

GridPP: Development of the UK Computing Grid for Particle Physics

Authors

See http://www.gridpp.ac.uk/collaboration_members.html

PACS classification numbers

07.05.-t Computers in experimental physics

07.05.Bx Computer systems: hardware, operating systems, computer languages, and utilities

07.05.Hd Data acquisition: hardware and software

07.05.Kf Data analysis: algorithms and implementation; data management

07.05.Tp Computer modeling and simulation

07.05.Wr Computer interfaces

29.50.+v Computer interfaces in elementary-particle and nuclear physics

29.85.+c Computer data analysis in elementary-particle and nuclear physics

Abstract.....	4
1 Introduction to Grid computing for particle physics.....	4
1.1 Particle physics computing	4
1.2 The LHC programme	5
1.2.1 The LHC computing challenge.....	5
1.3 Grid computing	10
1.4 The GridPP project	10
1.4.1 The LHC Computing Grid (LCG)	11
1.5 How the particle physics Grid works.....	11
1.6 Computing for non-LHC experiments.....	13
1.6.1 BaBar	13
1.6.2 CDF and DØ	13
1.6.3 UKQCD	14
1.6.4 Other	14
1.7 Grid projects.....	14
1.7.1 Middleware projects.....	14
1.7.2 The European Data Grid (EDG)	15
1.7.3 Enabling Grids for E-science (EGEE)	15
1.7.4 The UK e-Science programme.....	16
1.7.5 The UK National Grid Service (NGS).....	16
1.8 Summary	16
2 Summary of developments during GridPP1	17
2.1 Physics motivation	17
2.2 LHC Computing Grid	17
2.3 Middleware and applications	18
2.4 Planning and deployment.....	18
3 Outlook	18
3.1 GridPP2.....	18
3.1.1 Middleware and applications	19
3.1.2 Deployment and operations	19
3.2 CERN LCG.....	20
3.3 EGEE (Enabling Grids for E-Science).....	21
3.4 Conclusion	23

Figure 1: First level trigger event rates and sizes for particle physics experiments. The shaded quadrant in the bottom left corner covers the scale of data for which traditional computing models were used. On this scale all the finally selected raw data can be stored on the site at which it was produced and the computational analysis of the data could also be done on-site. Outside the dotted line, the four LHC experiments either have large event sizes (ALICE), large event rates (LHCb), or both (ATLAS and CMS) and are opting for a Grid solution for their offline computing requirements.6

Figure 2: The LHC Tiered Computing Model. Raw data is handled at the CERN Tier-0 centre. Both raw and processed data are then shipped to Tier-1 centres around the world. Each Tier-1 has a number of related Tier-2 centres: in the UK these are regional groups of Tier-3 institutes, such as universities or research centres. Estimated resources when first LHC data is analysed [1kSI2k = 3GHz CPU] Disk only.9

Figure 3: Schematic evolution of application communities benefiting from EGEE infrastructure. (Ack. the EGEE Consortium).....23

Abstract

The GridPP collaboration is building a UK computing Grid for particle physics, as part of the international effort towards computing for the Large Hadron Collider. The project, funded by the UK Particle Physics and Astronomy Research Council (PPARC), began in September 2001 and completed its first phase three years later. GridPP is a collaboration of approximately 100 researchers in 19 UK University particle physics groups, the Council for the Central Laboratory of the Research Councils and CERN, reflecting the strategic importance of the project. In collaboration with other European and US efforts, the first phase of the project demonstrated the feasibility of developing, deploying and operating a Grid-based computing system to meet the UK needs of the Large Hadron Collider experiments. This paper describes the work undertaken to achieve this goal.

1 Introduction to Grid computing for particle physics

1.1 Particle physics computing

Particle physics has always been at the forefront of the development of scientific computing, and with Tim Berners-Lee's invention of the World Wide Web [1] at CERN has influenced all computing. We start with a brief examination of how particle physics computing has developed over the last twenty years.

Twenty years ago particle physics computing was dominated by large mainframe computers, typically manufactured by IBM, situated at a few key centres. These were complemented by smaller institute-scale machines, such as the DEC-VAX range, used for testing and smaller jobs. So great was the need to communicate with and between these machines that particle physics became one of the first serious users of international networks. The computing paradigm then (as now) was dominated by submission of prepared jobs in batch mode, to run over large data sets stored close to the computing nodes, with typically a few hours or days turn around time.

Ten years ago, particle physics evolved to the use of modest clusters (~50 nodes) of small but powerful machines such as Sun, HP and IBM workstations. This led to the development of sophisticated, often bespoke, software suites to manage job submission and data storage. Each experiment would run its own clusters supported by in-house expertise. This model was quite appropriate to the data rates and simulation requirements of experiments at electron-positron colliders that were in full production at the time; for example, the ALEPH experiment had a total raw data production of 3.5 TB [2]. At this time networks were beginning to seriously evolve to higher capacities of 155 to 622 Mbits/s. Particle physics moved to a predominantly Internet Protocol-based model using the SuperJANET national education and research network, with a few residual private links to CERN and the US.

As new proton collider experiments began operating, data rates increased and the need for more computing power arose. For example, the DØ experiment at Fermilab produced 30 TB of raw data between 1992-1996 [2]. These higher data rates led to the transition to large clusters of commodity computers (PCs) linked by cheap Ethernet. These 'farms' have now become the norm, based upon the Linux operating system that is now the de facto standard for scientific computing. Such clusters are run by

large computer centres, such as the Atlas Centre located at the Rutherford Appleton Laboratory in Oxfordshire. Facilities have also been constructed at most University sites. The international networks had by now grown to use 10 Gbit/s backbones which were (and still are today) more than adequate, although this is expected to change as many end nodes become capable of generating data rates greater than 1 Gbit/s.

Thus, five years ago we had reached a model of reliance on commodity clusters of the order of a few hundred nodes, backed by gigabytes of storage, and connected by national and international networks based upon 10Gbit/s backbones with 1 Gbit/s access links. During this time we have also seen the evolution of the programming norm from Fortran to C and then to object oriented programming based upon C++ and sometimes Java.

However, the cluster computing model will not be sufficient to allow efficient analysis of data from the next generation of particle physics experiments, and in particular CERN's Large Hadron Collider (LHC [3]) programme. In the following section we examine the computing needs of the LHC and other current experiments, leading to an analysis of why Grid computing has been chosen.

1.2 The LHC programme

To be able to search for new physics at the scale of 1 TeV requires the Large Hadron Collider which is under construction at CERN. The LHC will collide protons together at 40 MHz with a total centre of mass energy of 14 TeV, making it the most powerful accelerator in the world for decades to come.

The LHC will house four experiments, positioned around the ring at points where the proton beams collide. Each high-energy proton proton collision will generate hundreds of particles that will shower out into the detectors to be recorded and analysed. When operating at full intensity, around 20 such collisions will occur each time the beams cross, giving further complexity.

Two of the detectors, ATLAS [4] and CMS [5], are general purpose, and will attempt to explore the complete range of physics at the TeV scale. These will be the largest experiments ever built with up to 10^8 readout channels. The LHC also provides a unique opportunity for more specialised experiments. A third experiment, LHCb [6], will study the heavy third generation b-quark, whose decays violate the combined symmetry of charge and parity conjugation. Finally, the ALICE [7] experiment will focus primarily on collisions of heavy ions, in order to study the creation of quark-gluon plasma.

1.2.1 The LHC computing challenge

The experiments carried out at the LHC will produce more data than any previous coordinated scientific endeavour. Careful analysis of all of this complex data will be required to achieve the full experimental sensitivity required in the search for the Higgs boson, supersymmetry, CP violation and precision tests of the Standard Model.

With the construction of LHC assured in the mid-1990s, attention turned to the offline computing problem. The Monarc project [8] was established in 1998 to create a blueprint for cooperation between the many distributed computing centres. Then, in 2001 the Hoffman Review [9] established the scale of the required resources. This

showed that the four major experiments at the LHC collider would each generate PetaBytes of raw data per year, to be processed into event summary data at CERN and then analysed by the worldwide community of LHC physicists. This has since been updated to an estimate of 15PB per year. At the time of writing, it is projected that analysis of the data will require 100,000 CPU (at 2004 levels of processing power)[10].

The LHC experiments are therefore of a different scale and complexity to previous particle physics experiments. The CDF and DØ experiments at Fermilab each accumulated tens of TB of raw, reconstructed and analysed data during their first run. In comparison, the ATLAS experiment will produce PB of data each year, requiring hundreds of times more CPU to analyse than the first run of CDF or DØ.

A comparison of first level trigger event rates and sizes for several past, current and future particle physics experiments is shown in Figure 1. This illustrates the scale of the online computing requirements after initial filtering.

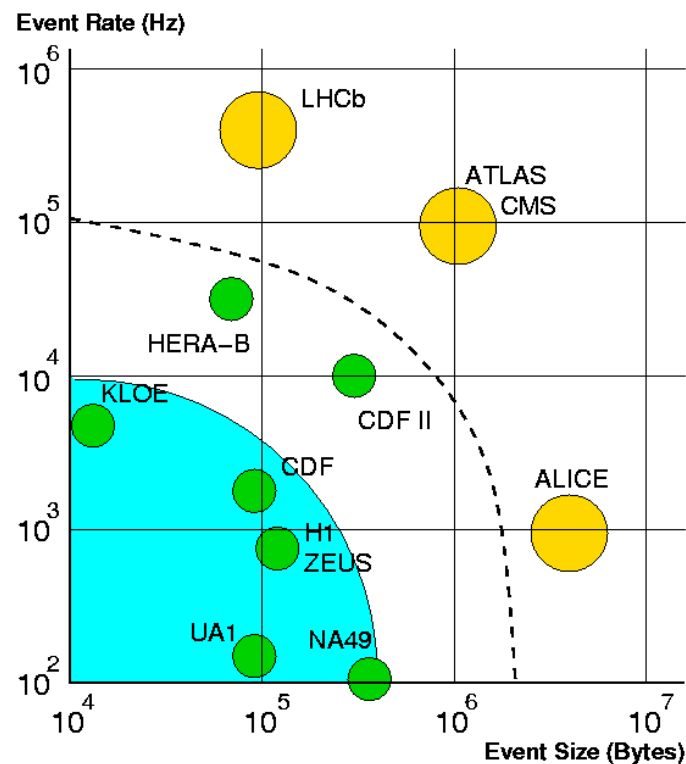


Figure 1: First level trigger event rates and sizes for particle physics experiments. The shaded quadrant in the bottom left corner covers the scale of data for which traditional computing models were used. On this scale all the finally selected raw data can be stored on the site at which it was produced and the computational analysis of the data could also be done on-site. Outside the dotted line, the four LHC experiments either have large event sizes (ALICE), large event rates (LHCb), or both (ATLAS and CMS) and are opting for a Grid solution for their offline computing requirements.

As discussed at the start of this paper, data analysis for previous experiments took place using central computing clusters. Although CPU and disk storage become

cheaper each year, there was insufficient funding available at the CERN centre to both build the LHC and meet its computing requirements. Substantially more resources, both in terms of hardware, technical support and funding, were distributed through CERN member's home states. Therefore, despite noting the superficially attractive option of concentrating all computing resources centrally at CERN, the Hoffman review concluded that this would be impractical for a number of reasons, including:

- it would rule out access to funding that would not be available at CERN;
- it would be difficult to exploit established computer expertise and infrastructure in national labs and universities;
- the wish to devolve control over the computing resource.

However, this would require a solution for distributed data and CPU management. In previous experiments, all file transfers had taken place by hand, with users needing to login directly to accounts at centres where they wished to run jobs or use data. Ad-hoc solutions for distributed worldwide Monte Carlo production had been developed by the experiments (e.g. Funnel for ZEUS and Centipede for H1). However, the Monarc project found that there had been few successes in distributed computing, with problems including; support and continuity at remote sites; maintaining and updating the same code at a number of sites; and hardware and operating system incompatibilities.

With data volumes increasing by at least two orders of magnitude, a more structured approach would be required to make LHC data available to the 5,000 physicists working on the collaboration worldwide in 500 research institutions and universities. Data would need to be distributed across the sites robustly (each file must be replicated on more than one site); efficiently (files should be stored close to where they are needed most often); and transparently (end users should not need to be concerned with where the data is stored or how). It would also need to be 'persistent' – available for the whole 15 years of the LHC project.

It was therefore decided to use a distributed, hierarchical model, based around the new technology of Grid computing. The Grid aims to integrate the computing power of sites worldwide, allowing a user to submit a computing job from anywhere in the world, for that job to run at a site somewhere else in the world on data that is not local to the user and for the results to be returned.

A Grid for the LHC would:

- incorporate regional and local computing resources into the worldwide computing effort. This has the advantage of allowing individual institutes and national centres to maintain and upgrade these resources.
- make efficient use of CPU worldwide by automatically balancing the computing load across available resources.
- allow users to run their Grid jobs and obtain access to their data without the need to apply in advance for computer accounts at every single Grid site and logon to the sites manually.
- store copies of data in regional centres. This is a more effective use of limited network bandwidth than copying them from CERN every time they are needed. Such a scheme would therefore balance proximity of the data to centralised processing resources with proximity of users to frequently accessed data.

- allow the effective involvement of scientists and students in each region, and utilise computing expertise from regional computing institutes, without the need for them to work physically at CERN.
- have no single points of failure, as there are multiple copies of data and computing jobs are assigned to available resources.
- cover all time zones, enabling easier provision of 24 hour support.

Figure 2 below shows the agreed model for LHC data processing. Data leaves the detectors themselves at ~ 1 PB/s, but this is reduced to a few hundred MB/s by the online trigger system and this is recorded to mass storage. The resulting datasets will be a few PB of raw data per year per experiment, or roughly 15 PB per year overall.

It is expected that copies of all of the raw data will be transferred to a set of Tier-1 sites at major national computing centres. Data reduction will then be performed both at CERN (Tier-0) and the Tier-1 sites. In this way, data sets needed by physicists will be spread throughout the world in multiple copies for resilience. Smaller Tier-2 sites (typically universities or groups of universities) will also take responsibility for analysis and simulated data production. The overall model therefore will lead to a mesh of data flows and processing steps spread over geographically distributed centres. The estimated load on the networks will be some tens of Gbits/s.

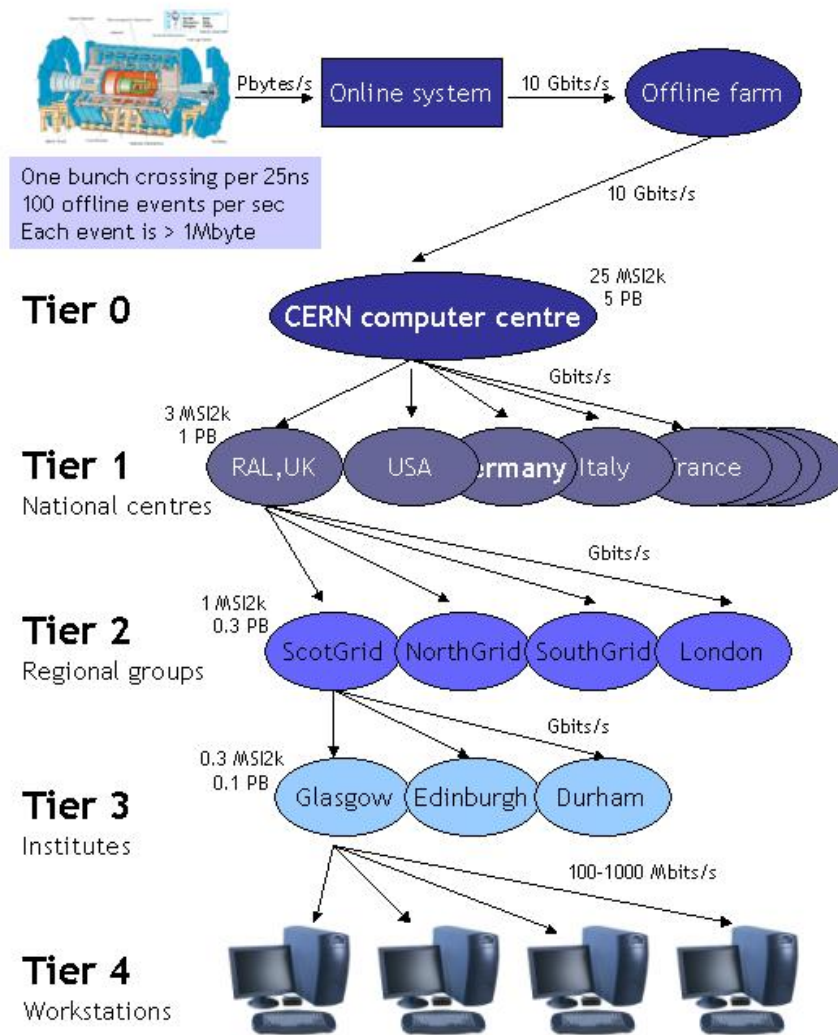


Figure 2: The LHC Tiered Computing Model. Raw data is handled at the CERN Tier-0 centre. Both raw and processed data are then shipped to Tier-1 centres around the world. Each Tier-1 has a number of related Tier-2 centres: in the UK these are regional groups of Tier-3 institutes, such as universities or research centres. Estimated resources when first LHC data is analysed [1kSI2k = 3GHz CPU] Disk only.

However, Grid computing for the LHC poses a number of challenges, including: deciding where computing jobs are run; who should have access to the data and computing resources; how access will be controlled; ensuring sufficient levels of network bandwidth; deploying software and updates across many different locations; managing data over long periods of time; and providing fast and secure access to that data. In addition, software for analysis needs to be developed that can use the Grid's resources, and interfaces built to allow particle physicists to submit their jobs to the Grid. In the next sections, we introduce Grid computing, the GridPP and LCG projects, and then consider how a particle physics Grid works in practice.

1.3 *Grid computing*

In its most general sense, Grid computing is the collaborative use of computing resources in multiple administrative domains. The resources in question may be computers themselves, data storage systems, computer networking, or more sophisticated facilities such as electron microscopes, telescopes and even particle accelerators. Coordination across institutions may be required either because of the uniqueness of some resources; to facilitate collaboration between teams of experts from several groups; or to aggregate resources to tackle large-scale problems beyond the scope of individual institutions. The data processing and analysis challenges posed by the experiments at the LHC have all three of these requirements.

In 1964 Martin Greenberger wrote an article for *The Atlantic Monthly* entitled “The Computers of Tomorrow” [11], in which he proposed that computing development would follow the same lines as the development of electricity, evolving into a utility that people could use for any kind of computational problem. The electrical power Grid provides a reliable, constant source of power that any machine designed with the appropriate wiring can plug into and use, irrespective of which power plant generated the power. Similarly, the computational Grid was foreseen to be a utility of processing power providing standard interfaces to which any computer can link to harness its power, with the source of the power completely transparent to the user.

Grid Computing gained much wider visibility in 1999 following the publication of “The Grid – Blueprint for a New Computing Infrastructure” by Ian Foster and Carl Kesselman [12]. This influential book painted a broad vision for Grid computing and the opportunities it presented across a range of scientific and industrial applications. We also saw the emergence of the Global Grid Forum, now a standards-defining body that is essential to focus effort and allow national projects to integrate with worldwide partners. Thirdly, the UK established an e-Science programme, described in Section 1.7.4 below. The GridPP project, the UK’s contribution to computing requirements for the LHC, was part of this e-Science programme.

1.4 *The GridPP project*

GridPP, the UK Grid for particle physics, is developing the computing tools and infrastructure to allow particle physicists to collaborate effectively when LHC data-taking begins in 2007. The six-year, £33m project, began in 2001 and has been working in three main areas:

- Developing applications to allow users to submit data to the Grid for analysis;
- Writing middleware to distribute jobs around a secure Grid;
- Deploying computing infrastructure at sites across the UK.

A key component in this development was, and continues to be, the use of new tools and techniques within current experiments. In this way the current experiments quickly benefit from new developments, which are tested in real production use.

The UK particle physics Grid currently contains some 3,000 processors, and has approximately 1 PB of storage available for the experiments.

GridPP is now in its second phase (GridPP2). Over the next three years, the lessons learned will be used to improve the deployed systems so that they can be scaled to the

levels required for the LHC. The infrastructure deployed in the UK will be tested continuously, both by current experiments and by the LHC data challenges to ensure that the final system is ready for the first LHC data.

1.4.1 The LHC Computing Grid (LCG)

In 2002, CERN established the LHC Computing Grid Project [13], to deploy Grid computing solutions at computing centres that have signed up to deliver computing resources for the LHC experiments. Existence of the necessary middleware is not in itself enough to solve the LHC computing challenge. Experiments also need software that is able to exploit the Grid infrastructure and guaranteed access to sufficient computing resources. Overall, the GridPP contribution represents almost 40% of the non-CERN resources allocated to the project and roughly 25 staff members funded via GridPP are active in each of the project's key areas.

At the time of writing, the LHC computing Grid had more than 10,000 CPU and more than 7 PB of storage capacity on disk and tape, at over 100 sites in 31 countries. This is approximately 5% of the processing capacity required for the LHC in the long-term.

1.5 How the particle physics Grid works

This Grid infrastructure is in general built from an aggregation of many different hardware platforms running different operating systems, and managed by different administrations. At the same time the Grid must appear to the user application with a uniform interface from the point of view of submitting and retrieving work, and such that the user does not have to have a specific knowledge of, or relation to, the resources. This is the function of middleware – to provide a uniform access layer of services, which buffer the user from the details of the underlying resources.

GridPP uses middleware developed as part of the European DataGrid (EDG, see Section 1.7.2) project. In association with CERN's LHC Computing Grid project (LCG) we have tested and hardened these and other components from the Virtual Data Toolkit [16]. Middleware development continues as part of the Enabling Grids for E-science in Europe project (EGEE, see Section 1.7.3).

Consider a practical example: a physicist in Sheffield University has prepared an analysis program to sort through a large data set which has been recorded by an LHC experiment. The physicist sends this into the Grid, whence a suitable computing resource must be located upon which to run the job – let us say that this is available in Karlsruhe. The data set required as input resides upon a big disk farm in the Rutherford laboratory in Oxfordshire. This data set must be identified as part of executing the job, and then copied to Karlsruhe prior to execution. When the job is finished the original user must be informed and the results sent back.

This process may sound simple, but now we are going to walk through the components needed to allow this to all happen in a "Grid like" way. The Grid aggregates resources that may be added or removed at any time – so there is no sense in which the user can either know of the resource they might use, nor be known personally by the resource provider or have a classical "account" upon such a resource.

- The user prepares a job. This needs a user interface, to help the user prepare the job, submit it, monitor it and retrieve the results. The user interface can be the experiment computing framework, or it might be a web portal or command line interface. Section S1 examines the applications that particle physics experiments have developed to allow data analysis and simulation on the Grid.
- For simplicity, let us assume the job is an executable binary, plus some specification of input and output files. We said that the user must send this to the Grid. What does this mean in practice? Let us imagine the user has an access portal (a simple Graphical User Interface (GUI) which will take the job specification and do the right thing with it). To take this first step there must then be a service somewhere on the internet, known by the GUI, to which the job specification can be sent, and which will be able to locate a “Grid resource” somewhere around the globe upon which the user has the right to run the job. This is known variously as a resource broker, or workload manager (WM), and is considered in Section S2.1. We can already see that since in any scalable system there must be the possibility of several such WMs, possibly provided by third parties, then there must be some form of standardised job submission interface and description language – we will see this theme of standard protocols and interfaces pervading Grid middleware.
- The first thing the WM will do upon receipt of the job is to check the identity and rights of the user. This leads immediately to the need for security services based upon digital credentials, which can prove identity, confirm that the user belongs to a particular experiment (or virtual organisation-VO) and perhaps confirm their role within that VO. These security services must be accessible by many of the Grid components. Security is examined in Section S2.6.
- Once the WM has performed authentication and authorisation checks, then it must locate suitable resources. It can only do this if there is a publication and discovery system to advertise resource availability. This leads to the need for monitoring and information systems, considered in Section S2.3. These must also allow the WM to verify that the user has the rights to use any resource it finds.
- Once a computing resource is found, then the job must be sent to it. This implies the need for a standardised job submission interface sitting on top of whatever actual job submission system runs locally. In addition a status logging system is required so that the user can query the progress of the job and retrieve results. These components were built on top of the Globus toolkit and Condor (see Section 1.7.1).
- At some point the job will run. It will immediately need to open the input file, which is of course stored remotely. The user’s programme will also most likely refer to some logical identifier (such as “the high transverse momentum lepton sample”) rather than a physical identifier. This leads to the need for a data replica management service, which manages and keeps track of logical to physical file mappings, as well as actual physical file replica locations. Data management is examined in Section S2.2.

- Once the physical files are identified, they will need to be pulled from the remote storage. This also requires a standardised way of accessing and transporting data from storage elements that comprise of mixtures of disk and mass storage archives. Section S2.5 looks at Grid storage.
- In order to ensure that data can be transported in a reasonable time, or even to select the best physical data set location, the data management services must be able to query the status of network connections in the same way as for CPU and storage availability. This information is also required as part of the operational monitoring infrastructure. Networks and monitoring are considered in Section S2.7.
- When the job finishes, the output file must be sent back to the user. The user will not in general be “online” and nor will there necessarily be an internet connection to the user’s site. This relies on the logging and bookkeeping services described as part of workload management.

1.6 Computing for non-LHC experiments

Although the move towards the Grid was predicated upon the requirements of the LHC experiments, nevertheless some of the existing experiment programmes are also able to benefit from the transition to Grid technology. In the US, the Particle Physics Data Grid has reported a 20-50% gain in resources for experiments through job scheduling over the Grid, from the use of opportunistic and distributed resources.

These non-LHC experiments differ widely, not only in the physics they address and the detectors and accelerators they use, but also in their computing models and their timescales. There are four such experiments that played a major part in GridPP1.

1.6.1 BaBar

The BaBar experiment situated at the Stanford Linear Accelerator Centre (SLAC) is studying the phenomenon of CP violation through the decay of B mesons. BaBar initially started with a central computing model. With the increasingly successful operation of the PEP-II accelerator, producing billions of B mesons, it became clear that a new distributed strategy, involving the use of large ‘Tier A’ sites at RAL and other national centres, was necessary. In 2004, BaBar produced over 2 billion Monte Carlo events with 30 sites in 6 different countries using its legacy system for event production. The experiment plans to use Grid resources to triple this rate, to keep up with the ever increasing performance of the Stanford Linear Accelerator. Section S1.3.1 gives more detail on BaBar’s use of the Grid.

1.6.2 CDF and DØ

The UK is involved in two experiments at the Tevatron at Fermilab (FNAL): CDF and DØ. These measure proton-antiproton collisions which are similar to those of the LHC but at lower energy (2 TeV in the centre-of-mass). These experiments have been running for many years, but are currently undertaking a new data taking period (Run II) with a large increase in luminosity and data processing requirements. The previous computing model had been centralised, however, the data processing requirements for Run II outstrip the capability of the FNAL system, including any imaginable upgrades to it. The DØ SAM system (Sequential Access to data via Metadata) has been adapted to Grid technology, and is now used by both DØ and CDF. DØ has reported that using

Grid protocols to transfer data has resulted in increased throughput by a factor of five. They were also able to reprocess 100 million events remotely, to meet an otherwise impossible publication deadline. CDF aim to have 50% of the experiment's computing outside Fermilab by 2005. Use of the Grid by CDF and DØ is examined further in Section S1.3.2.

1.6.3 UKQCD

The UKQCD collaboration is concerned with lattice calculations in respect of Quantum ChromoDynamics (QCD). Lattice QCD, as it is called, is performed by dividing space-time into a discrete lattice (in four dimensions) and then performing parallel calculations at the nodes and on the links. To do this requires very large and leading-edge parallel computing facilities and this work shares many features with the experiments. Large data sets produced on the lattice must be stored and replicated using distributed Grid technology. These data, which are in effect the raw "experimental" data, must then be processed repeatedly by remote collaborators for different investigations. This user processing step is performed on conventional architectures and is amenable to the use of distributed Grid resources. Section S1.4.1 considers QCDGrid.

1.6.4 Other

For completeness we mention the other experiments in which the UK is involved. Some have been running for many years and have evolved a computing model which works for their limited data rates, others are newly-emerging and will be able to embrace a Grid approach from the start.

In the first category lie the LEP and HERA experiments. These use systems whereby the data processing and data storage is essentially all done at the host accelerator site (CERN and DESY respectively). The LEP experiments (the UK is involved in ALEPH, DELPHI, and OPAL) have completed data taking, and their analyses continue within a system that works and with nothing to gain from modernisation. The HERA experiments are now entering phase 2 running and while data-taking continues, the computing resources are under increasing strain, and there is a motivation to adopt new techniques.

The second category of newly emerging experiments includes the Linear Collider groups (CALICE, LC-FI and LC-ABD), the neutrino factory studies and the precursor MICE experiment, Dark Matter searches and the neutrino experiments MINOS and ANTARES. These are generally at the stage of a few dedicated users running a small number of simulations, and are not yet at the stage of designing the overall computing system. When one of these matures it will be integrated into the GridPP project, however this did not arise during GridPP1.

1.7 Grid projects

Having set the scene, we now conclude this introduction with a brief description of relevant Grid projects, which set the context for GridPP and the rest of the paper.

1.7.1 Middleware projects

The middleware used by GridPP builds on that from the Globus and Condor projects. The Globus project [14] is focused on enabling the application of Grid concepts to

scientific and engineering computing. In particular it provides the Globus toolkit, which includes software services and libraries for monitoring, discovering and managing resources on the Grid, plus security and file management. Condor [15] is a specialized workload management system for compute-intensive jobs. Users submit their serial or parallel jobs to Condor, Condor places them into a queue, chooses when and where to run the jobs based upon a policy, monitors their progress, and ultimately informs the user upon completion. Condor and Globus components are packaged and tested as part of the Virtual Data Toolkit (VDT [16]) which is then used to provide components of the LCG/EGEE middleware deployed by GridPP.

1.7.2 The European Data Grid (EDG)

In January 2001 the European DataGrid [17] was founded as the flagship European project to develop a prototype Grid service. EDG was funded by the European Union and involved 6 major partners including the UK.

The EDG project served three distinct application areas (i) particle physics (ii) Earth observation and (iii) bio-medical imaging. Significant effort was devoted to working with these applications to Grid-enable their software. The project was successfully completed in March 2004.

EDG developed the higher-level middleware, on top of existing middleware, necessary to build a Grid that could be used by several distributed communities (see Section 1.5 for a description of how the middleware works). The middleware developed by EDG included:

- Workload management and logging services: to accept user requests, match them to resources around the globe, and to track and report progress.
- Storage and data management: required to access heterogeneous storage systems and to manage data replicas thereon.
- Resource information services: required to provide the much richer information set representing “computing”, “storage” and “network” elements needed by the workload management system.
- Installation and configuration: required to manage distribution of software and configuration of computing nodes.
- Virtual Organisation (VO) management: to allow resource providers to verify identity and capability of people belonging to large distributed communities. This was a key development without which a scalable Grid, serving many and varied users, could not be constructed.
- Network performance monitoring: to provide connectivity, bandwidth and stability data to the information services.

1.7.3 Enabling Grids for E-science (EGEE)

The EGEE project [18] is a successor to the EDG project, but aiming to deploy a production Grid for use by e-science. EGEE is funded under the EU framework VI infrastructure funds programme.

In practice EGEE builds strongly upon the previous EDG (and therefore GridPP1) and LCG software stacks, as well as components from several US projects. EGEE is funded at the level of €32m for two years starting in April 2004. GridPP is a key component of EGEE through its relation with the UK Tier-1 centre situated at the

Rutherford Appleton Laboratory, as well as many other groups who work within specific EGEE activities. GridPP will work closely with EGEE and the UK National Grid Service (NGS, see below) to deploy a consistent production infrastructure. More details about EGEE are given in Section 3.3.

1.7.4 The UK e-Science programme

The 5-year UK e-Science programme (2001-2006) was established by the then Director General of the UK Research Councils, Dr. John Taylor. It is a common programme across all UK research councils to develop the advanced collaborative computing required by today's scientific researchers. Over the full 5 years the UK will invest approximately £250m in a wide range of technological developments covering advanced video conferencing, visualisation, data storage, high throughput and high performance computing and computer networking and security. A key component is the engagement of industrial partners, coordinated through a cross discipline "Core Programme" focused on identifying common themes.

The pivotal role of the challenges facing the LHC was recognised at an early stage, and the particle physics community was already collaborating in establishing the EU-funded DataGrid project. The Particle Physics and Astronomy Research Council was given a sizable allocation of the e-Science investment with which it established two major projects: GridPP and the Grid for Astronomy (AstroGrid). In addition, as part of the process of modernising University infrastructure, significant computing resources were installed by the Higher Education Funding Council for England (HEFCE) and the Scottish Higher Education Funding Council (SHEFC) through initiatives such as SRIF (the Science Research Investment Fund).

1.7.5 The UK National Grid Service (NGS)

The UK's National Grid Service [19] began in April 2004 and is the first step towards deploying and operating a Grid infrastructure to support all UK researchers. The NGS builds on the experiences of GridPP and other projects to provide Grid-enabled access to computer clusters, data archives, advanced software, and the UK's High Performance Computing facilities. In addition to a set of core resources, the NGS serves as a means of resource sharing across UK academic institutions. The central services for the NGS are coordinated and supported by the Grid Operations Support Centre (GOSC), providing services such as the UK Certificate Authority (CA) and a central helpdesk.

At present the NGS and GridPP resources only partially overlap. However, the CA, helpdesk and operational monitoring are common efforts across both projects. This integration is expected to grow in future as the underlying Grid technologies become more coherent, allowing common services to support more disciplines, and as more GridPP institutions expand Grid support to cover a wider range of research areas. In the latter case, for example, rather than joining GridPP and the NGS, an institution would simply become an NGS partner supporting GridPP as well as a range of other projects.

1.8 Summary

In this introduction the concept of, and motivations for, Grid computing have been outlined. Grid computing aims to provide a new paradigm of computation for science. The particular need for Grid computing for the LHC experiments at CERN comes

from the unprecedented scale of data they will produce. Current and future particle physics experiments will use the same shared infrastructure. Only by distributing the computational load globally can physicists cope with the requirements of the experiments. In order to meet these requirements, we envisage more than 100,000 computers worldwide of which 10,000 will be in the UK by 2007 operating as a Grid production service.

We envisage Grid computing will do for access to distributed computing what the World Wide Web did for access to stored information: that is, provide the layer of uniformity and transparency to allow use of resources without the users needing to have an explicit relationship with the provider, or even care who or where the provider is. If this model pervades commerce and society in the future, then the developments made in the GridPP project described in this paper will have played a significant role in this evolution.

2 Summary of developments during GridPP1

2.1 Physics motivation

Due to the enormously high energy, high luminosity and complexity of the detectors, the LHC experiments will produce unprecedented amounts of data, estimated to be more than 10 PetaBytes per year, for offline analysis by teams of physicists all over the world. To analyse this data and to generate the Monte Carlo simulated data necessary to understand it will require huge amounts of computing and data storage facilities. These computing resources will be distributed across the world, linked together as a "Grid".

GridPP has built a prototype Grid in the UK that enables the LHC experiments to generate large amounts of Monte Carlo simulated data. This is being tested by current experiments in the USA with which the UK is involved: BaBar at SLAC and CDF and DØ at the Tevatron, FNAL. Several other smaller experiments have also started to use the prototype Grid. Having these non-LHC experiments involved in the development of the Grid has been enormously helpful, since these physicists are performing real physics analysis. This places quite different demands on the robustness and stability required of the systems, compared to simulation or testing.

Although GridPP has focused on the provision of computing resources for the LHC experiments and associated phenomenologists, the project continues to support the US-based and other experiments such as H1, ZEUS, UKDMC and groups preparing for new initiatives such as MICE and the Linear Collider. By 2007 it is expected that nearly all UK particle physics computing equipment will be part of the Grid and hence we have to ensure that *every* UK experiment can use the Grid for their analyses even though the resources they require will only be a small fraction of the total.

2.2 LHC Computing Grid

GridPP funding was instrumental in establishing the LCG project at CERN which is a Grid deployment project organised internally in four areas, Applications, Fabric, Technology and Deployment. The applications area is currently consolidating a number of software projects that are common across the LHC experiments, such as POOL, the common persistency framework and the Physicist Interface project that is now moving to ARDA (Architectural Roadmap towards Distributed Analysis). The

fabrics area is charged with constructing the Tier-0/1 centre at CERN where the LHC data will be reconstructed and significant progress has been made in the development of fabric management tools as well as progressive upgrade of resources. In the Grid technology and deployment areas, the first release of LCG software, LCG-1, based on the EDG middleware and elements of the US Virtual Data Toolkit was deployed to 28 sites worldwide by the end of 2003 and the second release LCG-2 has been rolled out to more than 100 sites worldwide.

2.3 Middleware and applications

GridPP was a major investor in the EDG middleware development project that ended in March 2004 and provided the leader or deputy leader in five of the eight relevant work packages. GridPP's main contributions were the releases of a number of middleware/software packages. R-GMA, the Relational Grid Monitoring Architecture, provides information about the Grid to brokers and users. OptorSim simulates the EDG architecture and allows optimal file replication strategies to be developed. Spitfire allows secure access to metadata catalogues while the Storage Element provides transparent access to mass storage systems. GridSite allows remote management of websites using Grid certificates.

In application development, one of the successes of GridPP1 has been the GANGA project, a joint ATLAS-LHCb initiative to provide a user-Grid interface allowing configuration, submission, monitoring, bookkeeping, output collection and reporting of Grid jobs. GANGA is likely to be adopted by other experiments in the future. Other major developments have been the DIRAC LHCb interface to POOL, AtCom, the ATLAS Commander Monte Carlo production tool and the use of R-GMA to monitor CMS applications. All LHC experiments have been involved in increasingly large 'Data Challenges'. In addition to the LHC experiments, GridPP has helped develop SAM, originally from the US DØ experiment, into a very successful joint CDF/DØ SAMGrid that is deployed at many sites worldwide to schedule data transfers from a central repository for local analysis. The other US experiment, BaBar, has developed a remote submission system that allows users to locate data and run analysis jobs at several UK sites. Finally the UKQCD Collaboration has successfully produced an operational Grid for distributed processing of lattice QCD data spread across 7 nodes at 4 sites.

2.4 Planning and deployment

GridPP has constructed a prototype LHC Tier-1 regional centre at RAL that also serves as a BaBar Tier-A centre. As of November 2004 the total resources comprised ~700 Intel CPUs (450Mhz-2.6GHz), ~ 80TB of available disk space and ~60TB of tape, with a theoretical tape capacity up to 1 PB. Not all of the resources were available via the Grid but that will ramp up over GridPP2. In addition to the Tier-1, GridPP is developing 4 regional Tier-2 centres, to harness the massive resources at the universities provided through the Joint Infrastructure Fund /Joint Research Equipment Initiative and SRIF.

3 Outlook

3.1 GridPP2

The initial three-year phase of the project, GridPP1 - "From Web to Grid", successfully established a prototype Grid in the UK as described above. The project

has been extended for a further three years to September 2007 as GridPP2 - "From Prototype to Production". This will build upon the achievements of GridPP1 and turn the prototype Grid into a production Grid of sufficient size, functionality and robustness to be ready for the first data from the LHC, due at the end of 2007. The GridPP2 project has been allocated £15.9m by PPARC with a further £1m set aside as a possible additional contribution to LCG during this period.

3.1.1 Middleware and applications

In GridPP2, similar areas of middleware activity will be supported as in GridPP1, but the emphasis will change from development to deployment of a full-scale production Grid. In GridPP1 several middleware development projects were undertaken, under the auspices of the EDG project. In GridPP2 these have been consolidated to concentrate on the areas that are of primary importance to UK physicists, such as Grid Data Management (Metadata handling) and Storage Interfaces. GridPP2 will also focus on areas where the UK has particular expertise and played a leading role in EDG such as Workload Management, Security, Information Services (R-GMA) and Networking. This future development work will take place as part of the EGEE project (see below).

Applications development in GridPP2 will continue with support for the LHC experiments ATLAS, CMS and LHCb and the joint ATLAS-LHCb GANGA project. US experiments will also be supported: BaBar, and CDF and DØ through SAMGrid. In addition to existing support for Lattice QCD theory (QCDGrid) there will be support for particle phenomenology, an area crucial to the success of LHC physics, through PhenoGrid. GridPP will also develop a generic 'portal' that will allow other experiments, not directly supported by GridPP, to access the UK Grid and its computing resources.

3.1.2 Deployment and operations

A very important component of GridPP2 will be the Deployment Team led by a newly appointed Production Manager. This team will consist of middleware experts and technical coordinators from the Tier centres who will be responsible for the deployment and operation of the production Grid. This will include development of the individual resources; the technical and organisational integration of the parts; and the roll-out of an operational Grid service to, eventually, a level of stability and an ease of use consistent with a production facility.

Initial models of LHC computing involved a hierarchical arrangement of tiered regional centres, starting from Tier-0 at CERN, to major Tier-1 centres in several countries, then to smaller Tier-2s and Tier-3s in institutes and departments. This model is being modified and the hierarchical nature becoming less distinct with the concept of virtual Regional Centres emerging. However the realities of bandwidth, data, and service hierarchies still motivate a multi-tiered approach and so the terms Tier-1, Tier-2 etc are retained.

In GridPP1, there was no specific funding for Tier-2s as the emphasis was on the development of the prototype Tier-1 Centre which provided sufficient resources during the first phase of the project. However, in future the Tier-1 Centre will not be able to provide all the resources required by the experiments. GridPP is therefore developing four Regional Tier-2s in:

- Northern England (NorthGrid, the Universities of Lancaster, Liverpool, Manchester and Sheffield, and the Daresbury Laboratory);
- Southern England (SouthGrid, the Universities of Birmingham, Bristol, Cambridge and Oxford, and the Rutherford Appleton Laboratory);
- Scotland (ScotGrid, the Universities of Edinburgh, Glasgow and Durham); and
- London (Brunel University, Imperial College London, Queen Mary University of London, Royal Holloway, University College London).

These will logically group together resources provided by the institutes in those regions. It is expected that the total resources at the Tier-2s will be about the same as at the Tier-1 Centre.

Technically, there may be little distinction between a single resource at a Tier-2 or elsewhere, but overall different centres will be better suited to different tasks. In terms of the envisaged distributed data caching of ESD, AOD and TAG data it is clear that this more refined data should be as close as possible to the physicists performing their analysis. TAG data will reside primarily on the Tier-3 resources, such as local workstations, and the required streams of ESD and/or AOD data will reside at the Tier-2s. These centres are already very important politically, financially and sociologically in providing ‘local’ resources and experience and providing a focus for leverage of external funds.

GridPP will be providing funds for hardware and manpower at the Tier-1 Centre and manpower support for the Tier-2s. GridPP will also provide a small amount of hardware so that each institute will have a common set of front end interfaces to the Grid but it is expected that the bulk of the Tier-2 hardware will continue to be provided by HEFCE through initiatives such as SRIF.

In addition there will be continued support for travel to meetings conferences and workshops and support for management positions. GridPP2 also includes the appointment of a full time Dissemination Officer to manage the GridPP website, run events, produce literature, liase with the press and work with other disciplines.

3.2 *CERN LCG*

Five years after it first appeared, offering a way to integrate widely distributed computing resources for the LHC, Grid Computing is still in a state of flux. Changes to the Globus architecture and the many projects working to develop Grid middleware support an argument that there is, as yet, no “Grid Computing” solution.

However, the past four years have seen significant progress towards the integration and alignment of widely distributed particle physics computing resources that have long been recognised as essential for effective exploitation of the LHC data. At the time of writing, the LCG2 service brings together over 10,000 CPUs at more than 100 sites in over 30 countries to provide a computing resource accessible to the LHC data challenges from a central resource broker. In addition to the CPU capacity, nearly 10PB of storage is available and there is a Replica Location Service (RLS) in place to maintain a consistent list of accessible files.

The LCG2 service was used by CMS in March-April 2004 for their data challenge [20] designed to test the full chain of LHC data processing from event reconstruction at the CERN Tier0 to analysis at a remote Tier1 centre [21]. Whilst there are areas of the service that can be improved, the Hoffmann review stressed the need for services to “evolve under the strain of real users attempting to get their work done” and we confidently expect the LCG2 service to improve as problems identified by the experiment data challenges are addressed.

Looking more widely, worldwide computing needs a worldwide support and operations infrastructure. Here again, the past four years have seen significant steps towards meeting the needs of 2007. The LCG project has established a Grid Operations Centre (see Section S3.4) at RAL and also a Grid User Support Centre at the GridKa centre in Karlsruhe, Germany. More recently, Taiwan’s Academia Sinica Computing Centre has established similar centres in an Eastern time zone as the next stage in development of a worldwide, round-the-clock service.

Seen from CERN, then, the situation now is that the LCG project, by setting up an initial production Grid service, has put in place the foundations of the overall Grid environment that will be required when LHC begins operations in 2007. With these foundations laid, we are well placed to exploit the expected developments and improvements in Grid middleware and other software components over the next two years.

3.3 *EGEE (Enabling Grids for E-Science)*

As introduced briefly in Section 1.7.3, EGEE is a two-year project as part of a four-year programme, which aims to integrate national, regional and thematic Grids to create a European Grid infrastructure. It is part sponsored by the EU and part by national resources. Where the DataGrid project led in prototyping Grid systems through R&D and in integrating with other technologies (e.g. Globus, Condor), EGEE now takes over in moving from prototypes to delivery of production quality Grid services for the European academic community. GridPP, along with co-workers in the UK e-Science programme are major contributors to the EGEE project.

EGEE encompasses over 70 partners from 27 countries organised into 12 regional or functional federations:

- CERN
- Central Europe including Austria, Czech Republic, Hungary, Poland, Slovakia and Slovenia
- France
- Germany and Switzerland
- Ireland and the United Kingdom
- Italy
- Northern Europe including Belgium, Denmark, Finland, The Netherlands, Norway and Sweden
- Russia
- South-East Europe including Bulgaria, Cyprus, Greece, Israel and Romania
- South-West Europe including Portugal and Spain
- NRENS (National Research and Education Networks)
- United States

The project is structured into the following activities (where the terms in brackets refer to EGEE formal activity groups):

- Grid operations, support and management (SA1). This is the largest (by funding and people) activity, dealing with production software releases and support, and a distributed operational model to bind national Grid centres into a coherent framework.
- The interface between EGEE and the European networks (SA2), comprise Géant and national education and research networks such as SuperJANET.
- Joint research activities to cover development of middleware to enhance the production infrastructure (JRA1, 2, 3, 4)
- Application support to help applications use EGEE (NA4)
- People networking to cover management, dissemination, training and international co-operation (NA1, 2, 3, 5)

A Grid Operations Centre provides both real-time monitoring of Grid status and archives of Grid usage.

The operational infrastructure is backed up in two key ways. Firstly, there is an extensive middleware development programme focused on engineering and integrating production quality Grid services. This is based on prototypes from previous R&D projects (e.g. DataGrid), core software from other projects and new software. A major aim of this activity is to move to Web Service based middleware. Secondly, a strong emphasis is placed on training, dissemination and outreach. A substantial set of training modules is being developed and delivered and a particular activity in developing new application communities in using the infrastructure is underway. An Industry Forum provides links to the commercial world and a number of relationships with related projects have been developed.

With resources derived from a large number of domains and funding sources there exist a variety of approaches. In order to harmonise contributions into a coherent whole while at the same time respecting local ways of working, EGEE also convenes a high-level group working on policies and best practice and involves a cross section of the wider Grid community in Europe.

As a starting point the initial applications exploiting EGEE are based on High Energy Physics and Bio-medical communities. The initial starting point for production services is based on the second release of the LHC Computing Grid infrastructure. However, it is a major deliverable of the project to expand the portfolio of disciplines benefiting from EGEE, as exemplified in Figure 3.

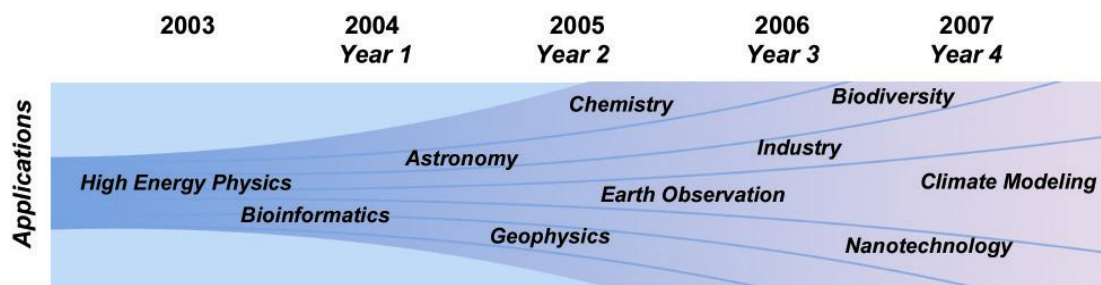


Figure 3: Schematic evolution of application communities benefiting from EGEE infrastructure. (Ack. the EGEE Consortium)

3.4 Conclusion

At the conclusion of the first half of the GridPP project, now is an appropriate time to ask, “Has GridPP built a real Grid?” Ian Foster devised a three point checklist for defining a Grid [22]: “A Grid is a system that - (1) Coordinates resources that are not subject to centralized control. (2) ... using standard, open, general-purpose protocols and interfaces (3) ... to deliver nontrivial qualities of service”.

The first of these is clearly met since the resources are owned and controlled by the participating institutes and not by CERN or GridPP. The second is certainly met by most of the middleware components coming from the Virtual Data Toolkit (Globus, Condor-G etc) and from EDG/EGEE which are both Open Source projects. The third is perhaps more subjective but the experiments have certainly conducted non-trivial data challenges over the Grid. CMS generated 75M events comprising 150 TB of data, ATLAS generated 7.7M events and 22TB of data while LHCb generated 186M events, for which over a quarter of the CPU was in the UK.

In conclusion, during the prototype stage, GridPP has demonstrated that the Grid computing paradigm works and has built a Grid of considerable size and complexity that is being routinely used by the experiments to generate significant amounts of simulated data. The challenge for the next three years is to expand the prototype Grid to full size with the necessary robustness and security in order to be ready for the first data from the LHC so that the full physics potential of this exciting new accelerator can be realized.

-
- [1] Berners-Lee T and Fischetti M 1997 *Weaving the Web* (Harper, San Francisco, USA)
- [2] MONARC Site and Network Architecture Working Group 1999 *MONARC Report on Existing Computing Architectures*
- [3] LHC Study Group 1995 *The Large Hadron Collider conceptual design*, CERN/AC/95-05
- [4] ATLAS Collaboration 1994 *ATLAS Technical Proposal* CERN/LHCC/94-43
- [5] CMS Collaboration 1994 *CMS Technical Proposal* CERN/LHCC/94-38
- [6] LHCb Collaboration 1998 *LHCb Technical Proposal* CERN/LHCC/98-4
- [7] ALICE Collaboration 1995 *ALICE Technical Proposal* CERN/LHCC/95-71
- [8] Campanella M, Perini L 1998 The analysis model and the optimization of geographical distribution of computing resources: a strong connection, *MONARC note* 1/98
- [9] Bethke S, Calvetti M, Hoffmann H F, Jacobs D, Kasemann D, Linglin D 2001 *Report of the Steering Group of the LHC Computing Review* CERN/LHCC/2001-004
- [10] LHC Collaboration 2005 *LHC Computing Grid Technical Design Report* CERN/LHCC/2005-024
- [11] Greenberger M 1964 The Computers of Tomorrow, *The Atlantic Monthly*, May 1964.
- [12] Foster I, Kesselman C, eds. 1999 *The Grid: Blueprint for a New Computing Infrastructure* (Morgan Kaufmann, San Francisco, USA)
- [13] CERN Council 2001 *Proposal for building the LHC computing environment at CERN* CERN/2379/Rev
- [14] Foster I and Kesselman C, 1997 *Int. Journal of Supercomputer Applications* **11(2)** 115-128
- [15] Tannenbaum T, Wright D, Miller K, and Livny M 2001 Condor: A distributed job scheduler. In Thomas Sterling, editor, *Beowulf Cluster Computing with Linux* (MIT Press, Cambridge, USA)
- [16] www.cs.wisc.edu/vdt
- [17] Burke S *et al* 2004 HEP Applications Experience with the European DataGrid Middleware and Testbed *Proceedings of CHEP2004 (Interlaken Switzerland, 27 September – 1 October 2004)*
- [18] <http://www.eu-egee.org/>
- [19] Richards A *et al* 2004 *Proceedings of the UK e-Science All Hands Meeting (Nottingham UK, 31 August - 3 September 2004)* 210-217
- [20] See, for example, <http://www.uscms.org/s&c/dc04/>
- [21] Strickland D 2004 CMS Status for GridPP Presented at *GridPP10 (CERN, Geneva, 2-4 June 2004)*
- [22] Foster I 2002 What is the Grid: A Three Point Checklist *Grid Today* July 20, 2002