GridPP - The UK Grid for Particle Physics

By 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]*Dept. Physics & Astronomy, 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*

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.

Keywords: grid middleware distributed computing data particle physics LHC

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 operation in autumn 2008. Protons, with energies of up to 7 TeV, will be collided at 40MHz 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

* Author for correspondence (g.n.patrick@rl.ac.uk)

the Worldwide LHC Computing Grid (WLCG). This is a hardware and software infrastructure that has to provide dependable, consistent, pervasive and inexpensive access to high-end computational capabilities.

2. Choice of Computing Model

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. Custodial copies of the data will also need to be kept for the lifetime of the LHC and beyond - a minimum of 20 years. A centralised approach of maintaining all of the computing capacity at a single location close to the experiments was not adopted because:

- The costs of maintaining and updating the resources are more easily shared in a distributed environment, where national organisations and funding bodies can provide local computing resources, whilst still contributing to the global goal.
- It is more easy to build redundancy and fault tolerance into a distributed system and minimise the risks from single points of failure that are more inherent in a centralised system. Multiple replicas of data can provide load balancing and the reassignment of resources in the case of failure of one component on the Grid.
- The spanning of time zones means that monitoring and support can be more readily provided. The LHC will operate around the clock for 8 months each year and the high availability of resources, both hardware and human, is essential.

Despite the above, the operation of a secure, large-scale, highly heterogeneous Grid with the complexities of worldwide data distribution and on-demand analysis, coupled with the continuous operation of the LHC, poses many challenges. This has meant that physicists and computer scientists have had to stage the development of software and infrastructure alongside a programme of large-scale simulation tests to build up the necessary operational and monitoring experience to feed back into the development of middleware.

3. Worldwide LHC Computing Grid - WLCG

The Worldwide LHC Computing Grid currently consists of 250 computing centres based on a hierachical tiered structure (Shiers 2007). Data flows from the experiments to the 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.

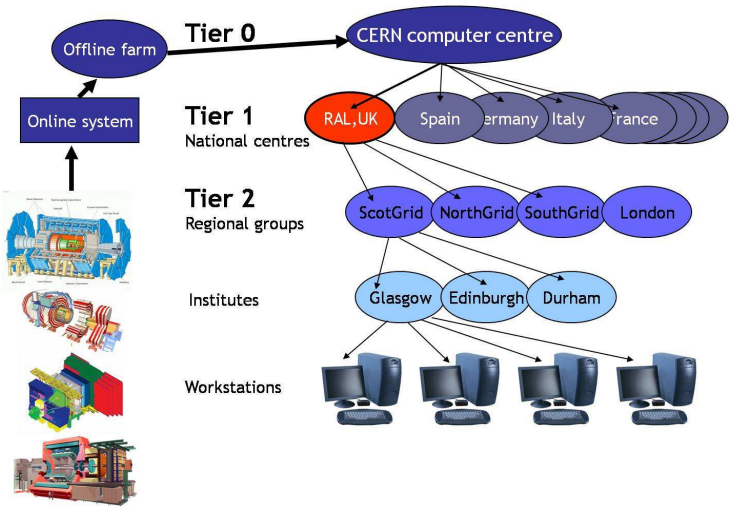

Figure 1. Hierarchical tier structure of the Worldwide LHC Grid

The LCG project has been through two phases. Phase I (2002-2005) had the objectives of building a service prototype based on existing Grid middleware, of running a production Grid service and of producing the Technical Design Report for the final system. Phase II (2006-2008) has focussed on building and commissioning the initial LHC computing environment. The LCG 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 based on the Advanced Resource Connector (ARC) middleware.

4. 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 the GridPP1 prototype stage in 2001-2004 (Faulkner et al. 2006) and this was then followed by building a production scale Grid during the GridPP2 phase of 2004-2008. With the initial tests and operation of the LHC starting in 2008, the project has just entered the GridPP3 exploitation phase for the period 2008-2011.

(*a*) *UK Tier 1 Centre*

The UK Tier 1 Centre is located at Rutherford Appleton Laboratory and the 2008 hardware configuration is based around a cpu cluster of 3,200 cores delivering 4,500 KSI2K and 340 disk servers with 5,400 drives 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.

A share of the raw data from the LHC experiments is transferred from the Tier 0 Centre at CERN via the LHCOPN to the Tier 1 Centre at RAL for processing. Here 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 (HSM) 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 developed at CERN (Presti et al. 2007). Each of the LHC experiments, ALICE, ATLAS, CMS and LHCb, use separate CASTOR instances to avoid resource conflicts and to minimise problems with data flow.

(*b*) *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 data 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.

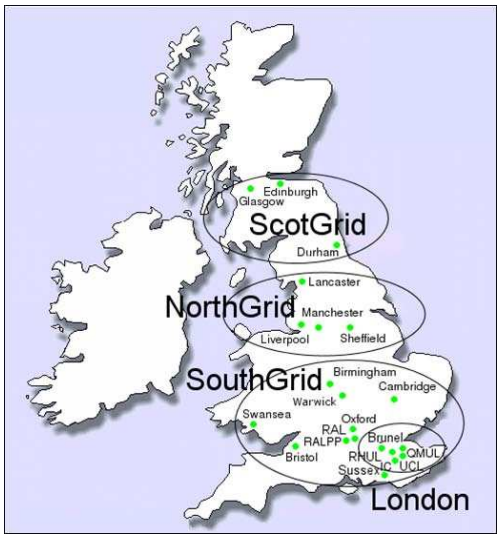

Figure 2. UK Tier 2 centres

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 ∼2,000 KSI2K and several hundreds of Terabytes of disk to small departmental clusters of a few machines.

5. User Access

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 issued by a National Certification Authority, which in the case of the UK is run 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 by issuing a command (*voms-proxy-init*). This contacts a VOMS server to acquire the user attributes/privileges and generates a Grid proxy with a VOMS extension for a limited lifetime (typically 12 hours).

6. 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 such as DataGrid, Globus, GriPhyn, IVDGL, EGEE and LCG.

(*a*) *Workload Management - WMS*

A Grid job is specified using a Job Definition Language (JDL), 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 (RB) 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 (CE), 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 (SE) and catalogued.

(*b*) *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 adopted by LCG. This protocol is designed to allow 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. Version 2.2 of SRM (Donno et al. 2008) introduced the concepts of storage classes and space tokens. These provide experiments with the funcionality to place data on different combinations of storage device to reflect the different retention policies and access patterns. The three main storage classes for experiment data are usually defined through the device mappings of *tape1disk0*=nearline/custodial, *tape1disk1*=online/custodial and *tape0disk1*=online/replica.

Data distribution between the Tier 0, Tier 1 and Tier 2 centres is performed using the gLite File Transport Service (FTS). This uses the concept of unidirectional channels to provide point-to-point queues between sites. Servers are located at the Tier 0 and Tier 1 sites with CERN providing the initial "export" of data and the

Tier 1 servers moving data between themselves and pushing/pulling data from the Tier 2 sites. File transfers are handled as jobs, providing prioritisation and retry mechanisms in the case of failure.

(*c*) *Distributed Databases and Catalogues*

In addition to the experiment data (physics "events"), experiments also require a large amount of non-event data, such as sub-detector conditions, calibrations and geometry descriptions, to be stored in relational databases 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 files (or their replicas) and this is achieved through the LCG File Catalogue (LFC) in the gLite stack. The LFC contains the logical file name (LFN) to physical file mappings along with any associated replicas on the Grid.

The necessary infrastructure for these worldwide distributed databases has been set up by the LCG 3D (Distributed Deployment of Databases for LCG) project which uses Oracle Streams technology to replicate the databases to the external Tier 1 centres outside CERN (Duellmann 2006). At the UK Tier 1 centre, there is a dedicated three-node database cluster for the conditions database of the ATLAS experiment (Viegas et al. 2008), while the LHCb conditions database (Clemencic 2008) resides on a two-node cluster. These database clusters have high availability and performance allowing a large-volume of transactions to be applied to the read-only copies at RAL.

The LFC back-end database system also runs on an Oracle Enterprise relational database. The ATLAS LFC is supported on one node of a three-node database cluster shared with the FTS services which run on the other two nodes. This cluster is highly configurable and if a node fails, the remaining resources can be shared across the applications. Similarly, the LHCb LFC service is deployed on the same database cluster as the 3D service, allowing these two services to share the same hardware and database configuration.

7. VO Specific Software

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 has an API which can be used via three methods: an interactive interface, in a script or through a GUI. 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 transfer management

database, providing a robust system to manage global data transfers. The agents communicate asynchronously through a blackboard architecture and the system can achieve disk-to-disk rates in excess of 500Mbps and sustain tape-to-tape transfers over many weeks.

8. Grid Deployment, Operation and Services

In order to achieve a good overall 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 any operational problems at sites within its domain. GridPP sites are covered by the UK/Ireland ROC. The Grid Operators are a distributed team in charge of providing active and continuous monitoring of the Grid services - this responsibility is rotated on a weekly basis between the ROCs. In the UK, overlapping with the ROC is a GridPP deployment team who provide more dedicated software support and track and resolve LHC VO specific problems and issues. This team consists of technical coordinators from each Tier centre plus experts in specialist areas such as networking, security, storage, gLite middleware and the experiment software. Strategic decisions and project-wide policies (including service levels defined in a MOU) are set by this team together with other GridPP experts and management who form a Deployment Board. Allocation of resources is agreed via a User Board and their use monitored as part of a GridPP project map.

At the cluster level, fabric is usually monitored locally by open source service and network monitoring programs such as Nagios and Ganglia. These can display variations in cluster load, storage 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 (usually daemons and cron jobs) which probe sites and publishes the results to an Oracle database. There is ongoing work to dynamically import the results of these Grid test jobs into the site monitoring frameworks so that middleware and software configuration problems can be highlighted via Nagios alarms. At the ROC level, problems are tracked by Global Grid User Support (GGUS) tickets issued to the sites affected or to the middleware/operations group. If a bug is discovered in any software component, it is reported in bug tracking software called Savannah where developers can obtain details of the problem and record progress on fixing it.

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 a VO-specific standpoint. Dashboards have been developed to present this information in a concise and coherent way. In addition, GridPP has found it hugely effective to also develop a suite of tests which regularly exercise the essential components in the experiment computing chains. In the case of the ATLAS experiment, jobs are sent every few hours to each UK site on the Grid and perform an exemplar data analysis (copy data to the site, run an analysis using experiment software, write, check and catalogue the results). The efficiencies are compiled and displayed on the Web for each site, providing a very informative service showing the user experience to system admininistrators.

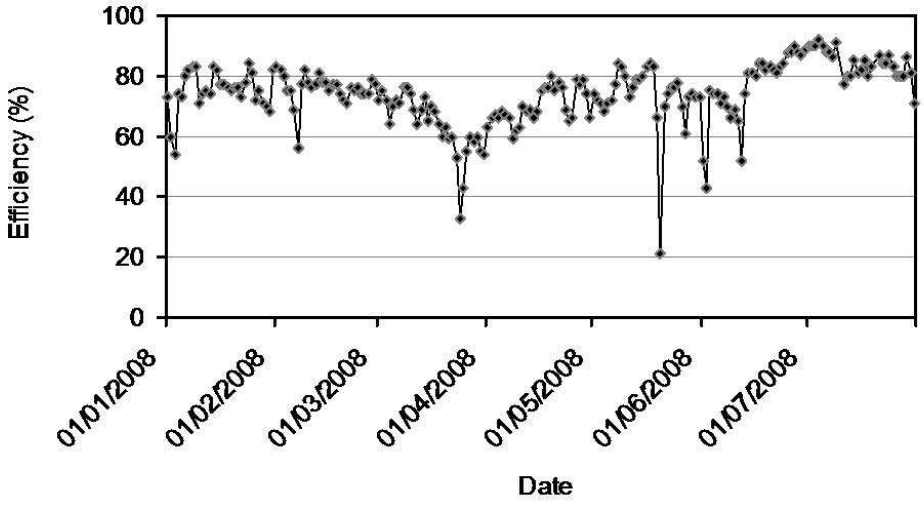

Figure 3. Percentage of successful ATLAS test jobs averaged over all GridPP sites

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.

9. Experiment Experience

Over several years, the LHC experiments have individually exploited and tested 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. The number of batch jobs submitted each month to the UK Tier 1 Centre during the build-up to LHC operation 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.

In addition to simulated data, the two GPD experiments ATLAS and CMS, have made extensive use of cosmic rays to test their subdetectors and data processing systems.

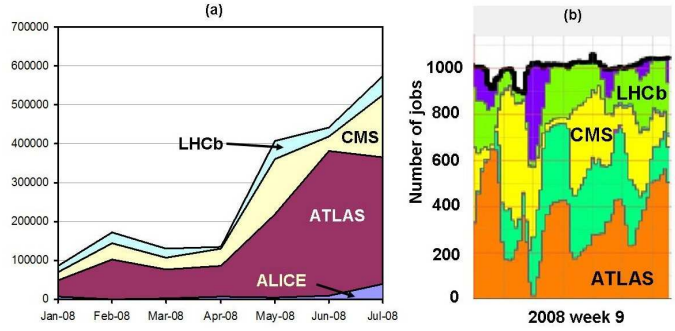

Figure 4. Batch jobs from LHC experiments at UK Tier 1: (a) submitted each month in buildup to LHC operation, and (b) simultaneous load during one week in Feb 2008

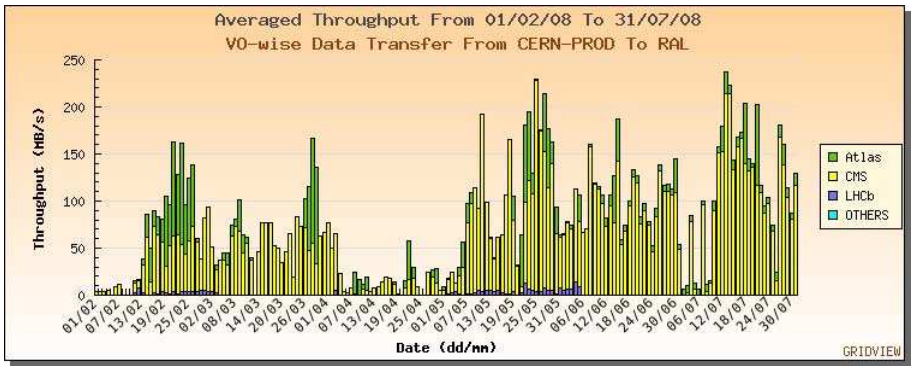

Figure 5. Averaged daily throughput from CERN to UK Tier 1 Centre

10. 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.

At the time of writing, the commissioning of the collider is well underway with two separate synchronisation tests of the beam transfer systems successfully performed in August 2008. Bunches of protons travelled down the transfer lines from the Super Proton Synchrotron through a single sector of the LHC. Particle tracks were observed in the sub-detectors of the LHCb experiment, marking the first arrival of real data from a proton beam.

The final day of AHM2008 should also mark the first attempt to circulate a beam through the entire LHC with proton-proton collisions expected in October; GridPP will then be ready to record the first real LHC collision data following seven years of development.

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.

References

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

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

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

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

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

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

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/.

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

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

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)

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.

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/

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.

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/.

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/.

Pordes, R., et al. 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/

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

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)

Donno, F., et al. 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.

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).