LHC Computing Grid

Distributed Production Environment for Physics Data Processing

www.cern.ch/lcg

The Large Hadron Collider (LHC), currently being built at CERN near Geneva, is the largest scientific instrument on the planet. When it begins operations in 2007, it will produce roughly 15 Petabytes (15 million Gigabytes) of data annually, which thousands of scientists around the world will access and analyse. The mission of the LHC Computing Grid (LCG) project is to build and maintain a data storage and analysis infrastructure for the entire high energy physics community that will use the LHC.

The data from the LHC experiments will be distributed around the globe, according to a four-tiered model. A primary backup will be recorded on tape at CERN, the "Tier-0" centre of LCG. After initial processing, this data will be distributed to a series of Tier-1 centres, large computer centres with sufficient storage capacity for a large fraction of the data, and with round-the-clock support for the Grid.

The Tier-1 centres will make data available to Tier-2 centres, each consisting of one or several collaborating computing facilities, which can store sufficient data and provide adequate computing power for specific analysis tasks. Individual scientists will access these facilities through Tier-3 computing resources, which can consist of local clusters in a University Department or even individual PCs, and which may be allocated to LCG on a regular basis.

How to participate in the LCG project

The LCG project encourages participation by academic institutes in the field of high energy physics with suitable computing infrastructures, as well as other public and private data centres willing to commit resources to the project.

Support centres are available in the following countries and regions:

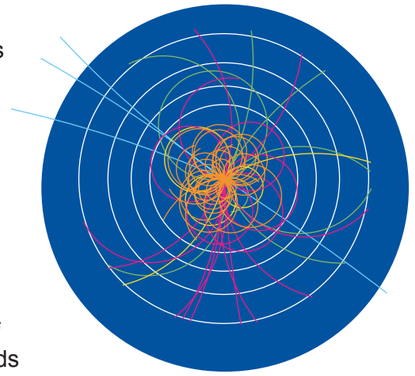
Australia, Austria, Belgium, Brazil, Canada, China, Czech Republic, Denmark, Finland, France, Germany, Hungary, India, Israel, Italy, Japan, Netherlands, Norway, Pakistan, Poland, Portugal, Romania, Russia, Slovenia, Spain, Sweden, Switzerland, Taiwan, Ukraine, United Kingdom, U.S.A

To contact one of these centres please send an email to

lcg.office@cern.ch

LCG Dissemination - January 2006

The goals of the LCG project



Discovering new fundamental particles and analysing their properties with the LHC accelerator is possible only through statistical analysis of the massive amounts of data gathered by the LHC detectors ALICE, ATLAS, CMS and LHCb, and detailed comparison with compute-intensive theoretical simulations. LCG provides tools and computing services to help the research physicists in their work. The goals of the LCG project include:

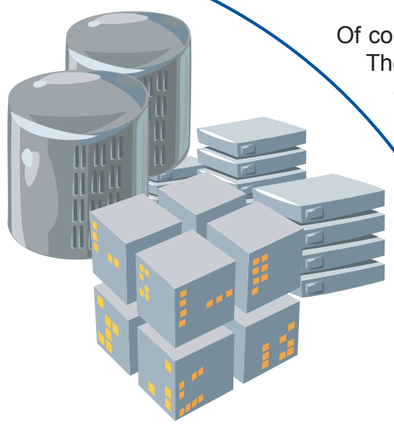
- Developing components to support the physics application software, such as general-purpose scientific libraries, tools for storing, managing and accessing physics data, adaptation of general scientific simulation frameworks and toolkits to the needs of LHC, interfacing batch and interactive analysis systems to a Grid environment.
- Developing and deploying computing services based on a distributed Grid model using resources from more than a hundred computing centres around the globe.
- Managing users and their rights in an international, heterogeneous and non-centralized Grid environment.
- Collaborating with the regional and national education and research network organisations (NRENs) to ensure high bandwidth data paths between the major LCG centres.
- Coordinating the programme of tests and pilot services for commissioning of the LCG service in 2007.

Rationale for using a Grid to process LHC data

When the LHC accelerator is running optimally, access to experimental data needs to be provided for the 5000 scientists in some 500 research institutes and universities worldwide that are participating in the LHC experiments. In addition, all data need to be available over the 15 year estimated lifetime of the LHC. The analysis of the data, including comparison with theoretical simulations, requires of the order of 100 000 CPUs at 2005 measures of processing power. A traditional approach would be to centralize all of this capacity at one location near the experiments. In the case of the LHC, however, a novel globally distributed model for data storage and analysis – a computing Grid – was chosen because it provides several key benefits. In particular:

- The significant costs of maintaining and upgrading the necessary resources for such a computing challenge are more easily handled in a distributed environment, where individual institutes and participating national organisations can fund local computing resources and retain responsibility for these, while still contributing to the global goal. As the use of Grid technology grows these resources can be provided from Grid infrastructures shared with other sciences.

- Also, in a distributed system there are no single points of failure. Multiple copies of data and automatic reassigning of computational tasks to available resources ensures load balancing of resources and facilitates access to the data for all the scientists involved, independent of geographical location. Spanning all time zones also eases round-the-clock monitoring and support.



Of course, a distributed system also presents a number of significant challenges. These include ensuring adequate levels of network bandwidth between the contributing resources, maintaining coherence of software versions installed in various locations, coping with heterogeneous hardware, managing and protecting the data so that it is not lost or corrupted over the lifetime of the LHC, and providing accounting mechanisms so that different groups have fair access, based on their needs and contributions to the infrastructure. These are some of the challenges that the LCG project is addressing.

LCG in context

The major computing resources for LHC data analysis are provided by the Worldwide LHC Computing Grid Collaboration – comprising the LHC experiments, the accelerator laboratory and the Tier-1 and Tier-2 computer centres. The LCG project manages the deployment and operation of the distributed computing services for LHC on behalf of the collaboration.

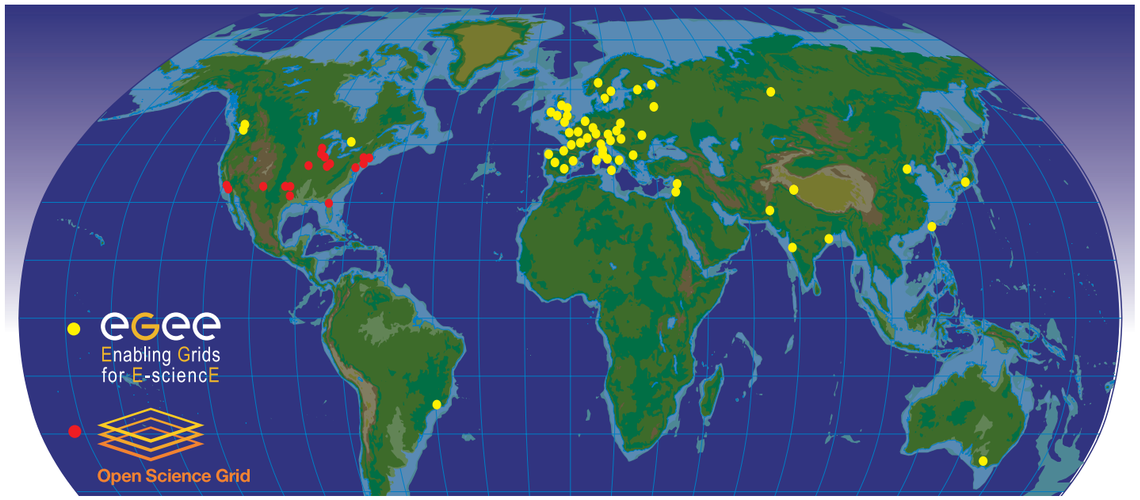
The computing centres providing resources for LCG are embedded in different operational Grid organisations, in particular EGEE (Enabling Grids for E-SciencE) and OSG (the Open Science Grid), but also several national and regional Grid structures such as GridPP in the UK, INFN Grid in Italy and NorduGrid in the Nordic region.

EGEE (Enabling Grids for E-SciencE). LCG is the primary production environment for this project, which started in April 2004 and aims to establish a Grid infrastructure for European science. With 70 partners from Europe, the US and Russia, EGEE is leading a worldwide effort to re-engineer existing Grid middleware, including that developed by the European DataGrid (EDG) project, to ensure that it is robust enough for production environments like LCG.

OSG (Open Science Grid) is a national production-quality Grid computing infrastructure for large scale science, built and operated by a consortium of U.S. universities and national laboratories. The OSG Consortium was formed in 2004 to enable diverse communities of scientists to access a common Grid infrastructure and shared resources.

The **Globus Alliance** involves several universities and research laboratories conducting research and development to create fundamental Grid technologies and produce open-source software. The LCG project is actively involved in the support of Globus and uses the Globus-based Virtual Data Toolkit (VDT) as part of the project middleware.

The LCG project is also following developments in industry, in particular through the **CERN openlab for DataGrid applications,** where leading IT companies are testing and validating cutting-edge Grid technologies using the LCG environment.



A map of the worldwide LCG infrastructure operated by EGEE and OSG.