CIF21 DIBBs: Domain-Aware Management of Heterogeneous Workflows Active Data Management for Gravitational-Wave Science

PI: Duncan Brown[1], Co-PIs: Ewa Deelman[2] and Jian Qin[3]

NSF Award ACI-1443047

[1] Physics Department, Syracuse University. [2] USC Information Sciences Institute. [3] iSchool, Syracuse University.

Large-scale scientific workflows are essential to LIGO's discoveries. To detect gravitational waves, LIGO data must be filtered through hundreds of thousands of signal models. This is repeated many times using simulated signals to measure the search's efficiency and to diagnose and fix problems with the detectors. Searches are also run multiple times to tune the scientific parameters for maximum sensitivity. Analyses are run by teams of scientists in distributed locations and are executed using heterogeneous computing environments, including in-house resources, the Open Science Grid, and XSEDE HTC resources.

LIGO was an early adopter of the Pegasus Workflow Management System (WMS) and HTCondor for its binary black hole searches. This project builds on the widely-used Pegasus WMS to address the problems encountered in large-scale, distributed science analysis. Based on LIGO scientist interviews, we prioritized the following developments:

- Hardening Pegasus' existing data re-use capabilities based on simple metadata (e.g. file URIs) and providing simple ways to integrate re-use with scientific workflow generation.
- Improving Pegasus Stampede Dashboard for visualization of workflow status and progress, and providing tools to integrate Dashboard into scientific workflows and results.
- Implementation cataloging of metadata as part of workflow execution in Pegasus WMS and use of Dashboard to provide file and metadata information to users.
- Development of an initial metadata model for gravitational-wave science.

These developments have already had significant impact on LIGO's search for gravitational waves. This work was used in the publications reporting the discoveries of the binary black hole mergers GW150914, GW151226, and LVT151012.

The significance of LIGO's discoveries demanded thorough review of all aspects of the detection. The PyCBC search produced the statement that the events were observed with $> 5\sigma$ significance. PyCBC generates workflows that are run by Pegasus WMS and uses the developments enabled by this project. Integration of the scientific result pages with workflow information significantly streamlined the review process. Our data management and re-use capabilities were used to combine together individual two-week analyses of LIGO data and generate the final results from LIGO's three-month observing run. Data re-use capabilities are now regularly used to re-run the analysis with different configurations to tune the search and to demonstrate the robustness of LIGO's detections.

Additional development to the Pegasus WMS now allows users to associate metadata with: (i) the workflow itself and for any sub-workflows; (ii) individual tasks in the workflow; and (iii) individual files produced by tasks in the workflow. Metadata is specified as a key value tuple, where both key and values are strings. Metadata is populated into the Stampede database for the workflow by Pegasus' workflow monitoring daemon. Users can identify static metadata attributes at workflow creation time that are populated automatically as the workflow executes. Pegasus also automatically captures metadata as output files are generated and associates them at the file level in the database.

In the final year of this project, we are completing Pegasus metadata integration into PyCBC, which will allow workflows to automatically identify and re-use intermediate data products that remain valid from prior analysis, while re-running only the parts of the workflow needed.