PRAS-DT: Portable, Reliable and Automatic Streaming Data Transfer with Globus Online

Christine Harvey, Richard Stockton College of New Jersey, Dr. Rosa Filgueira, University of Edinburgh,
Dr. Malcolm Atkinson, University of Edinburgh

Abstract

Modern science involves enormous amounts of data which need to be transferred and shared among various locations. For the EFFORT (Earthquake and Failure Forecasting in Real Time) project, large data files need to be synchronized between different locations and operating systems in near real time. There are many challenges in performing large data transfers, continuously, over a long period of time. The use of Globus Online to perform the data transfers addresses many of these issues. Globus Online is quickly becoming a new standard for high performance data transfer. Globus Online allows for periodic data transfers to be automatically run through scripts and for simplified secure data movement without the hassle of complex certificates.

Why Globus Online?

globus online

- Command Line Interface & GUI
- Globus Connect
- Endpoint Management
- Quick File Transfer
- Transfer Monitoring
- Secure

Introduction

The research for the EFFORT project has very specific data transfer requirements. Data needs to be easily transferred from multiple endpoints in London and Catania to the server in Edinburgh. In this project we need to choose and set up a mechanism for the server machine to receive data. This mechanism should be able to transfer data automatically, without human involvement and should be compatible with different operating systems. The mechanism should perform data transfers in near real time, files are generated every minute. One experiment could be running for several months or years. If any of the systems fail, data transfer should be easily resumed in an orderly fashion when the system is corrected.

Taking in these many considerations for the project, it became easy to narrow down the data transfer possibilities. SFTP is an easy to use, data transfer tool but only allows for single-stream file transfers [NASA 2012]. SFTP is not a reasonable solution for working with multiple endpoints. A more obvious solution might be to use GridFTP. GridFTP can be used for multi-stream file transfers and is extremely efficient [Khanna 2008]. The downside to GridFTP is that it is difficult to obtain certificates and establish trust on the systems, this makes it extremely difficult for non-experts to use.

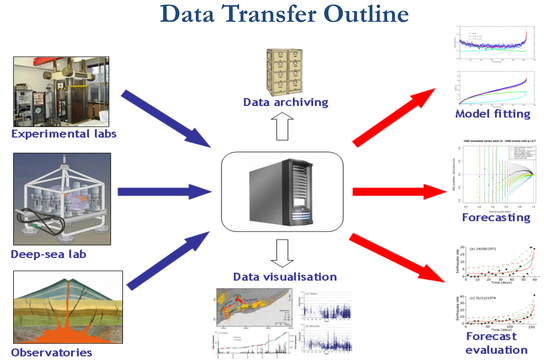
Globus Online is a convenient interface for transferring files between two or more end points. The system uses GridFTP to perform transfers and can be easily incorporated into scripts to run the transfers periodically [Foster 2011]. These scripts can be installed at either of the endpoints involved in the data transfer. Globus Online has all of the reliability and trust that goes with using GridFTP while avoiding having to handle certificates and difficult installations.

Globus Online combines the reliability, security, and efficiency of GridFTP with the convenience of a GUI or a simple CLI. For all these reasons, Globus Online was chosen to complete the data transfers for the EFFORT project.

Acknowledgements

The Open Science Data Cloud PIRE Internship Program
The National Science Foundation for funding through award #1129076
University of Edinburgh School of Informatics
Richard Stockton College of New Jersey

Data Transfer Outline



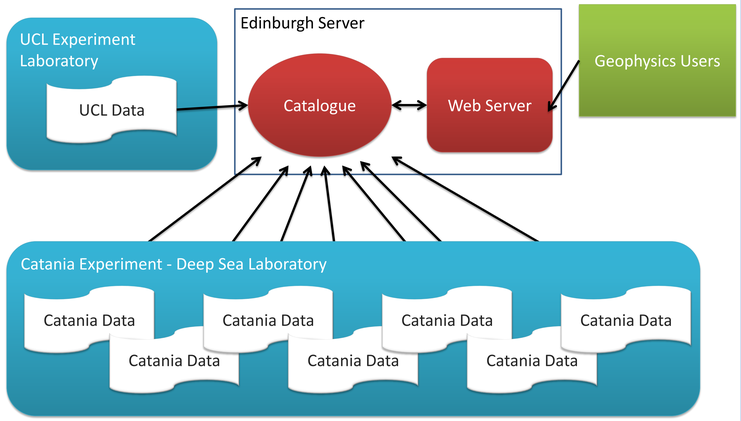
The EFFORT Project

The EFFORT (Earthquake and Failure Forecasting in Real Time) project is a collaboration between the University College of London, The University of Edinburgh Geosciences Department, and The University of Edinburgh School of Informatics. The focus of the project is to determine the predictability of brittle failure of rock samples in laboratory experiments and then see how this predictability scales to the greater complexity of natural-world phenomena. Data will be collected from controlled laboratory experiments which includes data from the UCL Laboratory and Data from deep-sea experiments in Catania. The UCL laboratory will complete experiments which includes traditional laboratory brittle creep experiments at the Rock Physics labs [EFFORT 2012]. In the future, data will also be collected from NERC funded UCL/Edinburgh Creep2 project which will undertake a series of low strain-rate brittle creep experiments in a seep-sea laboratory [CREEP2 2012].

The Edinburgh Informatics Research group is responsible for the data management in the EFFORT project which includes data transfer, data storage, and data access. In the controlled laboratory experiments there are two types of data, Time Driven Data (TDD) and Acoustic Emission Data (AE). The TDD reflects changes in bulk properties of the sample and are generated once per minute. AE data is processed to pick up individual acoustic emission events and compile a catalogue.

These files need to be transferred from both UCL and Catania without any manual interaction aside from the initial start of the data transfer.

The EFFORT Project



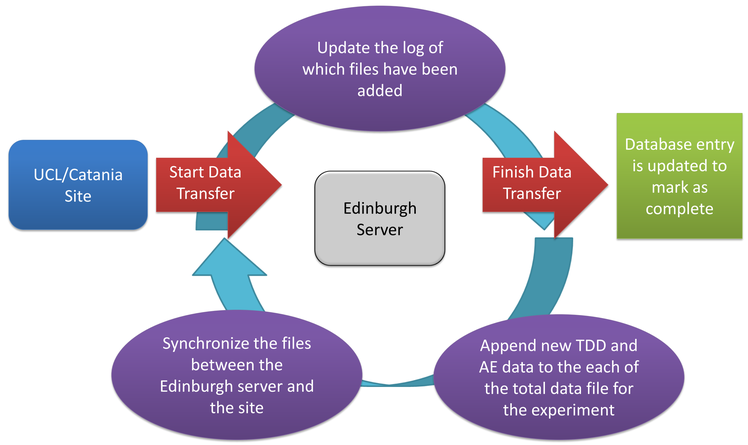
Methodology of PRAS-DT

The initial step in this process was creating the Globus Online account and setting up the necessary endpoints. Endpoints were created for the Edinburgh server and the UCL laboratory using Globus Connect. In the future, an endpoint for the Catania experiments will be created. Globus Connect makes it possible to begin transferring data within an hour of beginning the installation process.

PRAS-DT (Portable, Reliable and Automatic Streaming Data Transfer) is the data transfer mechanism that we have designed for EFFORT in order to satisfy the transfer requirements explained previously. PRAS-DT is an automated script that is designed to run for the life of the experiment. PRAS-DT bringns by running a shell script in the background that begins the data transfer and synchronization between the laboratory endpoint and the Edinburgh server. Following the data transfer, a workflow is run on the new files to append the new files to the total data file for the experiment and to log the files which have been added. This process ensures that the files are appended to the total data file in the correct order and that files are not skipped or appended more than once.

The data transfer and appending process is repeated every minute until the experiment completes.

PRAS-DT Workflow



References

NASA Advanced Supercomputing Division (March 7, 2012). How can I speed up my data transfers to/from NAS? Retrieved from NASA website: http://www.nas.nasa.gov/hecc/assets/pdf/training/Speeding_Up_Data_Transfers_2012_03_07.pdf.

NGS Globus Online. [Accessed July 31, 2012]; Available from: http://www.ngs.ac.uk/globus-online.

Allen, B., Bresnahan, J., Childers, L., Foster, I., Kandaswamy, G., Kettimuthu, R., Kordas, J., Link, M., Martin, S., Pickett, K., and Tuecke, S., Globus Online: Radical Simplification of Data Movement via SaaS. Submitted for publication, 2011.

Foster, I., Boverhof, J., Chervenak, A., Childers, L., DeSchoen, A., Garzoglio, G., Gunter, D., Holzman, B., Kandaswamy, G., Kettimuthu, R., Kordas, J., Livny, M., Martin, S., Mhashilkar, P., Miller, Z., Samak, T., Su, M., Tuecke, S., Venkataswamy, V., Ward, C., and Weiss, C. Reliable High-Performance Data Transfer via Globus Online, Preprint ANL/MCS-P1904-0611, June 2011.

Khanna, G., Kurc, T., Catalyurek, U., Kettimuthu, R., Sadayappan, P., and Saltz, J. A dynamic scheduling approach for coordinated wide-area data transfers using GridFTP. In Proc. of 22th International Parallel and Distributed Processing Symposium (IPDPS), Miami, Florida, 2008.

Kourtellis, N., Prieto, L., Iamnitchi, A., Zarrate, G., and Fraser, D. Data transfers in the grid: workload analysis of globus gridftp. In DADC '08: Proceedings of the 2008 international workshop on Data-aware distributed computing, pages 29–38, New York, NY, USA, 2008. ACM.