University of Glasgow | Experimental Particle Physics

GLAS-PPE/2009-02
19th March 2009

Department of Physics and Astronomy
Experimental Particle Physics Group
Kelvin Building, University of Glasgow,
Glasgow, G12 8QQ, Scotland
Telephone: +44 (0)141 330 2000 Fax: +44 (0)141 330 5881

GridPP - The UK Grid for Particle Physics

D. Britton[1], A.J. Cass[2], P.E.L. Clarke[3], J. Coles[4], D.J. Colling[5],
A.T. Doyle[1], N.I. Geddes[6], J.C. Gordon[6], R.W.L. Jones[7], D.P. Kelsey[6],
S.L. Lloyd[8], R.P. Middleton[6], G.N. Patrick[6*], R.A. Sansum[6] and S.E. Pearce[8]

[1]University of Glasgow, Glasgow G12 8QQ, UK,
[2]CERN, Geneva 23, Switzerland,
[3]National e-Science Centre, Edinburgh - EH8 9AA, UK,
[4]Cavendish Laboratory, University of Cambridge, Cambridge CB3 0HE, UK,
[5]Dept. Physics, Imperial College, London SW7 2AZ,
[6]STFC Rutherford Appleton Laboratory, Didcot OX11 0QX, UK,
[7]Dept. Physics, Lancaster University, Lancaster LA1 4YB, UK,
[8]Dept. Physics, Queen Mary, University of London, London E1 4NS, UK

Abstract

The startup of the Large Hadron Collider (LHC) at CERN, Geneva presents a huge challenge in processing and analysing the vast amounts of scientific data that will be produced. The architecture of the worldwide Grid that will handle 15PB of particle physics data annually from this machine is based on a hierarchical tiered structure. We describe the development of the UK component (GridPP) of this Grid from a prototype system to a full exploitation Grid for real data analysis. This includes the physical infrastructure, the deployment of middleware, operational experience and the initial exploitation by the major LHC experiments.

[0)] *Author for correspondence (glenn.patrick@stfc.ac.uk)

1 The Computing Challenge of the Large Hadron Collider

The Large Hadron Collider (Evans & Bryant 2008) will become the world's highest energy particle accelerator when it starts full operation. Protons, with energies of up to 7 TeV, will be collided at 40 MHz to recreate some of the conditions that prevailed in the Universe during the earliest moments of the "Big Bang". Positioned around the 27 km superconducting collider will be four major experiments - ALICE (Aamodt et al. 2008), ATLAS (Aad et al. 2008), CMS (Chatrchyan et al. 2008) and LHCb (Augusto Alves Jr et al. 2008) - which will record the particle interactions of interest. These four experiments contain a total of ∼150 million electronic sensors and the rate of data flowing from them will be about 700 MB/s, equivalent to 15 PB per year. The processing and analysis of these data will require an initial CPU capacity of 100,000 processors operating continuously over many years. This capacity will need to grow with the accumulated luminosity from the LHC and is expected to double by 2010 (Knobloch 2005).

Particle physicists have chosen Grid technology to meet this huge challenge with the computing and storage load distributed over a series of centres as part of the Worldwide LHC Computing Grid (WLCG).

2 Worldwide LHC Computing Grid - WLCG

Once the LHC is in continuous operation, the experimental data will need to be shared between 5000 scientists scattered around 500 institutes in the world. Copies of the data will also need to be kept for the lifetime of the LHC and beyond - a minimum of 20 years.

The Worldwide LHC Computing Grid (WLCG) currently consists of 250 computing centres based on a hierachical tiered structure (Shiers 2007). Data flows from the experiments to a single Tier 0 Centre at CERN, where a primary backup of the raw data is kept. After some initial processing, the data is then distributed over an optical private network, LHCOPN (Foster 2005), with 10 Gb/s links to eleven major Tier 1 centres around the world. Each Tier 1 centre is responsible for the full reconstruction, filtering and storage of the event data. These are large computer centres with 24x7 support. In each region, a series of Tier 2 centres then provide additional processing power for data analysis and Monte-Carlo simulations. Individual scientists will usually access facilities through Tier 3 computing resources consisting of local clusters in university departments or even their own desktops/laptops.

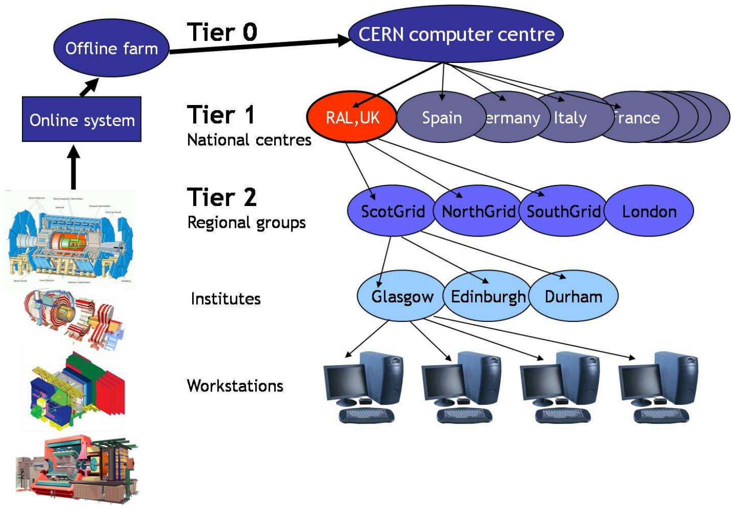

Figure 1: Hierarchical tier structure of the Worldwide LHC Grid

The WLCG is based on two major Grid infrastructures: the *Enabling Grids for E-Science* project or *EGEE* (Jones 2005) and the *US OpenScience Grid* (Pordes et al. 2008). The Scandinavian countries also contribute the NorduGrid infrastructure.

3 GridPP Architecture

The UK component of the Grid infrastructure for particle physics has been built by the GridPP collaboration through the joint efforts of 19 universities, the Rutherford Appleton Laboratory and CERN. The initial concept was first taken through a prototype stage in 2001-2004 (Faulkner et al. 2006) and this was then followed by building a production scale Grid during 2004-2008. With the start of LHC commissioning in 2008, the project has just entered its exploitation phase.

3.1 UK Tier 1 Centre

The UK Tier 1 Centre is located at Rutherford Appleton Laboratory (RAL in Figure 2) and the 2008 hardware configuration is based around a CPU cluster of 3,200 cores delivering ∼2,500 job slots and 340 disk servers providing 2.3 PB of disk storage. A Sun SL8500 tape robot provides 10,000 media slots, 18 T10K tape drives and a storage capacity of 5 PB.

After receiving a share of the raw data from CERN, the particle interactions (so-called "events") are reconstructed from the electronic information recorded in the various sub-detectors of each experiment. The extracted physics information is then used to select, filter and store events for intitial analysis. Datasets are then made available to the Tier 2 centres for specific analysis tasks.

A hierarchical storage management system is used for large-scale storage of data files at the Tier 1. This is based on the CASTOR 2 (CERN Advanced STORage manager) system (Presti et al. 2007). Each of the LHC experiments uses separate CASTOR instances to avoid resource conflicts and to minimise problems with data flow.

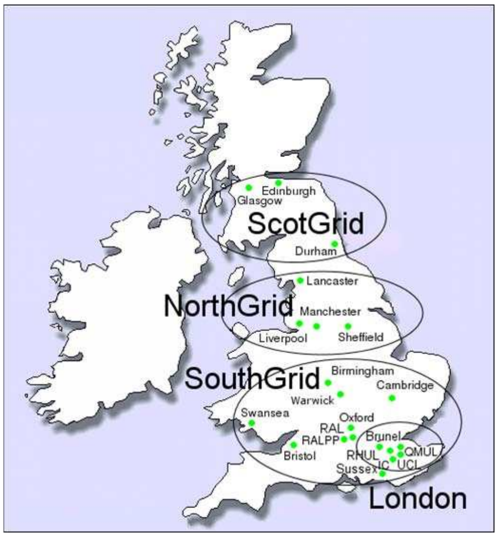

Figure 2: UK particle physics Grid centres (RAL marks the Tier 1 location)

3.2 UK Tier 2 Centres

GridPP has developed four regional Tier 2 Centres as shown in figure 2: LondonGrid, NorthGrid, ScotGrid and SouthGrid. These primarily focus on providing computing power for generating simulated Monte-Carlo data and on the analysis of data by individual physicists. Each Tier 2 Centre is a federation of computing facilities located in several institutes - for example NorthGrid consists of the universities of Lancaster, Liverpool, Manchester and Sheffield.

The distributed Tier 2 resources currently provide ∼10,000 job slots and approaching 2 PB of disk storage. The 17 individual sites vary in size from large centres such as Manchester providing 1,800 job slots (2,160 KSI2K) and 162 TB of disk, to small departmental clusters of a few machines.

4 Middleware

The middleware used in GridPP is the gLite distribution, currently version 3.1, (gLite 2008) from the EGEE project and adopted by WLCG across all of its sites. This consists of components from a number of different Grid projects.

4.1 Workload Management - WMS

A Grid job is specified using a Job Definition Language, based on the Condor ClassAd language, and this is submitted in a script together with the necessary program and input files through the Workload Management

System (Andreetto et al. 2008). A Resource Broker component accepts each job and matches it to a suitable site for execution. Other components transmit the job to the relevant site and finally manage any output. Sites accept jobs through a gatekeeper machine known as a Computing Element, which then schedules the job to run on a worker node within the cluster. Small output files are transmitted back through the WMS, while larger data files may be written to a Storage Element and catalogued.

4.2 Data Management

GridPP supports the small scale Disk Pool Manager and medium scale dCache storage solutions at the Tier 2 sites, in addition to the large-scale CASTOR system at the Tier 1 centre. The problem of file access across a heterogeneous Grid with multiple mass storage systems has been solved by the Storage Resource Management or SRM interface. This protocol is designed to provide access to large-scale storage systems on the Grid, allowing clients to retrieve and store files, control their lifetimes as well as reserving filespace for uploads, etc. The concepts of storage classes and space tokens were intoduced recently (Donno et al. 2008) to enable experiments to place data on different combinations of storage device and to dynamically manage storage space.

Data distribution between the Tier 0, Tier 1 and Tier 2 centres is performed using the File Transport Service (FTS). This uses unidirectional channels to provide point-to-point queues between sites with file transfers handled as batch jobs; providing prioritisation and retry mechanisms in the case of failure.

4.3 Distributed Databases and Catalogues

In addition to the interaction data (physics "events"), experiments also require a large amount of non-event data, describing the current running parameters and calibration constants associated with each sub-detector, to be stored and accessible all over the Grid. This time-varying data is essential for the reconstruction of physics events and is stored in a conditions database. Users and production programs also need the ability to locate data files (or their replicas) and this is achieved through the LCG File Catalogue (LFC). The LFC contains the mappings of logical file names to physical files, along with any associated replicas on the Grid.

These worldwide distributed databases have been set up by the LCG 3D (Distributed Deployment of Databases for LCG) project using Oracle Streams technology to replicate the databases to the external Tier 1 centres outside CERN (Duellmann 2006). At the UK Tier 1 centre, there are independent multi-node database clusters for the conditions databases of both the ATLAS experiment (Viegas et al. 2008) and the LHCb experiment (Clemencic 2008). This ensures high availability and allows for large transaction volumes. The LFC is also based on an Oracle Enterprise relational database, but the ATLAS and LHCb catalogues reside on multi-node clusters shared with other services. These are highly configurable allowing for all services to still run if a node fails.

5 User Access and Experiment Specific Software

Particle physicists access the Grid from a User Interface (UI) - a departmental or desktop computer with the user-level client tools installed. Authentication is based on digital X.509 certificates, which in the case of the UK are issued by the National Grid Service. Individual physicists belong to a Virtual Organisation (VO) representing their individual experiment and each computing site in the UK decides which VOs can use its facilities and the appropriate level of resource. A Virtual Organisation Membership Service (VOMS) provides authorisation information; specifically the roles and capabilities of a particular member of a VO. At the beginning of each session, a proxy certificate is obtained for a limited lifetime (typically 12 hours).

Experiments have also developed front-ends to simplify the definition, submission and management of jobs on the Grid. The Ganga interface (Maier 2008), a GridPP supported project, is the best known example and is used by the ATLAS and LHCb experiments. This allows jobs to be either run on a local batch system or the Grid and provides all of the facilities for job management; including submission, splitting, merging and output retrieval. Ganga can be used via three methods: an interactive interface, in a script or through a graphical interface. The adoption of Ganga enables a physicist to exploit the Grid with little technical knowledge of the underlying infrastructure. Over 1,000 unique users of Ganga have been recorded in 2008 as data analysis programs have been prepared for the switch-on of the LHC.

Similarly, experiments have written their own data management layers which sit on top of the standard Grid services. For example, the GridPP supported PhEDex (Physics Experiment Data Export) system of the CMS collaboration provides a data placement and file transfer system for the experiment (Tuura et al. 2008). This is based around transfer agents, management agents and a control database, providing a robust system to manage global data transfers. The agents communicate asynchronously and between centres the system can achieve disk-to-disk rates in excess of 500 Mbps and sustain tape-to-tape transfers over many weeks.

6 Grid Deployment, Operation and Services

In order to achieve a high quality of service across the wider Grid, the WLCG/EGEE infrastructure is divided into ten regions, each with a Regional Operations Centre (ROC) responsible for monitoring and solving operational problems at sites within its domain. GridPP sites are covered by the UK/Ireland ROC.

Computer hardware in the underlying clusters is usually monitored locally by open source programs such as Nagios and Ganglia. These can display variations in cluster load, storage consumption and network performance; raising alarms when thresholds are reached or components fail. The basic monitoring tool for the LCG Grid is the Service Availability Monitoring (SAM) system. This regularly submits Grid jobs to sites and connects with a number of sensors which probe sites and publishes the results to an Oracle database. At the regional level, problems are tracked by Global Grid User Support (GGUS) tickets issued to the sites affected or to the middleware/operations group.

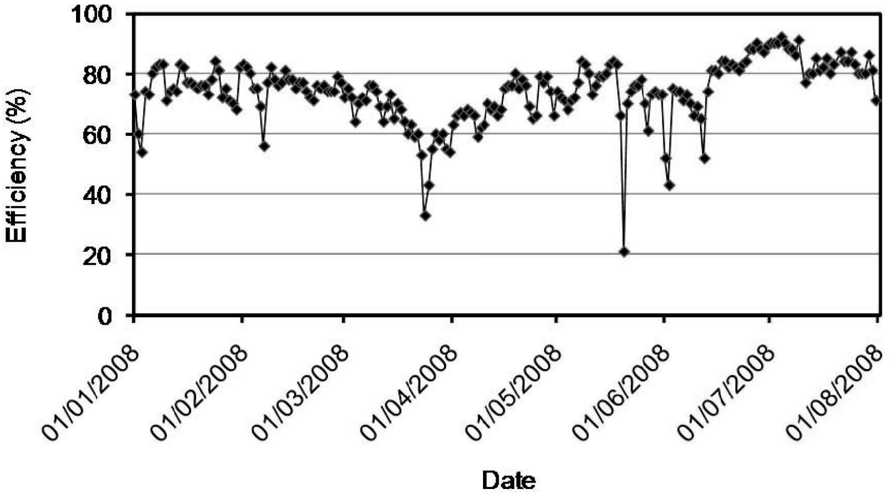

Figure 3: Percentage of successful ATLAS test jobs averaged over all GridPP sites

The LHC experiments also need to have a clear picture of their Grid activities (e.g. job successes/failures, data transfers, installed software releases) and so have developed their own SAM tests which probe the site infrastructures from their own specific standpoint. Dashboards have been developed to present this information in a concise and coherent way. In addition, GridPP has found it useful to also develop a suite of tests which regularly exercise the essential components in the experiment computing chains. For the ATLAS experiment, jobs are sent every few hours to each UK site on the Grid and perform an exemplar data analysis. As can be seen from figure 3, the overall efficiency (ignoring some specific incidents) across all GridPP sites has varied between 60% to 90% and the continuing challenge is to improve the robustness and reliability of the Grid for experiments.

7 Experiment Exploitation and Research Applications

Over several years, the LHC experiments have individually exploited the evolving Grid infrastructure through a series of "data challenges" and "service challenges". These have been based on large data samples, typically containing many millions of simulated events, which have been generated using the Grid and then also processed by the Grid using the same data processing chains prepared for real data.

In readiness for the switch-on of the LHC, all of the experiments simultaneously conducted a Common Computing Readiness Challenge (CCRC08) at the start of 2008, with phases in February and May. The number of batch jobs submitted each month to the UK Tier 1 Centre during the build-up to LHC commissioning is shown in figure 4(a), whilst the load of simultaneous running jobs on the Tier 1 during one week of CCRC08 is shown in figure 4(b).

In proton-proton mode at nominal luminosity, the four LHC experiments are expected to produce a total data rate of ∼1600 MB/s from the Tier 0 to the eleven external Tier 1 Centres. The UK share is estimated to be of the order 150 MB/s and from figure 5 it can be seen that this was achieved during the simultaneous challenges in February and May 2008.

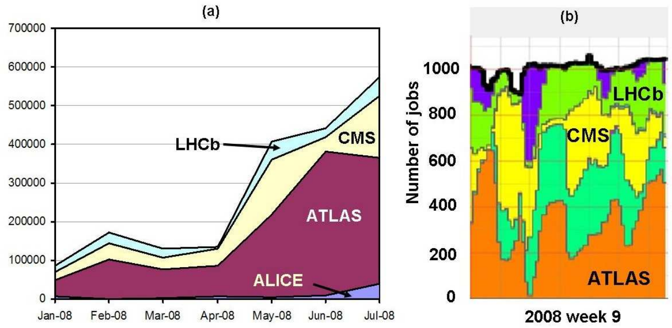

Figure 4: Batch jobs from LHC experiments at the UK Tier 1: (a) submitted each month in buildup to LHC operation, and (b) simultaneous load during one week in Feb 2008

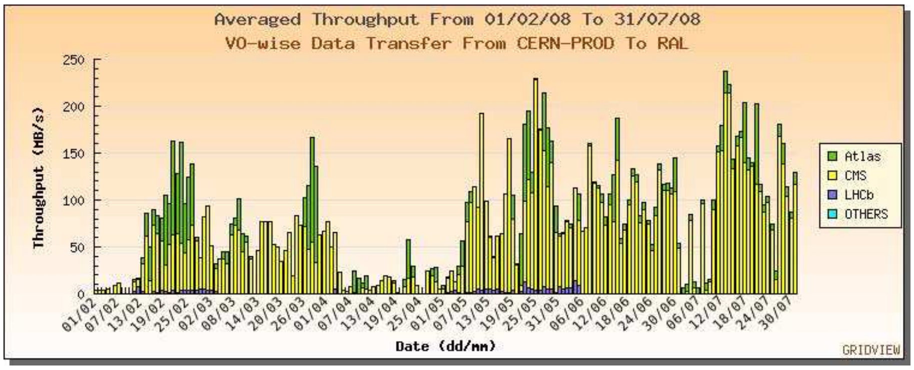

Figure 5: Averaged daily throughput from CERN to the UK Tier 1 Centre

Monte-Carlo techniques play a vital role in the modelling, analysis and understanding of particle physics data. As an example, we show the results of a very large number of simulations on the particle physics Grid. By simulating a statistical sample of events, equivalent to the size of the expected dataset, many thousands of times, we can study the precision with which a parameter will be measured when real data arrives. Typically, this is undertaken on the Grid by farming out many hundreds of simulation runs and then collecting the results to produce a set of statistical distributions or to show a trend. This is illustrated in figure 6, which shows the value of a systematic detector effect between various species of particle after they have been reconstructed in the LHCb experiment, together with a Monte-Carlo study of fitting a physics parameter, $\Delta A^{s,d}_{fs}$, which may be biassed by this effect. This parameter is important in studying the asymmetries between the decays of the neutral B meson (particle) and anti B meson (antiparticle), which ultimately should improve our explanations for the prevalence of matter in the Universe. To produce the plot in figure 6(b), each simulated model "experiment" required one million signal events to reach a statistical precision comparable to six months of data taking. Several models were investigated using different input configurations and each model required $\sim$500 "experiments" to be run - leading to a very large number of simultaneous jobs on the Grid consuming over 100,000 hours of CPU time.

Another example of the power of the Grid comes from the CMS experiment (CMS Collaboration 2008), which simulated, processed and analysed a sample of over 20 million physics events containing pairs of muons. The aim was to perform a measurement of the $\mu\mu$ mass spectrum and isolate any resonances after processing and filtering the data. A fit to the final subsample of events after track isolation is shown in figure 7, where the $Z^\circ \rightarrow \mu^+\mu^-$ decay can be seen prominently sitting on the background from quantum chromodynamics (QCD).

8 Outlook

A working Grid for particle physics has been established in the UK with the necessary resources for the early exploitation of the Large Hadron Collider. The LHC is poised to become the frontline facility for particle physics over the next decade and GridPP is a vital component of the worldwide infrastructure built to process and analyse the data recorded by the four main experiments.

The commissioning of the collider started with a series of injection tests in August and September 2008. On

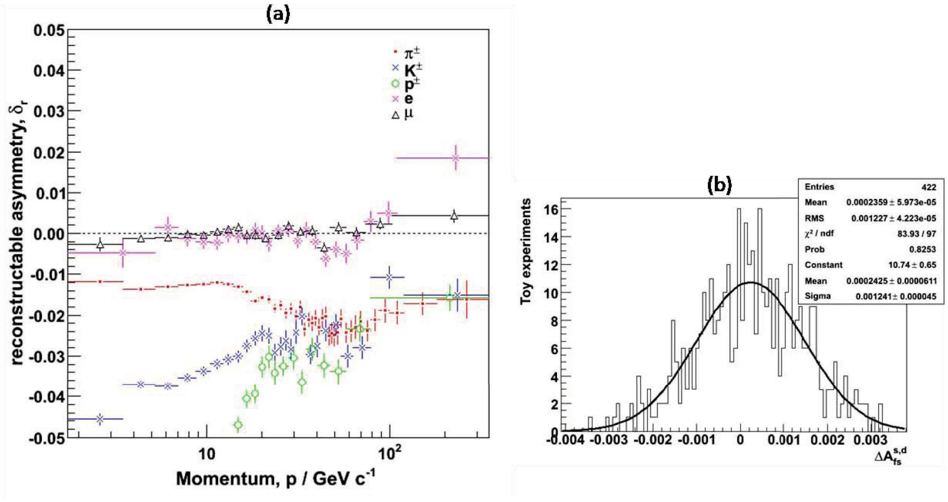

Figure 6: (a) Reconstruction asymmetry for different final state particles in the LHCb experiment versus momentum, and (b) example distribution of fitted values for the physics parameter $\Delta A_{fs}^{s,d}$ from a single set of 422 Monte-Carlo "experiments" (Lambert 2008)

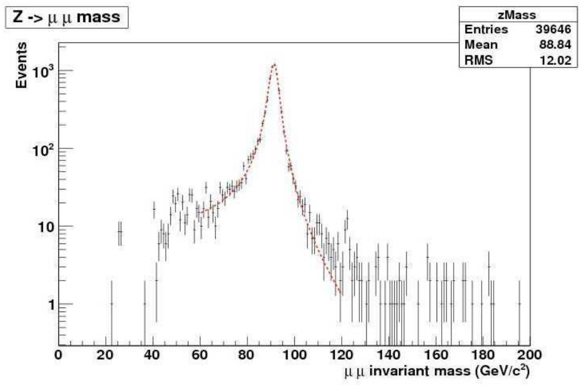

Figure 7: Fit to the Z, W and muon-enriched QCD backround from dimuon candidates, for an assumed luminosity of 25pb^{-1} in the CMS experiment (CMS Collaboration 2008)

10 September 2008 (the penultimate day of the AHM2008 conference) the first circulating proton beam through the entire 27 km of the LHC ring was achieved, closely followed by several hundred orbits. Particle interactions orginating from the beam were observed in all four experiments. A technical fault with the accelerator developed on 19 September and proton-proton collisions are now scheduled for summer 2009; GridPP will then be ready to record the first real LHC collision data following eight years of development.

Acknowledgments

We acknowledge financial support from the Science and Technology Facilities Council in the UK and from the EGGE collaboration. We wish to thank the many individuals in the UK, both in the Tier 1 Centre and the Tier 2 institutes, who have helped to build GridPP. This Grid project would also have not been possible without the major contributions from our WLCG and EGEE colleagues at CERN and around the world. The CMS collaboration and Rob Lambert are thanked for making available the results of their simulations.

References

[1] Aad, G., et al. 2008 The ATLAS Experiment at the CERN Large Hadron Collider. *JINST* **3**, S08003.

[2] Aamodt K., et al. 2008 The ALICE experiment at the CERN LHC. *JINST* **3**, S08002.

[3] Andreetto, P., et al. 2008 The gLite Workload Management System. *J. Phys.: Conf. Ser.* **119**, 062007. (doi:10.1088/1742-6596/119/6/062007)

[4] Augusto Alves Jr, A., et al. 2008 The LHCb Detector at the LHC. *JINST* **3**, S08005.

[5] Chatrchyan S., et al. 2008 The CMS Experiment at the CERN LHC. *JINST* **3** S08004.

[6] Clemencic, M. 2008 LHCb Distributed Conditions Database. *J. Phys.: Conf. Ser.* **119**, 072010. (doi:10.1088/1742-6596/119/7/072010)

[7] CMS Collaboration 2008 Towards a measurement of the inclusive $W \to \mu\nu$ and $Z^{\circ} \to \mu^{+}\mu^{-}$ cross sections in pp collisions at $\sqrt{s} = 14$ TeV. Report CMS-PAS-EWK-07-002, CERN.

[8] Donno, F., et al. 2008 Storage Resource Manager Version 2.2: design, implementation, and testing experience. *J. Phys.: Conf. Ser.* **119**, 062028. (doi:10.1088/1742-6596/119/6/062028). See also http://sdm.lbl.gov/srm-wg/doc/SRM.v2.2.html.

[9] Duellmann D., et al. LCG 3D Project Status and Production Plans. *Proc. of Computing in High Energy and Nuclear Physics*, Mumbai, Feb. 2006. See also http://lcg3d.cern.ch/.

[10] Evans, L. & Bryant, P. (eds) 2008 LHC Machine. *JINST* **3**, S08001.

[11] gLite 2008 gLite middleware, version 3.1 documentation. See http://glite.web.cern.ch/glite/documentation/R3.1/

[12] Faulkner, P.J.W., et al. 2006 GridPP: development of the UK computing Grid for particle physics. *J. Phys. G: Nucl. Part. Phys.* **32**, N1-N20. (doi:10.1088/0954-3899/32/1/N01)

[13] Foster, D. (ed.) 2005 LHC Tier-0 to Tier-1 High Level Network Architecture. See https://twiki.cern.ch/twiki/pub/LHCOPN/LHCopnArchitecture/LHCnetworkingv2.dgf.doc.

[14] Jones, B. 2005 An Overview of the EGEE Project. In *Peer-to-Peer, Grid, and Service-Orientation in Digital Library Architectures*, Lecture Notes in Computer Science **3664** pp 1-8, Berlin: Springer. (doi:10.1007/11549819_1)See also http://www.eu-egee.org/

[15] Knobloch, J. (ed.) 2005 LHC Computing Grid Technical Design Report, LCG-TDR-001 and CERN-LHCC-2005-024, CERN. See http://lcg.web.cern.ch/LCG/tdr.

[16] Lambert, R.W. 2008, LHCb Hybrid Photon Detectors and Sensitivity to Flavour Specific Asymmetry in Neutral B-Meson Mixing. PhD thesis, University of Edinburgh.

[17] Maier, A. 2008 Ganga - a job management and optimising tool. *J. Phys.: Conf. Ser.* **119**, 072021. (doi:10.1088/1742-6596/119/7/072021). See also http://ganga.web.cern.ch/ganga/.

[18] Pordes, R., et al. 2008 The Open Science Grid status and architecture. *J. Phys.: Conf. Ser.* **119**, 052028. (doi:10.1088/1742-6596/119/5/052028). See also http://www.opensciencegrid.org/

[19] Presti, G.L., et al. 2007 CASTOR: A Distributed Storage Resource Facility for High Performance Data Processing at CERN. *Proc. 24th IEEE Conf. on Mass Storage Systems and Technologies*, 275-280. (doi:10.1109/MSST.2007.4367985). See also http://castor.web.cern.ch/castor/.

[20] Shiers, J. 2007 The Worldwide LHC Computing Grid (worldwide LCG). *Comp. Phys. Commun.* **177**, 219-233. See also http://lcg.web.cern.ch/LCG/.

[21] Tuura L., et al. 2008 Scaling CMS data transfer system for LHC start-up. *J. Phys.: Conf. Ser.* **119**, 072030. (doi:10.1088/1742-6596/119/7/072030)

[22] Viegas, F., Hawkings, R. & Dimtrov, G. 2008 Relational databases for conditions data and event selection in ATLAS. *J. Phys.: Conf. Ser.* **119**, 042032. (doi:10.1088/1742-6596/119/4/042032)

List of Figure Captions

Figure 1: Hierarchical tier structure of the Worldwide LHC Grid

Figure 2: UK particle physics Grid centres (RAL marks the Tier 1 location)

Figure 3: Percentage of successful ATLAS test jobs averaged over all GridPP sites

Figure 4: Batch jobs from LHC experiments at the UK Tier 1: (a) submitted each month in buildup to LHC operation, and (b) simultaneous load during one week in Feb 2008

Figure 5: Averaged daily throughput from CERN to the UK Tier 1 Centre

Figure 6: (a) Reconstruction asymmetry for different final state particles in the LHCb experiment versus momentum, and (b) example distribution of fitted values for the physics parameter $\Delta A_{fs}^{s,d}$ from a single set of 422 Monte-Carlo "experiments" (Lambert 2008)

Figure 7: Fit to the Z, W and muon-enriched QCD backround from dimuon candidates, for an assumed luminosity of 25pb^{-1} in the CMS experiment (CMS Collaboration 2008)