). However the summary speakers for each stream were positive regarding the quality. A lot, more than 50%, of the papers referred to preparatory work for one or other LHC experiment of course. One comment I overhead is that theory is now starting to be put into practice as real (calibration) data starts to flow from the detectors now being assembled, especially true of course for the online teams.

Attendance was estimated at 450 to 480 but in addition a number of people attended the pre-conference LCG Service Challenge meeting and then returned to CERN. Others participated in the pre-conference mini-Computer School. All told, the organisers estimated they catered for some 500 attendees.

Plenary Sessions

Monday

The conference was opened by the TIFR Director, who spoke of the Institute's history and how and when it became involved with HEP although its science programme is much wider than that. He noted that this will be the last CHEP before LHC startup and TIFR was very proud to host the conference. He was followed by a number of other local dignitaries who noted successively the importance of grids and grid software for modern experimental teams not only in HEP but also in other disciplines and, increasingly, beyond science. It was noted that TIFR is a Tier 2 site for CMS with a gateway for the ALICE group in Calcutta.

The first talk of the conference was given by **Jos Engelen, the Chief Scientific Officer of CERN, who spoke about the status of the LHC.** He showed photos of the ongoing magnet installation and inter-connection work. He displayed an extract of the LHC dashboard which shows the advancement of the installation and testing and spoke of the pilot run planned for 2007. He then covered the status of the experiments.

Jamie Shiers then reported on the readiness of the computing infrastructure for LHC. Since the last CHEP, the requirements have crystallised into the Technical Design Requirements (TDR) documents of the experiments as well as that of the LCG project. These requirements are exercised in a series of LCG Service Challenges in a production-like environment. All these, including targets and how to measure them, are defined in the LCG Memorandum of Understanding (MoU). Although it is expected that initial luminosity will be lower than the eventual target, this will be compensated by experiments using a more open trigger so LCG must plan for full data rates from the beginning of LHC commissioning. He described several issues such as site and user support and the need for of minor and major service upgrades when the LCG service will be expected to be continuously 52 weeks per year. The recent Service Challenge achieved an aggregated 1GBs to the tier 1 sites and up to 200-250MBps to individual sites. For how long these can be maintained and indeed ramped up to the full throughput needed for LHC data taking and distribution is debatable but Jamie is "cautiously optimistic" although there is a considerable amount of work to be done.

Paris Sphicas from CMS described the readiness of the experiments' software. He confirmed Jamie's fears of a data-intensive startup despite expected low luminosity, explaining the physics which is expected and noting that least one experiment (ALICE) is expecting to publish 2 papers within 1-2 weeks of first collisions! Turning to the software, the common software applications area projects (SPI, ROOT, POOL and the simulation tools) are well advanced and well used across the experiments. All except CMS have well-defined frameworks with CMS in the process of redefining theirs. GEANT is deployed across all experiments with very good progress in data generation. So-called fast simulation is less well advanced and at different levels in different experiments. All are aiming at outputting results at the AOD level. Reconstruction, triggering and monitoring are based on the corresponding frameworks, all aiming at commonality between online and offline. The advancement of the software for calibration and alignment is variable across the experiments with many open questions – how much detail should the simulations contain, how to manage the bookkeeping, etc. On documentation, his only comment was that the ATLAS documentation was extremely impressive and he urged his colleagues in the other experiments to check it out. Analysis software is progressing with a common understanding of the format of the data to be used. Moving on to user analysis, since the 90s, the development of ROOT and that of powerful laptops have made in-flight data analysis compete with in-flight entertainment systems and with wifi developments in the 2000s we have "physics in a meeting" – Microsoft's new term CPA (continuous partial attention). What's left to do – more realism in terms of detector performance, real commissioning experience, In summary, the overall shape is “OK” but there is always a difference between theory and practice. Much software is in place and performance starts to be the main challenge. Deployment has begun in earnest but there are still many milestones ahead.

Dr.. Jhunjhunwala ended the morning with a talk on networking in India. He explained why it had boomed recently (lowered costs to $7 per month) and how it could “reach the unreachable” for which he estimated it would have to drop to $2 per month. For broadband to become common in homes, his estimate is that it should cost no more than $6 per month and he foresaw 50,000,000 connections by 2010 if this could be achieved. He described cable-based networking and NetPC devices which would make networking affordable. Rural India would remain a problem – 700 million people in hundreds of thousands of small villages but he hoped that developments in wireless broadband technology would help. Local internet kiosks could be another mechanism. He displayed a device produced in Bangalore costing $250 which permitted basic remote tele-medicine functions such as temperature and blood pressure measurements (audio and digital).

Tuesday

Beat Jost gave a very full review of future Data Acquisition Systems, not only for the LCG experiments but also for other, non-CERN planned or projected experiments up to the ILC in 2017. He believes that Ethernet will remain a dominant technology but at yet higher speeds. ILC and CLIC experiments will need trigger-free DAQ because of very short, sub-nanosecond bunch spacing.

Liz Sexton-Kennedy of Fermilab (CMS) gave a workman-like review of event processing frameworks, which mostly consisted of reading her slides, and then a very detailed feature review of the frameworks of some running experiments. Her conclusion was that these were only partially successful with respect to their goals. She finished with a similar review of the features planned for the LHC experiment.

Martin Purschke presented some experiences from the PHENIX experiment at RHIC at BNL which is taking data at rates which approach those of the LHC experiments To avoid adding a second level trigger they boosted their data taking rate 60 times above their initial 20 MBps target to several hundred MBps. Techniques used include data compression and buffering. He related their experiences in taking and then analysing their data at these rates. Lessons learned include –

- increased data rates help resolve difficult triggering problems common in heavy ion experiments
- data compression is key
- don’t be afraid of taking more data than you can analyse immediately.

Tony Hey, now at Microsoft gave a talk on e-Science and Cyberinfrastructure. He presented Microsoft’s commitment to working with the open source, common standards community. He gave an example of a UK chemistry team who use e-Science to link their various research activities but he claims that the particle physics community has never really been involved with e-Science but says we don’t have to because we “are smart people and know where [we] are going”. He listed 6 key elements for a global cyberinfrastructure for e-Science and says web services can provide a framework. He hopes to push Microsoft in the direction of entering this world of open standards and interoperability with other players in this space.

David Axmark, co-founder of the company behind MySQL, gave an overview of the open source product. The company has grown from 2 people to almost 300, selling mostly support and consulting. They also sell propriety licences for anyone who wishes to develop the code in another direction. He described various storage engines behind MySQL developed and optimised for different environments ranging from HEP to banking. He presented a list of new features of version 5.0 and others coming in future releases. He explained how publishing the full source code provides excellent and timely feedback and makes his product so good (his words). Can one complain about a sales talk from an open source company? He was certainly more “commercial” than the preceding speaker from Microsoft. More seriously, should it be a CHEP plenary talk?

Completing a trio of “commercial” speakers, **Alan Gara of IBM spoke about supercomputing**. He started by explaining why supercomputers have become more common – the intersection of invention and the marketplace. He showed a number of the scientific challenges for which high performance computing is required. He then briefly described some recent supercomputers such as ASCI Purple and Blue Gene/L. BlueGene has 32 TB of RAM. Although this sounds a lot it can be a limiting issue because you have to divide it by a large number of nodes.This is on purpose: to double the memory you can either double the number of nodes and stay with the same amount of memory per node or double the memory per node. Apparently the price is the same. IBM decided for the first solution, which gives you more CPU power. He showed how CMOS has provided performance advances but notes that power is now the problem and we have to hope that yet again innovation will provide the answer.

Wednesday

Harvey Newman from Caltech reviewed the status of worldwide networking for HEP. He noted the recent rapid growth in traffic, following the predicted roadmap given by various reviewers in previous CHEPs. This follows a trend in the world at large. He reported briefly on the report produced by an ICFA sub-group on networking which Harvey chairs. The growth is fuelled by new technologies, by the rapid spread of dark fibre, high speed links and 10Gb links will be common by the time LHC starts production. Network infrastructures are advancing in many countries, including some of the emerging nations (see an interesting slide in his talk which lists these by country). In Europe GEANT2 is transforming research networking. Similarly the US networks serving LCG participants are being enhanced. Major issues for 2006 include improved monitoring (see slides). He described in detail a successful demonstration of high speed networking at Supercomputing05. A focal point of future work is to understand and start to close the so-called Digital Divide, referring to the work of Les Cottrell and others (see report below on some of Les’s work). He hopes and believes that initiatives taken by the HEP community in third world and emerging countries may lead to network improvements in a wider sense in these countries, and quoted some successes, for example in Brazil. In Europe, inter-country dark fibre is being exploited to connect to GEANT, for example in Poland and Ukraine. Africa is still a major concern, not only in terms of famine and disease but also in terms of the lack, which is growing deeper, of networking. Some initiatives have been proposed, including some supported by CERN (WSIS conference, Africa@home). He ended on an optimistic note, network is getting not only faster but spreading geographically and emphasis is being placed on closing the Digital Divide.

Les Robertson then reviewed the status of the LCG service. He covered the mission, the roots and the history of LCG, relating the milestones to previous CHEPs. He showed the links between LCG and the phases of the EGEE project and how discussions have started to see what might follow Phase II. He explained how the US Open Science Grid participates and how future funding is being proposed. He then compared the technology advances to those predicted by the PASTA studies of the 90s and concluded that by and large these were relatively accurate, the exception being the use of object databases, although the cost predictions were way out. In CHEP Padua there were optimistic views on how easy and quickly it would be until we had a production grid. In CHEP Interlaken, there was a feeling of frustration due to many problems but Les believes recent performance is more optimistic and he compared LCG performance to the well-known Gartner curve of highs and lows of performance of a new project over time (see overheads). He ended with the feedback from the Service Challenge meeting held the previous weekend in TIFR and the plans for SC4.

Ruth Pordes of Fermilab described Grid Interoperability as seen from the US. Consider a grid as a shared resource where the users have agreements on what they offer and what they can use. Interoperability inevitably means collaborative and includes human and social factors. Ruth claims this leads to gateways, resource interfaces and overlapping services, all using a common network fabric. Interoperability often involves in-depth testing, demonstrating that it works in practice, as well as continually talking among the partners. She noted how LCG was a pioneer in showing interoperability between grids in different continents and also the work of the Datatag project. She explained how there was linkage between EGEE and OSG with some basic interoperability requirements and agreed validity checks. This will lead to more cross-grid VOs (virtual organisations). The trend is for an increasing number of grids, at all levels from campus to national to international. This only emphasises the importance of interoperability and she noted at the end that the agenda of the GGF meeting being held in the same week in Athens appears to show that GGF thinks this also.

Wednesday morning ended with a talk by **Rene Brun on "Root in the era of multi-core CPU**s". Although 11 years old, there are still many developments in ROOT. He assigns partially the success of the project to initial antipathy to it but it is now accepted and heavily supported in a number of labs and he thanked his colleagues in the LCG project for this. He showed an interesting plot of human resources which had been invested thus far. Assuming vendors' claims for multi-cores are correct, there will be an impact on ROOT – imagine the power of a laptop in 10 years (32 cores? More?). But to use this, applications must be multi-threaded – a future development of PROOF for example. Rene believes in this timescale a laptop or similar would be a more attractive alternative of a set of batch jobs submitted to a grid. Also, the powerful multi-core PC, or a local cluster of these, is inherently less complex than a grid and could be therefore an easier alternative. Rene then turned to the use of shared libraries; although he could not imagine having got to the present place without them, he believes their inherent inefficiency (link time, load time) means that we need to think of another method for the future – BOOT, a bootstrap software system, a very small subset of ROOT. By means of use cases, he described how it works; for details, see Rene's overheads. He admits there is still a lot to do to make BOOT a reality and at this stage he raises this pre-project only to gauge reaction, including possibly negative reaction but he firmly believes it would lead to major gains in productivity and efficiency. He made a rendezvous for next CHEP when he hoped he would be able to demonstrate this. In the questions, a leading member of the ATLAS offline collaboration gave a stern warning about the dangers of thread-unsafe user applications built above possibly thread-safe frameworks.

Thursday

On Thursday afternoon, **Wolfgang von Rüden gave a public lecture on "From WWW to the Grid**" in which he tried to draw some parallels on how the latter may one day affect our everyday lives just as the web does now.

Friday

Randy Scobie, the new chair of IHEPCCC, gave a short review of its activities, including it's use of HEPiX for providing feedback on technical issues.

A speaker from the Indian Research Group of Google gave a talk on data mining but he admitted that I was probably too commercial for this audience – so why did he not fix this before coming! Anyway, he explained why Goggle was the best search engine for avoiding indexing spam and how and why they use commodity computers. He noted that power consumption is a major issue and will become more so.

The **President of India came on site to give the valedictory address**. A trainee rocket scientist at TIFR 40 years ago, his talk, given using powerpoint which had prepared the previous day, showed signs indeed of a scientific background. It was also illustrated with screen shots of LCG monitoring. His theme was the knowledge grid and how it could be used to help in Indian development. He ended by offering three missions for TIFR over the next 2 years –

- computers in particle physics, including in the studies of the standard model in LHC experiments
- space and particle physics research
- energy research, using the knowledge gained in helping in building the LHC

Distributed Event Production and Processing

Monday

LCG Service Challenge – Jamie Shiers: more detail on the morning's overview with emphasis on the results of SC3 and plans for SC4, which will add analysis use cases to the work load. In the first run of SC3, they didn't make the goals on stability or throughput but in the rerun, most sites reached and often surpassed their nominal data rates, eg. IN2P3 and RAL nearing 200MB/s and SARA achieving 250 MBps. In recent disc to Tier 1 tape tests, DESY had achieved their target of 100MBps. Now the task was to make throughput peaks into average rates. Among the lessons learnt were that rolling out stable production systems takes a long time and there is a need to improve grid operations and grid support. Questions:

> Q. How long can T0-T1 export be down until there is a problem?
>
> A. Few days buffer – but monitoring and information to debug is the key: shouldn't be down for more than the targets in the MoU.
>
> Q. Are tier-2 roles understood?
>
> A. Experiments are understanding them. Simulation role is fairly well understood. Calibration and analysis is not yet fully understood.

LHC@HOME- Jukka Klem: he first explained the architectures and basic principles of BOINC ("volunteer" computing - people give spare cycles on their machine to some scientific group to do number crunching) on which this is based. Job redundancy is important – sending jobs to multiple computers and comparing and validating the results. Cool screensavers + credits (a leaderboard is published) make people interested in the application (so they run yours rather than the several others listed in the examples). It is suitable for parallelizable applications with preference for CPU bound apps, though some BOINC apps do use a lot of data (not LHC@home where the actual application is SIXTRACK to simulate LHC particle beams – looking for conditions that cause chaotic movements in the beam bunch -> understand conditions that lead to beam loss.). To date they have 15,000 active users running on 25,000 hosts. They have produced 800 CPU hours of 1 KSpf2K (a 2.8 GHz Xeon) – a factor of 100 times in computing power that would otherwise have been available for the application.

Massive data processing for ATLAS combined test beam - Frederik Orellana: Review of experience of the variety of grid systems that Atlas uses for the 2005 combined test beam reconstruction and analysis.

1. Initial model tested: all on local CERN resources. Local batch system, local conditions DB. Noted problem of overloading both.
2. Use of LCG production resources. Fine.
3. Nordugrid @ Switzerland(?) resources. Using local copies of everything including conditions DB.

In all cases, "special" production-like jobs for the task are submitted via a GUI designed for the task. The idea here is to allow sub-groups of people to do "small-scale" production-like jobs (testing e.g. particular calibration) without bothering the general ATLAS physics production staff. "Put the small production into the hands of physicists".

Database Access patterns in ATLAS computing model - A. Vaniachine: Calibration and alignment data is critical to the success (and a prerequisite) of basic reconstruction and analysis. It's a major component of the System Commissioning. Goal is "closed loop" calibration – i.e. iterative calibration Initially will focus on steady state calibrations. Diagram of (database) calibration dataflow in Atlas TDR. Three applications: geometry, conditions and TAG data. Testing with realistic data sets. Issues and lessons: chaotic grid jobs create varying load on conditions DB servers – sometimes overload them – limit of number of concurrent connections. Opportunistic grids (i.e. not production grids) used for a while but are not really sustainable in the long run. DB server indirection – use "logical" DB name in software – middleware finds the the real DB. C.f. replica location in data file management. Participation in LCG 3D for replicating the TAG data using streams and caching. Particaption in OSG edge-service dynamic databases. Project DASH to grid-enable MySQL and ANL project to grid-enable Postgres. OGSA-DAI. Development should focus on grid / database integration: there is a gap.

ATLAS computing model – Roger Jones: Review of ATLAS computing model. Computing model well evolved. Few unknowns: calibration and alignment, analysis, event size are still not fully clear**.** Review of Tier responsibilities. Implies large T-T1, T1-T1, T1-T2 data movement. Simulated data at T2 moving to T1. Data flow pictures + T1-view pictures. Rough bandwidth requirements. Reprocessing a ttier-1 understood, but still issues to be resolved – most of these came up at the SC workshop before:
- tier 1 recalling from storage, file pinning, pre-staging
- different storage classes at the tier-1: major issue at workshop

How should ESD data be streamed? TAG access (ATLAS have same data both in file and DB format): tier-1 will have all TAG, and it should be on RDBMS, since users will run large queries on it. Tier-2 will have TAG for data it stores, and can use the 'file' format for TAG – slow but not heavy queries. File TAG can be used to directly access AOD data file. Production is well understood. Analysis model needs further development. Some issues will require real data to be resolved.

ATLAS experience on large scale productions on grid – Gilbert Poulard: Review of data challenge experience. Data challenges were suggested in review to validate the experiments computing models. Started DC1 to validate reasonable set of software. DC2 added new ATLAS production system. Rome tested simulated data.Diagram of production system: it can make use of multiple grids (each has an executor) – Don Quixote (DM system) can transfer data between these grids. Executor – one for each grid flavour**.** DMS system – global cataloguing of files. Uniform view of grid-specific data management tools. Allow moving between grids. 3 grid flavours: LCG-2, Nordugrid, OSG Grid 3. Production stats. Monitoring – fairly even distribution of jobs over all sites. Observations of grid systems: submission rate irregular. Still takes too much manual intervention – the rate drops at weekends! Use of Condor-G executor to submit jobs directly to LCG resources – bypasses RB. This doubled job rate. DC2 Experience: the exercise helped debug middleware. Data management systems were lacking. The criticality of some of the services was underestimated (e.g. MyProxy, RB, etc)**.** 3-Grid solution works, but takes a lot of manpower: improvements still needed. Future: Moving to continuous production mode on SC4 service (starting before). Several more exercises planned this year – including real cosmics.

Studies with ATLAS trigger DAQ – Gokham Unel: Review of architecture of ATLAS trigger DAQ system. Review of rates at the various trigger read-out steps. Validation of components: the read-out rate is above requirements. Even for worst case, data readout time is 1% of allocated trigger time. Plenty of time for trigger logic to run. Reliability testing went well. Conclusion: ATLAS TDAQ works to spec and meets requirements.

Use of distributed and object based file system at BNL - ? for Robert Petkus: Review of filesystem in use and considered for use by BNL. Pros/cons for each of them.

- NFS: mature / ok rate BUT insecure, poor scalability.
- AFS: good for static software
- Panasas: fast, but reliability problems.
- dCache: good, responsive team, but pnfs is a bottleneck and single point of failure. No user/groupo quota management.
- Xrootd: fast performant, no single point of failure, scales well. Not yet merged with SRM.

Moving to distributed storage. No single impl provides everything. USATLAS will use dcache. Star will use xrootd (SRM version when its there). Panasas disappointing. AFS for static software. NFS continues (badly) for home dirs.

OSG-CAF – Single point of submission to the OSG for CDF – Matthew Norman: OSG CDF analysis facility. “Single point of submission”. CDF’s portal to OSG resources. Local CAF jobs are managed by Condor scheduler. Want same for gid jobs. Users use Kerberos, don’t want to have to know about grid certs. glideCAF (see other talk): user submits jobs to Condor as normal. glideCAF component talks to OSG schedukers and brokers and arranges for generic pilot job to be ‘glided in’. Once it starts on the worker node, it calls back to glideCAF Condor and get the real job to run. Issues with firewalls ~solved by GCB ‘proxy’. Input tarballs can be large for CDF software – caching solution with distributed web-caches (squid) saves bandwidth and load on central software server. These squids can also be used as part of the frontier service. Nice flow diagram. Software retrieval via parrot (locally mounted http filesystem). Via squid cache. System scalability – Condor C being investigated.

GridX1: Canadian PP grid – Ashok Agarwal: GridX1 resources integrated into LCG with all gridx1 resource appearing as a single LCG compute element. Job submission from LCG subsequently goes via the grid X1 interface. Underlying system relies on Condor G. Lots of jobs processed.

Tuesday

GlideCAF - A Late-binding Approach to the Grid - SARKAR, Subir: CDF analysis framework. Uses Condor glide-in mechanism to get jobs to worker nodes. Kerberos authentication upon submission. CAF system uses tools to provide interactive monitoring of your running jobs – web browser. CDF previously – most analysis done at central Fermi CAF. Simulation done offsite. Changing to use analysis facilities aoutside fermilab -> Grid. Glite-ins are pilot jobs – submitted to grid, they start a Condor startd on the node (in user space) and then pull the real job from the Condor pool. The WN joins the cdor pool temporarily until the job is done, then leaves it. Features: Fast failure - if a submission / WN fails, you know it early: only the pilot job fails, since it doesn’t pull a user’s job from Condor until it is happy. Basic VO ‘sanity’ environment checks can be run in the same way before committing the job to the resource. Currently uses single “CAF” proxy – this is being addressed – allowing Condor to send the real user’s credentials to the worker node. Plan is to use glexec on CE for this. Issues: Firewalls get in the way, Condor UDP over WAN. Both are being resolved using software solutions that shouldn’t involve things being installed on the resource sites.

A generic approach to job tracking for distributed computing: the STAR approach - Dr. FINE, Valeri: STAR job tracking and distributed monitoring. Built on Jakarta log4j framework – this has multiple language bindings all reading same config file. Interesting approach. STAR SUMS submission system adds extra ‘appenders’ to log4j family to add specific things – global jobID, sending logging to their specific logging database. Production manager can define things that will be put in the user’s logfile (jobID, host/IP, etc) and these will be appended to whatever the user has chosen to log.

Grid Deployment Experiences: The current state of grid monitoring and information systems - Mr. FIELD, Laurence: Review of all the major node / job / service monitoring strategies. They all has rather similar architecture but are fairly incompatible – although there exists some specific ‘bridges’. Apel, Lemon, RGMA. MonaLisa, MDS, gridCAT. Interoperability matrix: it’s a mess. Need general system: common components. Agree on common schema for this information. Common sensors that can publish to any monitoring fabric. We need to consolidate, but it’s difficult to do this.

> Q. General agreement on this point?
>
> A. People agree something needs to be done – we can’t share monitoring data – trying to spark off more discussions.

Data and Computational Grid decoupling in STAR – An Analysis Scenario using SRM Technology-Dr. HJORT, Eric: Use of DRM (a lightweight SRM) for handling the data control to and from jobs. The past: SUMS submission to local batch system at BNL. Today: SUMS is still the interface, but CondorG makes the grid resources available. Problem: (input and output) sandbox was too big. This creates problems for the data management. No way to manage the local scratch space. DRM is solution. Some sites install it explicitly: this is just an SRM. therwise, DRM can be glided-in and started in user space. You still need to know what scratch space it will manage, but once that is known, it can manage it for you for the length of the job. DRM dies with job after completing its final transfers. enefits: copy's are queued reasonably. Files don't go through gatekeeper→ protection of gatekeeper against 'bursty' job submission. Performance has been tested and is good - sufficient for STAR.

The LCG based mass production framework of the H1 Experiment - Mr. WISSING, Christoph: Review of H1 gridified MC production framework. HERA luminosity upgrade – this has substantially increased the computing demands. 1 year (2005) was more than entire pre-2000 integrated luminosity. Previous production chain: local based. Current production chain: same system now submits to LCG grid via DESY resource broker. Agent based production system, submitting to grid, retrieving results, validating. User-friendly web monitoring tool to see what the state is. Early problems noted – LCG 2.6.0 was much better. Automated agent system takes care of resubmissions – very low terminal job failure rate. 120 Million events in ½ year. Peak 20 Million per week. c.f. total set of 400 Million events produced on H1 previously. Looking to extend system for analysis tree production jobs. Prod framework is ready -> grid has arrived on H1.

Experience Supporting the Integration of LHC Experiments Computing Systems with the LCG Middleware - Dr. CAMPANA, Simone: Review of EIS team's work. Main focus: integration of middleware into experiment frameworks – support to experiments for this. Focus is production rather than end-user support. One FTE for each LHC experiment. Actual tasks are wider: integration, user support, infrastructure expertise: filling the gaps? Tools developed where gaps were seen -> where approipriate these had been fed back into the grid middleware: e.g. g-peek to see current output of job. Monitoring tools. Common VO box services developed / understood here for VO box deployment. Expt specific things done (more on slides):

- Alice integration
- ATLAS support for DC2 / Rome
- LFC evaluation for POOL CMS
- LHCb DC04 support, monitoring
- Support for Biomedical / Geant4 / Unosat.

Focus now: integration. Following up on operational issues.

CMS Monte Carlo Production in the Open Science and LHC Computing Grids - Dr. GARCIA-ABIA, Pablo: Experience of MC production framework on grid. New framework coming is, new event model – supports grid better. Improving robustness of framework. Production chain: generator -> simulation -> digitization -> reconstruction. Publication makes use of phedex. Data published in global refDB and local pubDB. Analysis tools use these for finding data subsequently. Production workflow: Preparation: submit via RB with logical file constraint. Start on WN. Download files from SE Run Summary file. Problems: staging in/out, no CMS software, local config problems. Poor application error reporting. Slow submission rate. Improvements made:

- Sandbox size reduced improves submission rate.
- Better staging procedure. More robust – retries, etc.
- Output zip archives -> less, larger files.
- Pile up support – explicit copying of pile up to jobs to make it available locally for jobs. Critical for digitization and reconstruction jobs.

Other ideas: local user-space CMS software install. Local pile up installation. Production operation experience: Processing lagged assignment. Various problems: manpower at sites, lack of dedicated resources. LFC migration was smooth and helps a lot.→ significant performance gains. Invaluable experience gained. Robustness of production frameworks very important.

Development of the Monte Carlo Production Service for CMS - Dr. ELMER, Peter: CMS MC production system – the details. Building on lessons from previous production system – see earlier talk. Inspired by Babar Sprite system (see Babar talk DEPP 299 D. Smith)→ easy to run .. minimal babysitting Prodmanager / prodagent framework: Prodagants per grid (or special ones per site/host). ProdManager: allocates jobs to prodagents. Interaction with policy manager. First prototype becoming available (prodAgentLite) – will run full MC chain in a couple of months. Much more robust error handling – catches and log environment failures, catches and logs application errors. Full reporting in summary file. Extend to analysis: "MyFriend" concept. Personal (user) agent to automate boring task of tracking submitted jobs / basic data management. Lives at T2s. Connect to it, give it instructions and disconnect. ProdAgent is prototype of this type of agent. Summary: CMS preparing new MC production system using new EDM, framework and data management system. Highly automated, robust. Ramping up in a couple of months.

Italian Tiers hybrid infrastructure for large scale CMS data handling and challenge operations - Dr. BONACORSI, Daniel: CMS computing model – tiered approach. Review of tier responsibilities. INFN T1: CNAF. Review: 375 TB disk online, 134 TB tape. Etc. What happens at T1s: Grid production. Analysis. Data distribution to elsewhere. Understanding and using Phedex. Multi-tier inbound . outbound rates. Understanding role of T2s in CMs computing model: too much detail. INFN tiers are being operated in the WLCG. Much involvement. Good experience from SC3 exercise. Good T1 transfer rates, then focus on job submission. Rate plots. Understanding the system. Various T2: understanding the system and the components. Good rates. Stability issues. SC3 was good test of T2 interaction with T1s. Good way of building up tier's know-how. T1: hybridism in infrastructure. Supporting lots of experiments, lots of tier-2s. Good to gain the experience. Need metrics for success: continuity (uptime), efficiency, robustness, experience sharing WLCG ready for CMS. INFN T1 centre fast gaining experience. T2s gaining experience too. -> experience shows that the tiers can 'keep the pace' Focus is on robustness / automation. Do the same with less effort.

Wednesday

DIRAC, the LHCb Data Production and Distributed Analysis system - Dr. TSAREGORODTSEV, Andrei: Description of services, resources and agents. Users only see services. Much of the DIRAC agents are deployed on VO boxes. Example of DIRAC 'config service' to configure all DIRAC services. Workload management is running in PULL mode from central DIRAC task queue. Cataloguing uses LFC in global catalog mode (with redundancy). Agents run in user space. Example pilot agents who are submitted to worker nodes – these check that the environment is OK and then pull a real user job from the central task-queue. This has advantages that the user's jobs are isolated from submission and environment failures since the job is never pulled if there is a problem.

Geant4 simulation in a distributed computing environment - Dr. PIA, Maria Grazi: Migration GEANT4 MC production from local batch system to grid experiences. Try to maintain transparency for user. Running on grid – there were problems – these can be hard to track down since the conditions are not reproducible. Definitely worth the effort. Gain more than the pain.

GANGA - A GRID User Interface - HARRISON, Karl: User analysis framework. Scripting and GUI submission and control mechanisms. Good uptake.

BaBar simulation production - changes in CM2 - Dr. SMITH, Douglas: Babar data model. Moving from database (Objectivity) data model to ROOT based model. DB problems: harder to manage. Need to be DBA to manage the files. Much work: many more files – need some development yto framework to handle all the files now. This was an effort. Comparison: positive – it was worth the effort – the data model scales much better now and is much easier to manage.

Long-term Experience with Grid-based Monte Carlo Mass Production for the ZEUS Experiment - Dr. STADIE, Hartmut: See H1 talk on Tuesday – same idea. HERA luminosity upgrade -> greater demand on computing at DESY→ move to grid solutions for monte carlo production. Took some effort – had to improve robustness of production system. Works well now and runs with minimal effort (< 1 FTE).

LCG 3D project status and production plans - Dr. DUELLMANN, Dirk Duellmann: Architecture: streams based replication from T0 master to T1 sites. Frontier caching (servlet with hsquid web-caches) to cache specific query resultsets at the T2. Plan defined. Infrastructure ready at T0 and T1 for testing. Experiments need to start testing.

ATLAS Tier-0 Scaling Test – Branco, Miguel: Arch of T0 described – what is being tested on Castor. Two parts to test.
Internal part

1. Writing data to tape cache: RAW from pit, reconstructed from recon farm.
2. Reading data from tape cache – RAW + recon by the migrator, RAW to send to recon farm
3. Writing RAW+recon to the tape itself.

Export to T1
Internal part – reached nominal rates on first try p- this was a great success.
External – several problems, config, etc – limited monitoring. Took a fair effort.
SC rerun succeeded at rates. Good experience – many problems shaken out and fixed.

Distributed Data Management in CMS - Peter Elmer: CMS computing model assumes that data is produced at the rate of 225 MB/s. This goes to worker nodes where reconstruction data is added. 280 MB/s goes to tape at Tier 0 and to Tier 1s. Data management is supposed to provide tools to discover, access and transfer event data. General principles:
- optimisation for read access
- optimisation for large bulk access
- minimisation of dependencies of jobs on the worker node
- site-local configuration should remain local

7PB of data in 10^6 files is going to be produced per year. Data processing workflow includes:
- Dataset Book-keeping System (DBS) - what data we have
- Data Location Service (DLS) - where is the data
- Site Local Catalogues - physical location of files at the site
- Access to the files through storage systems (dCache, CASTOR).

Output produced is registered in DBS and DLS
DBS:

- data definition
- data discovery
- used by distributed analysis tool (CRAB) and Monte Carlo Production systems
- scopes: one dbs instance describing data CMS-wide (global scope), replicated to instances with a more "local" scope.
- technologies used for DBS include CERN Oracle RAC for CMS, CGI "pseudo" server and Client CGI API (business logic + web service + client web service API).

DLS

- locate replicas of data in the distributed computing system
- map file-blocks to storage elements (SE's) where they are located
- very generic (not CMS specific)
- currently CMS prototype is used (python service with MySQL backend and client tools, no authentication/authorization mechanisms)
- LCG LFS is evaluated.

Local data access: DLS has names of sites hosting the data but not the physical location of files at the sites. Local File Catalogues provide site local information. DLS exists today as a CMS prototype (placeholder for evaluation of Grid catalogues using a python service with a MySQL backend + client tools, no authentication/authorization mechanisms. There is an evaluation of LCG LFC. Local file catalogues in XML, MySQL, trivial. Storage System can be dCache, or Castor

The CMS Computing Model - Dr. HERNANDEZ, Jose: Tiered architecture. Presented tiered data flows. Principles:: keep it simple! Optimize for common case. Explicit data placement – move jobs to data, data does not move to jobs. Summary: distributed computing model based on experience with previous one

Thursday

PhEDEx high-throughput data transfer management system - REHN, Jen: Traditional approach manual – took lots of effort. Phedex: what it does, transfer volumes. It's a distributed agent application. Scalability tests – tested without the actual transfers running to stress underlying phedex system – it can handle 50k files per hour – this is easily enough for CMS needs. srm-cp results: 50 % of transfers done with srm-cp fail on first try – various reasons. 10% “unknown”. They were all complete on retry eventually. Plans: improve dataset subscription mechanism. Easier agent management. Decentralise transfer database? Currently transfers have no prioritization.

Belle Monte Carlo Production on the Australian National Grid - LA ROSA, Marco: Belle monte carlo production. Description of the Australian national grid programme. New programme. Very heterogeneous environments – they're not really set up for grids, and the sites say what the specific access mechanism will be. Each site uses a different mechanism for getting the jobs to run and getting the data to and from the jobs. Heavy use of Xen virtual machines to hide to heterogeneity of the underlying sites from the jobs.

Software Components and Libraries

The “Software Components and Libraries” parallel track in CHEP06 addressed the areas of

- data management,
- reflection in C++,
- mathematical, statistical and analysis libraries,
- detector geometry representation,
- graphics and visualization,
- Monte Carlo event generators

As we are approaching the startup of the LHC it has been no surprise that most of the talks where related to the LHC computing. The track actually started with an overview of the LCG Applications Area (AA) software projects (#258). The goal of the AA is to provide common software solutions to the LHC experiments, where “common software” is defined when at least two experiments use or have expressed their intent to use it. The software is delivered as a result of coordinated use of resources from the experiments, the PH and IT departments at CERN, thus minimizing duplication of effort. Special attention is being paid on managing the configuration of the third-party software which the various AA deliverables are based on. There is also a strong focus on integration elements such as reflection mechanisms, plug-in management and scripting.

The AA projects are SPI (development infrastructure and configuration management services), ROOT (the software foundation of the LCG applications), POOL (the persistency framework) and SIMU (the simulation project).

Data Management

The Data Management session opened with a historical review of the use of databases in HEP (#12). While in the 90s the community had been largely convinced that Object Databases meet the requirements of the LHC computing, the commercial growth of the ODBMS technology did not eventually happen at the rate it had been anticipated. Therefore, despite the initial success of the use of commercial ODBMS solutions in HEP, the community eventually turned to home-made solutions for storing the experiment data. At this point in time the overall baseline followed by most of the HEP experiments can be summarized as follows:

“We stream the event, reconstruction and analysis data into ROOT files, we store the book-keeping, configuration, conditions and event meta-data in home-built systems based on top of commercial or open-source relational databases”.

The ROOT I/O system (#114) is used by most of the HEP experiments either directly or indirectly through the POOL framework for storing C++ objects in files. The major development efforts over the last couple of years focused on the optimized support of STL containers, data compression and thread safety. The TTree mechanism has been extended to allow for chaining, bitmap indexing and the storage of persistent object references either native (TRef) of external (pool::Ref).

The increasing use of RDBMS in HEP and in LHC computing in particular has been the driving force for the development of CORAL (#329), the relational software domain of POOL. CORAL provides a C++ API for accessing relational databases without the need of issuing SQL. Its components serve multiple RDBMS technologies (Oracle, MySQL, SQLite, Frontier) as well as functionality related to the distributed deployment of relational databases such as secure authentication, client-side monitoring, service indirection, connection pooling and fail-over mechanisms, the latter being a direct contribution from ATLAS (#32). Among the major advantages of the use of CORAL is the enforcement of “best practices” that are built-in the system.

CORAL has been eventually the basis of the implementation of various experiment-specific applications, such as the ATLAS event-level meta-data system which is using the POOL Collections (#81), as well as common applications such as the POOL Relational Storage Manager (#330), and COOL (#337), the LCG conditions database project.

The COOL project is already in its second year of development and its focus is currently shifting towards deployment. Its 3-dimensional meta-data model (item/time/version) originates from the early implementations of the conditions database before the COOL project was launched. Among the advantages of COOL is its flexibility in the definition of the data payload which may reside in the same database as the validity intervals or outside, in a different medium, addressed using persistent object references. COOL has already been integrated within the ATLAS and LHCb software frameworks. Its flexibility allowed its introduction into the LHCb High Level Trigger system where a specialized version of an LHCb reconstruction/analysis program is used (#168).

The software team of the BaBar experiment pioneered the concept of the Conditions Database as it is currently understood in HEP. Their early implementation was based on Objectivity/DB, but they are moving towards a technology independent model (#352). In particular, MySQL and ROOT are the envisaged technologies for the master replicas of the conditions. ROOT files will be used as the read-only distributed replicas.

Reflection in C++

One of the central software components in the LCG software is the C++ Reflection package (#185). It provides the introspection capability that is missing from the language itself (contrary to languages with built-in introspection capabilities, such as Java), which is absolutely required for persistency and interactivity purposes. The package has been recently integrated into ROOT and can be delivered both as part of ROOT itself, or as a standalone deliverable. The generation of the necessary “object dictionaries” is based on the gcc-xml tool.

JIL (#397) has been developed for the needs of the GlueX experiment and provides introspection for persistency purposes. The reflection information is generated from the compiler-generated debug information into XML, which in turn is used for the automatic generation of object streamers.

Mathematical, Statistical and Analysis Libraries

Among the foundations of any analysis task is the set of the mathematical and statistical libraries used. The mathematical libraries introduced recently in ROOT (#227) attempt to integrate in a set of standalone libraries of increasing functional complexity all the necessary mathematical tools and algorithms that are used by the various physics computations. Special attention is given to performance optimization with classes tailored to the needs of specific use cases. Well-established and successful algorithms from GSL or HEP-specific software components have been either imported or wrapped. The fitting libraries have been extended to include the C++ version of the MINUIT package.

Useful algorithms that are or can be used in HEP applications have also been developed in completely different kind of research fields such as operations research (#198). An example is the several pattern recognition tasks that can be accomplished with the use of the various linear and mixed-integer programming packages that are available either commercially or a open-source.

The problem of pattern recognition is also addressed by the StatPatternRecognition package (#208) developed for the needs of the data analysis of the BaBar experiment. It is based on Boost Trees which have demonstrated a superior performance with respect to other methods such as neural networking.

The PAX toolkit (#367) attempts to assist a physics analysis tasks by providing a framework for comparison all the various physics senarios that are compatible with the reconstructed set of tracks and vertices.

The Statistical Toolkit (#305) provides an extensive set of classes for performing Goodness-of-Fit tests. Different use cases require the use of different tests and therefore much attention is put on documenting the guidelines for the choice of the test to be used. The toolkit is now regularly used in the on-line environment of some experiments.

The Phystat repository (#465) has been set up by the FNAL Computing Division in order to accommodate the existing plethora of mathematical, statistical and analysis algorithms used in Physics in a single centralized archive. The responsibility for maintaining the code and documentation is on the submitters and the added value comes from the users submitting feedback.

Detector Geometry Description

Every detector simulation framework in HEP has a geometry definition component. The one used by ROOT (#383) is based on a model of volumes with a hierarchical structure. This allows the assembling of complex structures where global and local mis-alignments can be accommodated. Special attention has been given to the minimization of the memory consumption when building a geometry. A major advantage of the ROOT geometry modeler is that it can be used in both simulation and reconstruction programs.

Detector geometry, in order to be easily exchanged across applications, has to be specifiable in a format which is neutral to the programming language or application which is used. This need is addressed by two projects, where XML is used as the source of information. The first one is GDML (#259), where XML files can be generated automatically or written and edited by hand. There are converters for Geant4 and ROOT geometry modelers.

For the needs of the STAR experiment, the above concept has been extended (#105) to benefit from the programming capabilities that an XML parser offers when processing a file. It is based on the ATLAS Generic Detector Description.

Graphics and Visualization

Given that ROOT is currently the main player in the domain of interactive analysis, its 3D graphics component (#93) is an important one for the community. It is based on OpenGL for displaying objects, profiting from its rendering and “camera” capabilities. It has demonstrated good performance especially because it can take advantage of the hardware acceleration. The next major development step is the support for animations, which will allow, for example, following of tracks or showers in event displays.

ROOT has been the implementation basis of GLED (#381), a framework for distributed computing and event visualization, which is in turn is the basis of the ALICE event visualization environment (#380).

In STAR the ROOT framework has been complemented with Qt in order to create a framework for supporting complex interactive analysis sessions (#158). The framework is used in both on-line and off-line applications and most from its components are not experiment-specific.

The event visualization program of ATLAS (#69) is based on OpenInventor. It is extensively used for debugging purposes and it is going to be used as the event-display program at ATLAS Point 1.

In ATLAS an analysis framework has been developed based on the Java Analysis Studio with the aim of minimizing resources at the user's workstation, while at the same time providing a user with full analysis and interactivity capabilities (#332). This has been achieved using a 2-Tier architecture, where on the client side runs a JAS plugin while on the server side the actual off-line framework.

Monte Carlo Event Generators

The parallel track ended with a short sessions on Monte Carlo event generators. There has been a presentation of GENSER (#432), a project of the LCG AA. Its role is the development, validation, maintenance and documentation of the generators and associated services that are used by the LHC experiments, in a centralized manner.

The CEDAR Collaboration maintains HZTool (#122), a FORTRAN-based library of routines which enable the comparison of experimental data with predictions from the Monte Carlo generators. It is going to be replaced by a C++ based framework, which will make use of standard interfaces, such as AIDA and HepMC.

Finally, there has been a presentation of Eclipse (#222) as an environment to a physicist's development work. Its main advantages is that it supports multiple programming languages and that can be fully customized and easily interfaced with external tools. It has and impressive IDE and lots of potential uses, which come though at the expense of a large computing resource consumption.

The following sessions are reported on in more detail by various correspondents.

Evolution of Databases in Physics – Jamie Shiers: a review of the topic from the use of RZ and ZEBRA in the 80s, the arrival of OO in the early 90s, leading to LHC++ and the introduction of object databases such as O2 and Objectivity, the latter being chosen for adoption at a number of sites, principally SLAC and CERN. But the object DB market did not take off and gradually sites moved back to the mainstream databases such as Oracle.

COOL – Andrea Vassali: COOL is the LCG Conditions Database, storing non-event detector data which varies with time and which may come from online or offline sources. The data may also exist in different versions. COOL grew out of similar activities which were presented at CHEP04 by a number of experiments. The teams got together after that meeting and developed a common solution – COOL - merging the best ideas from the different bases. COOL is based on a meta-data model with 3 dimensions, including version. The first production release was in April 2005 and the latest release, last month, contains many new features, including some added after considerable performance testing. COOL is fully deployed at ATLAS and starting deployment at LHCb.

Java Analysis Studio as an interface to the Atlas Offline Framework - Julius Hrivnac
Atlas Analysis software is written in C++ with Python API (Python XML-RPC server). Java Analysis Studio (JAS) connects to Python using XML-RPC; results are sent using XML (defined with XMLSchema). JAS is a desktop analysis application based on FreeHEP. Basic scenario: you can run both remote and local python scripts from JAS; you can mix remote and local python. JAXB model is used to create java objects based on XMLSchema. Advantages:

- Light local client, 3 minutes to install
- fully interactive GUI
- scripting and API in several languages
- easily extensible by plugins.

Problems:

- python wrapper over c++ is incomplete
- undocumented
- unstable but/because improving
- c++ is still there

SSH tunnel is used for security.problems of isolation of clients, there's the risk that one client can take most of the resources and block other, lighter users. But this is a general problem of Athena rather than the Python XML-RPC server.

Computing Facilities and Networking

DNS load balancing and failover mechanisms at CERN – Vlado Bahyl: Load-balanced machines are watched by the observer using SNMP and this method supports many protocols. An interesting note on HTTP: DNS load balancing does not provide sticky sessions, so the stateful applications should always stay with the same host. We have to rely on the client for that. However, in apache httpd the "ServerName" directive can be used. In discussions between Vlado and Michal Kwiatek, Vlado suggested that DNS could be used for load balancing in the J2EE Public Service. Indeed, they should investigate this option, as it would be very straightforward to implement. Vlado mentioned that DNS load-balancing works well for HTTP and that there are no issues with browsers caching DNS replies.

Quantifying the Digital Divide - Les Cottrell, **SLAC**: :A scientific overview of the connectivity of South Asian and African Countries. One of the measures checked was minimum RTT (round trip) from US (best case scenario). They were checking both from US and from within the countries that were monitored. Apparently connectivity is improving by 40% per year in Europe. Russia is 6 years behind Europe. South Asia and Africa are 10 years behind Europe and not catching up. Universities in Asia in Africa have worse bandwidth than households in US and Europe (1 Mbit). In Pakistan, outages show fragility of the infrastructure (power outages among other problems) India is still behind although better than Pakistan and Africa but there were no concrete numbers (few sites monitored). Africa and South Asia still have to rely on satellites.

Storage for the LHC Experiments - Roger Jones, Lancaster University: It was agreed at HEPiX in Kalrsruhe (spring 2005) that a Storage Task Force should be formed. From CERN Bernd Panzer-Steindel and Helge Meinhard participated. This Task Force created an interim report; a final report will be presented at HEPiX in CASPUR, Rome in April. The interim report contains a a thorough discussion of disk issues (various types of disks have been analysed including disks optimised for large transactions and others optimised for many short I/Os) and less complete discussion of archival media. As an important conclusion, this report provides guidelines for large procurements of hardware. Other Conclusions :-

- better information exchange is needed
- testing should precede large hardware procurement
- while smaller site cannot always negotiate borrowing of hardware for test, CERN certainly can; the knowledge gained in this process should be passed on to the other sites
- RAID 5 is considered a good option; RAID 6 is expected to provide higher security of the data
- SAN was described as "flexible, but expensive"
- use of NAS/DAS was recommended with the best value/price ratio (management issues regarding SAN were not mentioned)
- some comparison of prices (in euros per effective 10 TB)
  - NAS/DAS 13500-27800
  - SAN/S 22000-26000
  - SAN/F 55000
- disk price should reach the same level as tape space by 2010
- rumours that the exponential growth of disk space per $ will stop by 2010 were noted in the report, but not confirmed

Openlab-II: where are we, where are we going? - Francois Fluckiger: openlab is a framework for collaboration with industry on R&D. Partner commitment from commercial companies is 3 years of contribution of hardware, software, services or money. In openlab-I, which finished at the end of 2005, 5 partners (IBM, Intel, HP, Enterasys and Oracle) contributed the total amount of $1.5M. The five companies did not compete; openlab is not a collection of disjoint projects. The companies actually cooperated to achieve common technical objectives: HP and Intel contributed 64 bit Itanium 2 processors

and later Xeon processors, IBM contributed a SAN FS Storage Tank, Oracle 10g databases were used, Enterasys and Intel contributed 10Gb ethernet and Voltaire (the only conributor, which meant 1-year commitment) contributed Infiniband connections. The main project of openlab-I was to enable LCG software on 64 bit architectures. Indeed, the results were important for SC-3. The results are published. Openlab-I also invited summer students and also helped educational initiatives such as GridCafe.
The lesson learnt from Openlab-I was twofold :-

- it's difficult to choose appropriate technologies to be evaluated, as the companies tend to push for specific technologies for marketing reasons
- it's difficult to avoid the expectations in companies that openlab will enable them to impose their technologies at CERN.

Two partners have already signed for openlab-II. Discussions with other partners are under way. Two major projects are going to be started:
1. Platform Competence Centre: PC virtualisation and enabling of 64 bit computing.
2. Grid Interoperability Centre: collaboration with EGEE II and involving partners in Grid Middleware.
A smaller project on Computer Security has also been briefly mentioned.

Network Information and Monitoring Infrastructure (NIMI) - Igor Mandrichenko, FNAL: Fermilab uses NIMI for semi-real time monitoring of their network. Thanks to constant checks, devices that are connected to the network are recognised within 15 minutes. It means that NIMI provides fully detailed information about all devices connected to the network 15 minutes ago. This data is stored in a PostgreSQl database (historical data since 2004 takes up 5GB on disk) and PosgreSQL backup solutions are used to backup the data to disk and CD backup server. Security checks are performed on detected machines; those considered insecure are automatically cut off (it has been mentioned that even 300 machines could be disconnected to protect on-site security). Database downtime would not mean network downtime, but would mean considerable threat to the security of the site, so has to be avoided. The team has been very happy of the quality of the database services based on PostgreSQL. In fact, they successfully use it even for more critical services such as backup service.

Benchmarking AMD64 and EMT64 - Hans Wenzel: Three families of processors, several processors of each, have been tested: Opteron, Athlon and Xeon. Both computational power and power consumption have been compared. Tests were done in three modes:

- 32/32: 32-bit legacy mode; machines installed with 32-bit OS, memory limited to 4GB
- 32/64: 32-bit compatibility mode: machines installed with 64-bit OS, but software run in 32-bit compatibility mode; memory per process is limited to 4GB, but the total memory limit is lifted.
- 64/64: full 64-bit mode.

Test machines were installed with Scientific Linux 32-bit or 64-bit. Dual-core processors generally promise double the computing power with fewer console and network connections. This test was meant to determine if this is really the case.The following software was used for the tests:
- OSCAR (only 32 bit)
- ORCA - CMS reconstruction program (only 32 bit)
- ROOT (both 32 and 64 bit)
- PYTHIA (both 32 and 64 bit)
Results for OSCAR showed full linear scalability of the software when number of processes was increased (up to the total number of cores). Performance was 4% worse in 32/64 mode as compared to 32/32 mode. Results for ORCA showed no linear scalability; apparently ORCA is IO intensive. Performance was 15% worse in 32/64 mode as compared to 32/32 mode. Results for ROOT showed that Opterons run 32% faster in 64/64 as compared to 32/32. Results for Pythia showed 20% boost when application was run in full 64-bit mode, as compared to 32/64. Also full linear scalability was observed when Pythia was run in parallel on dual core, 4 cpu machine. Power consumption benchmark, although difficult to set up, have been performed. Dual-core processors consumed only a fraction more than single-core processors, which is a

very good and promising result. Conclusion: legacy 32-bit applications can be safely run on 64-bit machines.

Introduction of a Content Management System in a HEP environment - Carsten GERMER

A Web Office has been working in DESY since 2003. It's a project between IT and PR staffed by

1 Concept/Programmer
1 Concept/Editorial
1 Programmer
1 Technician for hardware and system (joined later)

External funding has been assigned for consulting (technical and graphics). Handling graphics in-house proved impractical and had to outsourced. Started with dozens of websites in different technologies. Incorporation of existing web functionality was a requirement. Software chosen:

- Python application server "Zope"
- Contenet Management System "ZMS" - designed for scientific institutions, runs on Zope, is fully open source and has a large community.

Phase 1 (almost 2 years):
- single machine (with failover)
- Sites built with common functionality, but no centrally managed objects

Lessons learnt:
- web sites soon started to be used in production (hard to apply patches),
- more technical support was needed (hardware and system technitian joined the team),
- helpdesk support was needed for end-users.

Phase 2
- 6-node machines
- Central templates, if changed, changes visible on all the website
- Extra functionality through special objects, for example:
  - user and groups from DESY-Registry
  - Plugin for MS-Exchange calendars
  - browsing AFS
  - embedding content from other web sites (on-line)

Networks for ATLAS trigger and data acquisition - Stefan Stancu, CERN Three networks will be needed by ATLAS:
- control network: no special throughput requirements, 1Gb is enough
- 2 DAQ networks:
  (a) 100 Gb/s throughput with low latency and zero data loss (connection-less protocols are used)
  (b) 50 Gb/s throughput to be used with TCP.

Copper Ethernet 1G and 10G will be used based on devices available from multiple vendors. To achieve 100Gb throughput more than 10 links of 10Gb ethernet have to be provided. Moreover, networks are constructed with alternative paths. This needs to be correctly used by protocols. If a device breaks down, all network is still available, but throughput drops. If you send management traffic (which is normally based on connection-less protocols) together with your normal traffic, your management traffic will be penalised when throughput drops due to hardware failure. This is why separate management networks will be used by ATLAS. Tools to crosscheck and populate installation database are currently being developed.

Grid Middleware and e-Infrastructure Operation

The gListe File Transfer Service: Middleware Lessons Learned from the Service Challenges - Paolo Bodino: FTS is built around the concept of channel, which represents a point-to-point link between two sites. A channel has the following properties that can be set up using channel management interface:

- state (active, inactive, drain, stopped or halted)
- number of concurrent files transfers
- number of TCM streams
- VO share (used not to starve a VO; if there is only one VO, then it gets the full link).

Job represents the transfer request and is identified by a GUID. Supported transfer types are:

- SRM Get & Put + 3rd party GridFTP (only this one was used for SC3)
- SRM Copy

Support for Castor2, dCache and DPM have also been mentioned. Users can have one of the following roles:

- Submitter (can submit jobs)
- Administrator (full privileges)
- Vetoed User (has been revoked any privileges)
- Channel Manager
- VO Manager

FTS currently provides 3 interfaces:

- user interface to submit jobs
- administration
- monitoring

Simple retry logic has been used in SC3 (each problematic request was retried up to a predefined number of times). A more complex strategy is available in gLite 1.5; it evaluates transfer attempts history. The goals of the rerun of SC3 have been fully achieved. Goals for SC4 include

- run the service with least possible administration effort
- new functionality is going to be added (SRM2, integration with experiments)

A scalable Distributed Data Management System for ATLAS - David Cameron: ATLAS data model is the following: raw data goes to CERN Computer Centre. Then, reconstructed and raw data goes to tier 1 centres, where it is periodically reprocessed. From there data goes to tier 2s. Atlas uses 3 computing grids: LCG, OSG and NorduGrid, but they want to provide a single point of entry. The software that provides this single point of entry is called DonQuijote (DQ). Currently DQ2 is used and it's a scalable solution (DQ was not). It provides the concept of versioned datasets. A dataset is also the unit of data movement. 4 central catalogues used are used by ATLAS. They are physically split for partitioning because different usage patterns are expected. However, they are logically centralised. Central catalogue servers and clients are implemented in python; apache mod_python is used. MySQL databases are currently used at the backend. Oracle back end is going to be tested before SC4.

Software Tools and Information Systems

38 abstracts, 30 oral presentations, main themes:

- software management (release, distribution)
- collaboration tools & information management
- GUIs & frameworks
- Compilers, optimization, parallelization

Semantic Webs for HEP: need an ontology which is a way of describing terms & relationships**;** contains domain classes & objects, hierarchy, properties, ranges, logical relationships e.g. if a exists then there exists >0 of b. Focus on hep education, could be experimental ontologies. One example is protoge 2000

- a stanford tool
- a gui for knowledge management
- outputs rdf (resource description framework) schema and owl, a web ontology language

e.g. could model "gravity affects electrons", subclass of force relates to subclass of fermions. Need to work on how to use this semantically web rich domain knowledge for education. Another application is 'Frend of a friend' (FOAF). RDF for people relationships - rich information. Extract information from hepnames database - name, address, email, phone. People must add meta-data for themselves such as areas of interest; chemists and biomeds already doing this. For a mature technology - see foafamatic.

Use software snippets to improve application performance: Provide small code snippets to compiler writers with compiler options. Get good feedback compared to large complex projects**,** e.g. itanium + 3d rotation & transformation - few compilers get it right. **B**ad use of inlining**.** Agile use of snippets - quick & easy - test single compiler features. **B**enchmarking very important.

ROOT GUI builder: A powerful ROOT IDE exists with object browser and a session viewer for creating GUIs. Mdi(multiple document interface). Now many good widgets. There are plug-ins for editors. It is a cross-platform gui lib. Extendible architecture.

GUI Application Design: Building a GUI based on user goals, tasks and actions. Ideally a single button click to achieve important actions. Used ‘3 click rule’ as a firm objective**.** Performed a task analysis to map user requirements. Based on Use Case analysis**:** most frequent - toolbar - 1 click**.** For critical tasks have a backout option. Balance the paradox of complexity against easy to use and learn
Use progressive disclosure to hide complexity. Use of toolbars, tooltips, wizards, macros and sub-menus all good.

ROME – an analysis framework generator: Based on ROOT**.** Separates framework from analysis Generates the framework automatically. Experiment independent. See http://midas.psi.ch/rome.

The LCG SPI Project in LCG Phase II: An environment for LCG applications**.** Driven by the experiments in the form of the architects forum which channels requests. Increasing usage pattern: in Savannah 176 projects, 1500 users, 13500 bugs, 3k tasks.

- External software
- Testing frameworks
- Software distribution, build and release
- QA activities

Worm and P2P Distribution in ATLAS Trigger/DAQ: 6GB of software per release to 600 nodes in various locations. Distribution tools not always available outside CERN**.** Using the Nile worm to distribute tools and commands. Using BitTorrent to distribute large amounts of software.

ATLAS and CMS Release Processes: Many parallels**.** Solving the same problems Developers widely distributed, many lines of code. Using CVS, NICOS for nightly builds, tag collector, back-end database Some differences in tools enough to prevent reconciliation.

CMS software distribution on LCG and OSG Grids: Problematic**,** 2000 people in 37 countries across 160 institutes. Must distribute 72 RPMs; 1.5GB->4GB unpacked**.** Publish with glue schema**.** Need both OSG & LCG. Use XCMSI to monitor s/w installations**.** Clients sometimes missing crucial software such as PERL modules or even gcc. Use NICOS for nightly controlled builds**.** If root privileges are needed then cannot install over grid.

Packaging and Distributing CMS Software: Using SCRAM for software configuration, release & management. Using DAR (Distribution After Release) - robust, lightweight, good for runtime packages**.** Using XCMSI for large full development sets.

HyperNews: A HEP discussion tool providing faster feedback than web forums**.** Shared and archived, unlike email. Nicely structured tree of forum->discussion->thread->replies**.** Used extensively by BABAR experiment (~250K postings)**.** Allows very fast searching (0.1s)**.** Can easily download and use from http://hypernews.slac.stanford.edu. Increasingly used by ATLAS and CMS**.** Does not yet support RSS or VO authentication. Complementary to savannah.

Web Lecture Archive Project: Used for recording lectures**.** Driven by University of Michigan and CERN. Started in 1999**.** Captures seminars, slides & video streaming, all synchronized. www.wlap.org contains >700 web lectures. Can view as podcast via iTunes and see on IPods**.** Need to add metadata by hand to improve search. All based on w3c standards**.** Intended for HEP and education; commercial stuff out there e.g. at&t. wlcd lecture capture system has robotic speaker tracking**.** Intended to keep <10K $.

CERN equipment management integrates safety aspects – EDMS: Without asset management there is no LHC ! Goal: to track the whole lifecycle of equipment**.** From design->manufacture->installation->maintenance. ~450,000 pieces of equip**,** 100's users**.** ORACLE dependent; much written in PL/SQL.

CERN Document Service: orthogonal to arxiv.org**.** >2k publications per year**,** >10k conference contributions. CDSWare free from cdsware.cern.ch**.** Stores multimedia**.** HEP ontology would be useful**.** Written in 95%python, mysql, 5%lisp.

Report on RTAG-12: Collaborative tools need significant improvement at CERN**.** Final report was produced April 2005**.** The 9 recommendations need to be followed urgently**.** Collaborative action is needed.

Track Wrap-up
- A very interesting and diverse track
- Definite emphasis away from development and towards QA, performance and deployment
- Strong tendency to use python for a variety of tasks
- Widespread use of WIKIs
- Collaboration tools are becoming increasingly important
- Not much input in this track from LHCb and ALICE

Online Systems

Monitoring and Triggering an online trigger with PVSS – Eric van Herwijnen: rather than controlling hardware, Eric has used PVSS to control the LHCb online trigger farm. PVSS is also used to extract the counters and rates and then ROOT is used to display and manipulate the resulting histograms.. This is all wrapped up into the Gaucho programme. The actual monitoring sensors use DIM. There is then a PVSS backend implemented as a set of DIM clients subscribing to the counters and histograms produced by ROOT; the aggregation is published by PVSS where PVSS acts as a DIM server. Screen shots showed how the rates could be displayed by sub-farm, by node or by job. Eric reports good performance of PVSS in this environment although for scalability, he needs to move histogram manipulation outside ROOT. PVSS was chosen for this (unusual) task partly at least because of integration with the hardware control environment of LHCb.

ATLAS Trigger Monitoring – W.Vandelli. Need to monitor 140M channels of the various detectors! Will produce 10GB monitoring data per run. The framework was tested on 700 nodes of LXBATCH during 2005. Although the LXBATCH network interconnections are slower than the ATLAS online trigger farm will have, and this caused some saturation effects, the success of this test allowed them to proceed with their development with confidence.

ATLAS DAQ and High Level Tests on LXBATCH – Doris Burckhart: a description of the successful tests last year of ATLAS's event builder and HLT. A number of issues were uncovered, the most important of which is the need for fault tolerance, every time a node crashed, that trigger system crashed. The speaker acknowledged the support of IT and the LXBATCH support team.

CMS s and Data Quality Monitor – C.Leonidopoulos: 1000 node farm substituting for the traditional level 2 and level 3 trigger. 150KHz in, 150Hz out. They have built a scheme with a web interface so that anyone in CMS anywhere can check on data quality with a variety of tools.

Posters

Following on from the previous meeting, more emphasis had been put on posters and there were 2 healthy poster sessions in two distinct sessions with large numbers of posters in each..

Web Services with GridSite and C/C++/Scripts - Andrew McNab, University of Manchester
Michal was not able to catch the author by his poster, but he thought it looked interesting nonetheless: it's a web server architecture based on Apache Web Server that enables X.509 authentication through use of mod_ssl and mod_gridsite, which adds support for GSI proxies and extraction of any VOMS attributes by intercepting the underlying OpenSSL callbacks. Mod_gridsite is loaded into Apache at starup and accesses to data structures of all other Apache components. In this particular scenario it pipelines requests to mod_cgi. It could be checked if the same approach would work to enable Apache Web Server and mod_jk to be run in front of java based grid applications (FTS, Fireman), which are currently run on tomcat and authenticate users through a tomcat extension.

Alan Silverman
8th March 2006