An open source Grid for science

Dr Sarah Pearce charts the development of GridPP2, the next stage in the Grid technology research…

Over the next three years, UK physicists will develop a computing Grid equivalent to the world's second largest supercomputer. GridPP2 is the UK's contribution to an international collaboration by particle physicists, aimed at analysing an imminent 'data deluge' from CERN, the European Laboratory for Particle Physics.

In 2007, CERN will introduce the Large Hadron Collider (LHC). LHC will allow scientists to penetrate further into the structure of matter and recreate the conditions prevailing in the early universe, just after the 'Big Bang'. In the search for answers to questions such as why particles have mass, the LHC's detectors will produce 10 petabytes of data each year – equivalent to a stack of CDs twice the height of Mount Everest.

Rather than deal with this data on expensive super-computers, based at a few institutions and in high demand, LHC will use a computing Grid. More than 70,000 PCs, spread out at 100 institutions across the world, will allow scientists from different countries to access the data, analyse it and work together in effective international collaborations. Whereas a PC using the web provides information or access to services, such as banking or shopping, a PC on the Grid offers its computing power and storage. Individual scientists using the Grid won't need to know where the data is held or which machines are running their programs.

Physicists and computer scientists in the UK and CERN have been working on the original GridPP1 project for the past three years, setting up a prototype Grid. Known as a 'testbed', 17 UK sites and more than 100 computers have been used to develop the key components of the Grid. It is linked to other Grid testbeds worldwide, and has been tested by analysing data from US particle physics experiments in which the UK is involved. The GridPP1 testbed has therefore enabled UK particle physics to progress 'from web to Grid'.

In December, the Particle Physics and Astronomy Research Council announced £16m extra funding for the GridPP2 project. This next stage will extend the testbed across the equivalent of 20,000 1GHz personal computers in the UK.

Three years from now, GridPP2 will allow physicists to use the Grid in their day to day work, submitting complex computing jobs to be run much faster than would otherwise be possible. In this way, GridPP2 is moving 'from prototype to production'.

GridPP has computing hardware at 17 UK institutions. The Rutherford Appleton Laboratory near Oxford is the GridPP Tier-1 centre. Based around a large existing data tape store, it is providing a major contribution to both computer resources and staff. The centre also acts as a starting point for a worldwide Grid Operations Centre that will monitor the status of all the centres and the progress of thousands of jobs distributed around the world at any given time. The other GridPP institutions are grouped into four Tier-2 centres – London, NorthGrid, SouthGrid and ScotGrid – through which they are integrated into the Grid.

'Middleware' is the key to a successful Grid, and much of GridPP2's effort will be spent in this area. Middleware allows the software being used by the scientists to talk to the Grid's hardware, distributing computing jobs efficiently around the network. It also deals with issues such as security, ensuring that only authorised users can access the Grid. Middleware consists of many different computer programs, conducting the machine-to-machine negotiations that enable the smooth running of the Grid. Programs can act as 'agents', presenting 'metadata' about who is using the Grid, what data they need and what resources are available. Other programs are 'brokers', striking deals between machines for access to data and resources, and breaking data up into small packages for processing. Still further programs do the 'housekeeping', monitoring quality of service and routing the data around the network.

For security, GridPP uses digital certificates issued by the UK e-Science Certification Authority, and other European equivalents. The Certification Authority checks the identity of users, providing them (and computer hosts) with electronic certificates that can authenticate them to the Grid. Based on this authentication, users can then be authorised (or not) for a range of tasks through their

Computer generated image of the LHC tunnel

membership of a 'virtual organisation' (VO). For example, physicists working on one of the LHC experiments might belong to that VO, and so be allowed to read and work with its data. GridPP has also used this concept to produce a web service called GridSite, that can let users edit a website remotely, using their certificates as authentication.

'Middleware allows the software being used by the scientists to talk to the Grid's hardware, distributing computing jobs efficiently around the network. It also deals with issues such as security, ensuring that only authorised users can access the Grid.'

In September 2003, GridPP's testbed was incorporated into the LHC Computing Grid – the first worldwide production Grid. GridPP2 will also become part of a European initiative called Enabling Grids for E-science in Europe (EGEE). This aims to support the EU's European Research Area by bringing together local and national Grids, and those for different types of science. As well as particle physics, many other disciplines are developing Grid technology for high performance computing – ranging from bioinformatics and climate simulation to the nanoscale design of new materials and the integration of large engineering projects involving many partners. EGEE will attempt to integrate all these Grids, creating a tool for e-science in Europe and operating with other Grids around the globe.

Within the UK, GridPP2 will also collaborate with other parts of the UK's e-science programme, such as AstroGrid for astronomy. Many of the tools developed by GridPP could be useful for other disciplines – for example, GridPP is working with clinical researchers on the potential for using its computer security tools in the health service. In addition, many companies in the computing industry have made their interest in Grid technologies clear, both for large computing problems and to make more effective use of their existing computers. GridPP is open to opportunities to work with industry and discuss experience of current Grid development issues and solutions adopted.

This is an exciting time for Grid computing, and for GridPP in particular – indeed, it was recently cited in the Daily Telegraph as one of the key scientific projects to watch in 2004. We are moving from developing a prototype, towards making a useful, functioning tool for e-science collaborations. For the next three years and beyond, GridPP will help the UK make the most of new opportunities, both in fundamental physics and more widely in science and industry.

Dr Sarah Pearce
GridPP Dissemination Officer
Physics Department
Queen Mary
University of London
Mile End Rd
London E1 4NS
UK
Tel. +44 (0)20 7882 5049
Fax. +44 (0)20 8981 9465
s.pearce@qmul.ac.uk
www.gridpp.ac.uk