

Grid Physics, the Virtual Data Grid, and LIGO

Patrick Brady (University of Wisconsin–Milwaukee) and
Manuela Campanelli (The University of Texas at Brownsville)

Between 28 December 2001 and 14 January 2002, the three largest interferometric gravitational-wave detectors in the world were listening for signatures of cataclysmic astrophysical events in our Galaxy and beyond. For many involved in the LIGO (Laser Interferometer Gravitational-wave Observatory) project, this was an event of grand proportions which demonstrated that we are truly on the brink of gravitational-wave astronomy. Yet the data run was just the beginning. About 13 Terabytes of data was recorded and will be analyzed over the coming months. Scaling these numbers to full scale scientific operations, the experiment will generate several hundreds of Terabytes of data per year.

The variety of sources which could produce gravitational waves call for a variety of search techniques to be applied to the data stream. For example, searches for stochastic background gravitational waves require minimal computational power—a standard desktop workstation is good enough—yet, the search must access the data from all the detectors. At the other end of the spectrum of computational requirements are searches for continuous signals from spinning neutron stars. These signals are extremely weak, and require coherent accumulation of signal-to-noise for long periods of time; this introduces the need to account for earth-motion-induced Doppler shifts and internal evolution of the sources. Thus, the variety of signals and their weakness lead to an analysis problem which can use essentially infinite computing resources.

But the computational requirements are only half the story. LIGO computing facilities and scientific users reside at many different national and international centers and universities. For the LIGO Scientific Collaboration (LSC) [2], therefore, accessing these large datasets and performing an efficient analysis on them requires a dynamically distributed computational infrastructure, including tools to manage storage, migration and replication of data, job control, and cataloging of the many data products. The LIGO Data Analysis System handles these problems on a scale consistent with the LIGO-I mission, however *Grid Computing* [3] provides a new computational infrastructure to extend and enhance current capabilities to a level consistent with the expected requirements.

LIGO is only one of several physics experiments expecting to generate vast amounts of data which must be carefully analyzed using complex algorithms requiring enormous computational power. For this reason several LSC member

institutions, including California Institute of Technology (CIT), The University of Texas at Brownsville (UTB), and University of Wisconsin–Milwaukee (UWM), are participating in a multi-experiment project, sponsored by the National Science Foundation, to build the first Petabyte-scale computational grid environment for data intensive experiments. The Grid Physics Network (GriPhyN) project is a collaboration of both experimental physicists and information technology researchers. Driving the project are unprecedented requirements for geographically dispersed extraction of complex scientific information from very large collections of measured data, flowing from four experiments in high-energy and nuclear physics (two large hadron colliders at CERN, CMS and ATLAS [4]), gravitational waves (LIGO [1]) and astronomy (the SDSS project [5]). To meet these requirements, GriPhyN researchers will develop the Virtual Data Toolkit (VDT) containing basic elements to construct the first Global Petascale Virtual Grid.

The virtual data concept aims to unify the view of data in the distributed Grid environment. It will not matter if the data is raw or processed, or if it was generated from an hadron collider experiment or a gravitational-wave detector. The virtual Grid will enable data access and archival at nodes distributed around the globe while storing meta-data which make the data self-describing. The VDT developed by GriPhyN will be deployed on the Grid to directly manage these fundamental virtual data objects instead of complex data pipelines. In the case of LIGO, the VDT will be capable of executing deep searches for gravitational waves using many machines distributed around the world, while making the results available to the scientists in a transparent fashion. Once deployed, the Grid tools currently under development will significantly enhance scientists' ability to carry out the necessary analysis of LIGO data. In fact, prototype data replication tools (being developed by Scott Koranda at UWM) are already moving data from the archive at Caltech onto spinning disks at UWM for analysis using the UWM system.

GriPhyN is not the only data grid project, although it is one of the largest and probably most advanced in the world. Similar projects are now active in Europe and Asia. In September, the NSF announced the additional award of \$13.65M over five years to a consortium of 15 universities and four national laboratories to create the International Virtual Data Grid Laboratory [7, 8] (iVDGL). The iVDGL, to be constructed in partnership with the European Union, Japan, Australia and eventually other world regions, will form the world's first true *Global Grid*. The iVDGL will provide a unified computational resource for major scientific experiments in physics, astronomy, biology, and engineering. The iVDGL will therefore serve as a unique computing resource for testing new GriPhyN computational paradigms at

the Petabyte scale and beyond. Management of the iVDGL is integrated with that of the GriPhyN project. The international partners are investing more than \$20M around the world to build computational sites as part of the consortium. Moreover, the NSF award of iVDGL is matched by \$2M in university contributions, plus funding for Computer Science Fellows by the UK e-Science Programme [9]. Of this total award, \$2.11M will go to universities affiliated with the LIGO Laboratory to develop Grid Computing centers at the three GriPhyN/LSC institutions (CIT, UWM and UTB) and at Pennsylvania State University (PSU).

A significant challenge for science in the 21st century is data management and analysis. Just as large database technology has revolutionized the commercial world as the backbone of many information intensive enterprises, so virtual data, Grid computing and transparent access to a world of computing resources will revolutionize science in the coming decade.

References

- [1] The Laser Interferometer Gravitational-wave Observatory web site: see <http://www.ligo.caltech.edu>.
- [2] LIGO Scientific Collaboration web site: <http://www.ligo.caltech.edu/LIGOWeb/lsc/lsc.html>.
- [3] The Computational Grid is described in the book “The Grid : Blueprint for a New Computing Infrastructure” edited by Ian Foster and Carl Kesselman – Morgan Kaufmann Publishers (1998) ISBN 1-55860-475-8. Many more references can be found at the following web site: <http://www.aei-potsdam.mpg.de/~manuela/GridWeb/info/grid.html>.
- [4] CMS and ATLAS are two large hadron colliders at CERN, the world’s largest particle physics center near Geneva in Switzerland (see web site: <http://www.griphyn.org/info/physics/high.html>.)
- [5] The Sloan Digital Sky Survey (SDSS) project (see web site at: <http://www.sdss.org/sdss.html>) is the most ambitious astronomical survey project ever undertaken. The survey will map in detail one-quarter of the entire sky, determining the positions and absolute brightnesses of more than 100 million celestial objects. It will also measure the distances to more than a million galaxies and quasars. Apache Point Observatory, site of the SDSS telescopes, is operated by the Astrophysical Research Consortium (ARC).

- [6] The Grid Physics Network web site: <http://www.griphyn.org>.
- [7] The International Virtual Data Grid Laboratory web site: <http://www.ivdgl.org>.
- [8] The Outreach Center of the Grid Physics Network web site: <http://www.aei-potsdam.mpg.de/~manuela/GridWeb/main.html>.
- [9] e-Science is an equivalent project to iVDGL in the UK (see web site at: <http://www.e-science.clrc.ac.uk/>).