

Big Data: Enabling Big Science

The ability to manipulate huge data sets (terabytes and even petabytes in size) has become a key enabler of advances in science. Climate scientists can run models at resolutions that were simply unthinkable in the past, and they need even higher resolutions. Large scientific instruments, such as the Large Hadron Collider (LHC) at CERN, the Laser Interferometer Gravitational Wave Observatory (LIGO), and the Sloan Digital Sky Survey (SDSS) are producing terabytes per day, and in the case of the LHC are projected to produce several petabytes per year. Manipulation of data sets this size requires secure, scalable, and robust middleware tools. The DOE SciDAC DataGrid Middleware Project has produced the tools that are the de facto standard for Grid projects the world over for these types of problems.

The SciDAC Earth Science Grid (ESG) project uses the Replica Location Service for tracking the locations of replicated data and GridFTP for movement of that data. Don Middleton, PI for the ESG project had this to say:

"GridFTP has been integral to the success of ESG, which uses it at several levels in order to manage and deliver climate research data to a worldwide community. Not only does the tool provide us with a reliable, high-performance transfer capability, it also delivers a secure one. ESG provides access to data that is stored on numerous systems at several sites, and in today's environment of much-heightened security requirements we could not be operating without GridFTP. Behind the scenes, GridFTP provides the data transport fabric that invisibly and reliably serves our users"

Dr. Scott Koranda and his team within the LIGO project have built higher level services that use the RLS and GridFTP and hope to leverage future development as well:

Every single bit (literally) of LIGO data is being replicated from the LIGO observatory sites to the main data archive at Caltech using Globus GridFTP. In addition other data products (files containing downsampled or fewer data channels) are being generated at the LIGO sites and replicated to Caltech, so that the overall data transfer rate to Caltech with GridFTP is well over 1 TB per day.

Further the reduced data sets are being replicated from Caltech to the University of Wisconsin-Milwaukee (UWM), The Pennsylvania State University (PSU), the Massachusetts Institute of Technology (MIT), and the Albert Einstein Institute (AEI) in Potsdam, Germany. All this is being done with Globus GridFTP. Likewise, the GEO data is being replicated from Germany to UWM, PSU, and Caltech.

This replication of data using GridFTP is enabling more gravitational wave data analysts across the world to do more science more efficiently than ever before. Globus GridFTP is in the critical path for LIGO data analysis.

We have been able to leverage Globus GridFTP so effectively in part because of the excellent client API. We have created a tool we call the Lightweight Data Replicator (LDR) that includes a customized client built on top of the Globus GridFTP API, without which our robust tool would not be possible.

Soon we hope to upgrade LDR to use the latest GridFTP-enabled server developed by the Globus team. Specifically we want to use our own plugin so that we can deliver particular channels of LIGO data, pre-processed, using the GridFTP protocol. This will allow LIGO scientists even faster and more efficient access to the data they wish to analyze.

These SciDAC DataGrid Middleware sponsored middleware tools are also in use across several international collaborations. Projects such as the EU Enabling Grids for E-Science (EGEE), the LHC Computing Grid (LCG), the NorduGrid, and Quantum ChromoDynamics Grid (QCDGrid) to name a few. The DataGrid Middleware tools allow the QCD scientists to focus on the science and not the middleware. In the words of Dr. Richard Kenway, PI for UKQCD project:

The Globus Toolkit is at the heart of the software that runs the UKQCD Grid. The toolkit furnishes us with the middleware layer on which we have built high level, user-facing applications. The availability of the Globus Toolkit has significantly reduced the development effort required within the project, eliminating the need for us to design and implement complex software modules. It has also reduced the support overhead; since we are reusing tried and tested software that has been exercised by academic and corporate users from across the world."

In short, these tools have become the de facto standard for data management the world over, they are currently enabling science on a scale impossible without these tools, and scientists are counting on support and further enhancements of these tools to advance their science even further.