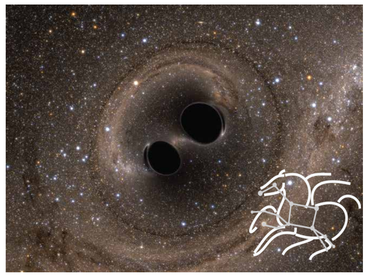


CIF21 DIBBs: Domain-Aware Management of Heterogeneous Workflows Active Data Management for Gravitational-Wave Science

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.

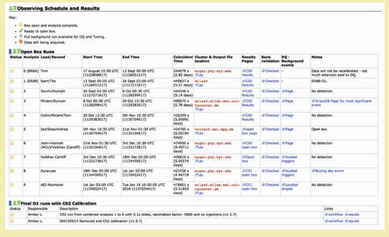
Gravitational Waves and Workflows

The direct detection of gravitational waves by the Advanced Laser Interferometer Gravitational-wave Observatory (LIGO) marks a transformative moment in 21st century science. LIGO detected two binary black hole mergers during its first observing run (September 2015 - January 2016). These detections captured the imagination of the public and provided scientists with a completely new way of studying the universe.

Large-scale scientific workflows are essential to LIGO's discoveries. LIGO was an early adopter of the Pegasus Workflow Management System and HTCondor for its binary merger searches. These systems have allowed scientists to mange increasingly complicated data-analysis workflows, but limitations have impeded efficient collaboration between distributed teams of scientists. LIGO scientists have struggled to discover, share, reuse, and verify the data products from their analyses. These problems are common in many scientific domains.

This project builds on the widely-used Pegasus WMS to address these problems. We use LIGO's search for binary mergers to identify and test solutions to the challenges of managing data and metadata in large-scale heterogeneous workflows.

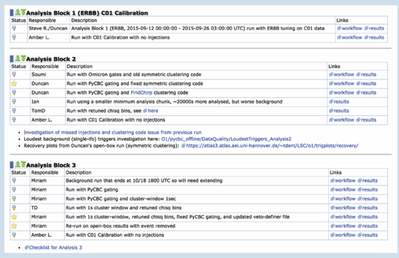
LIGO scientists have historically used wiki pages to track the status and provenance of scientific results. The image to the right shows the page tracking the results of LIGO's binary merger search in the first observing run. These pages are simple for domain scientists to create and maintain, but human error can cause them to be incomplete or to get out of date. The final scientific results of a search are displayed on static web pages linked from the wiki.



Motivation and Research Problem

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. These analyses are performed by teams of scientists in distributed locations and are executed using heterogeneous computing environments.

Domain scientists often do not see workflow and data management as part of "doing science" and may not recognize the benefits that infrastructure development can bring to them.



The complexity of computationally-intensive science poses challenges for data management. Motivations for keeping documentation of data and analysis results include trust, accountability, and continuity of work. Research reproducibility relies on metadata that represents code dependencies and versions and has good documentation for verification. The wiki page at left illusttrates the number of re-runs needed in a typical LIGO analysis. Individual workflows may contain hundreds of thousands of tasks.

We conducted interviews with LIGO scientists to determine what metadata needs to be captured to track, reproduce, and reuse computational results. Above all, scientitsts reported that:

- It was difficult to reuse results of previous analyses (even their own) and often entire workflows were re-run when there were minor parameter or configuration changes.
- Metadata for gravitational-wave data, workflows, and outputs tend to differ from those currently available in metadata standards. Metadata for scientific analyses can evolve rapidly, especially when the analysis is under active development.
- Scientists were frustrated with their ability to track the status and progress of workflows in HTCondor and typically used command line tools, rather than GUI tools for this task.

These findings were used to inform our metadata and infrastructure development activities.

Successes and Milestones

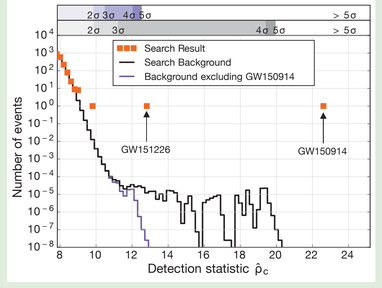
Based on LIGO scientist interviews, we prioritized and completed the following developments:

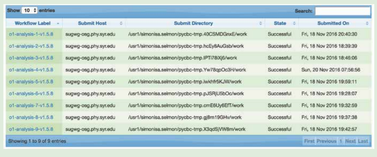
- Hardening Pegasus' existing data re-use capabilities based on simple metadata (e.g. file URIs) and providing simple ways to integrate re-use with scientific workflow generation.
- Improving Pegasus Stampede Dashboard for visualization of workflow status and progress, and providing tools to integrate 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 black holes and the ability of LIGO scientists to reliably use the resources of the Open Science Grid (OSG).

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 > 5σ 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. The image at top right shows the workflow section from a result page which links to the Pegasus Dashboard.

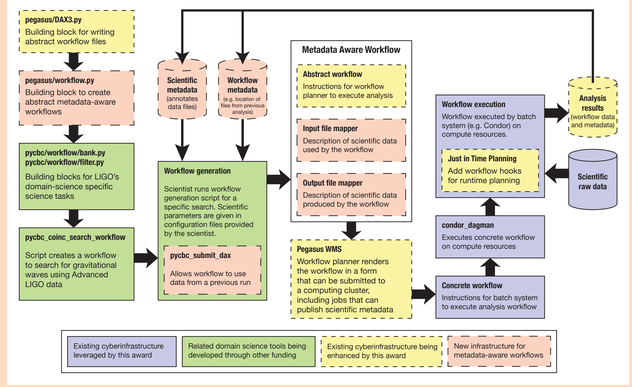
LIGO analyses are run in two-week blocks, which are combined to produce a final result. Our data management and re-use capabilities were used to combine together the individual analyses to generate the final results for the observing run (shown lower right with the discovery of GW150914 and GW151226), as well as to re-run the analysis with different configurations to tune the search and to demonstrate the robustness of LIGO's detections.





PyCBC searches are now cataloged in the Stampede Database. This is currently performed on a per-user level with simple metadata. No user expertise is required to manage this. The image at left shows the database of an undergraduate student who tested search improvements using the OSG.

Integration of Metadata into Pegasus Workflow Management



The Pegasus WMS now allows users to associate metadata with:

- The workflow itself and for any sub-workflows.
- Individual tasks in the workflow.
- 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. This development is ahead of schedule.

An initial metadata model has been developed based on user interview and the PyCBC workflow generation toolkit has been extended to allow the use of Pegasus metadata. We are currently working on extending LIGO's compact binary search workflows to add the metadata from our model into the workflow. The final stage is to integrate this together with our data re-use capabilities. The system design diagram shown at left illustrates how the components of our project integrate together.

Summary, Challenges, and Future Plans

The cyberinfrastructure developed has already had an impact on LIGO's ability to detect gravitational waves. The enhancements to Pegasus WMS and the Stampede Dashboard allow users to manage their analyses and re-use the components of prior searches. All production PyCBC searches now create Stampede Dashboard pages (such as the one shown at right) without any need for user intervention. The Stampede Databases have been tested at scale in LIGO's first observing run with workflows ranging from tens to hundreds of thousands of independent, heterogeneous tasks. These developments have been released to the community in Pegasus 4.6 and 4.7.

One unexpected (and happy) challenge we encountered was the discovery of gravitational waves in the second year of this award. This has delayed integration of Pegasus' new metadata features and our metadata model as PyCBC development was frozen for six months.

In the final year of this project, we are completing 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.