Contribution to the EAC Meeting Report by the LIGO-GriPhyN Working Group.

Amit Agarwal Shishir Bharathi, James Blythe, Ewa Deelman, Yolanda Gil, Carl Kesselman, Gaurang Mehta, Sonal Patil, Srividya Rao, Gurmeet Singh, Marcus Thiebaux, Karan Vahi,
Information Sciences Institute, USC

Stuart Anderson, Kent Blackburn, Masha Barnes, Philip Charlton, Phil Ehrens, Ed Maros, Greg Mendell, Albert Lazzarini, Mary Lei, Isaac Salzman, Peter Shawhan,
California Institute of Technology

Scott Koranda, *University of Wisconsin Milwaukee and NCSA*

Maria Alessandra Papa, *Albert Einstein Institute, Germany*

Alicia Sintes, *Balearic Islands University, Spain*

Accomplishments in FY2002

The Laser Interferometer Gravitational Wave Observatory (LIGO) is making great strides towards performing scientifically significant data analysis using Grid resources. As part of the collaboration between the LIGO and the GriPhyN projects, we have focused on a specific LIGO problem: the pulsar search (shown below). The data needed to conduct the search is a long stretch (~4 months, $2x10^{11}$ points) of a single channel—the gravitational wave strain channel. The power spectra of the small segments are stacked to make a large frequency-time image, perhaps $4x10^5$ on each side. The pulsar search consists of searching for coherent signals in this image. A source would appear on the frequency-time image as a wavering line, whose frequency might be 1 kHz, but modulated by a few parts in 10^6 over periods of 1 day and a few parts in 10^4 over periods of 1 year. In addition, if the source exhibits any secular variations due to slowing down of its rotational period, these will be encoded in the data, as well.

The pulsar search is very compute and data intensive and requires more resources than those available in the LIGO Scientific Collaboration. In order to take advantage of the Grid resources, we had to integrate LIGO's existing data analysis into the Grid environment. The LIGO Data Analysis System (LDAS) can perform a wide range of sophisticated and computationally intensive data analysis. We have developed an infrastructure in which LDAS can be accessed as a Grid resource, and in which LDAS can schedule its jobs on the Grid. To achieve integration of LDAS into the Grid environment, we have designed and implemented Globus interfaces both for data staging (via GridFTP) and computation scheduling (via GRAM). These interfaces also make use of the Globus Security Infrastructure to provide secure access to resources.

The pulsar search conducted at SC 2002 used LIGO's data collected during the first scientific run of the instrument and will target a set of 1000 locations of known pulsar as

well as random locations in the sky. The results of the analysis were published via LDAS to the LIGO Scientific Collaboration and thus will become accessible to LIGO scientists.

Pegasus, which stands for Planning for Execution in Grids, was developed at ISI as part of the GriPhyN project. Pegasus is a configurable system that can map and execute complex workflows on the Grid. Pegasus has been integrated with the GriPhyN Chimera system. To support the LIGO pulsar search, Pegasus was also be configured to perform the generation of the abstract workflow based on application-level metadata attributes. Given attributes such as time interval, frequency of interest, location in the sky, etc., Pegasus is currently able to produce any virtual data products present in the LIGO pulsar search.

Pegasus uses the Metadata Catalog Service (MCS), newly developed at ISI, to perform the mapping between application-specific attributes and logical file names of existing data products. AI-based planning technologies are used to construct both the abstract and concrete workflows. Pegasus also contains a Virtual Data Language generator that can populate the Chimera catalog with newly constructed derivations. Finally, Pegasus generates the necessary submit files and sends the CW to DAGMan for execution. The LIGO pulsar search is performed using the Pegasus system, LDAS and compute and storage resources at Caltech, University of Southern California, University of Wisconsin Milwaukee, University of Florida, and NCSA. During SC 2002, we conducted over 58 pulsar searches which resulted in a total of 330 tasks, 469 data transfers executed and 330 output files being produced. The total runtime was 11:24:35.

Plans for FY2003

During the coming year, there are three general areas of work that we plan to address. The scope of the effort will exceed the currently foreseeable FTE level, so they need to be prioritized in the near future. The LIGO-GriPhyN applications development group is planning a face-to-face meeting in late February. The general areas of research are listed below.

1. We will take the prototype code developed for the SC2002 demonstration project and produce from it a robust, stable utility that will be used by LIGO scientists to stage analyses on systems outside LIGO Laboratory's LDAS environment. There is a second class of search that has an immediate use and need for this, namely, the search for a gravitational wave stochastic background. Many autonomous jobs need to be run over the same data to obtain on-source/off-source analyses using time shift and FFT techniques. In addition, the periodic sources search that was showcased for SC2002 needs to be rendered useful for a larger number of scientists.

2. Additional API development is needed to enable a better (i.e., more seamless) integration between LIGO's LDAS and the globus grid toolkit. The specific area of effort that will be targeted is addressing the interoperation of the authentication schemas.

3. A different but important effort will be to develop better and more extensive data replication services for LIGO and its international partners. Part of this effort will involve cooperation among US and European groups in Germany, UK, Italy, and France. There is at present a simple, non-robust infrastructure based on unix utilities (principally shell scripts and rsync) to replicate and move data subsets among the different projects. This infrastructure is termed “DAS” for network data analysis servers. We want to revisit this in the context of grid-based data transport and replication. Requirements need to be identified, a specification for the new infrastructure need to be developed and then the integration of existing grid tools needs to take place to replace “NDAS”. A successful execution of this task would be a model of iVDGL cooperation from the GW community.