Extreme-Scale Data Lifecycle Management as a Service

Ian Foster
University of Chicago and Argonne National Laboratory

Introduction

Extreme-scale computing and scientific instrumentation are resulting in dramatic increases in the volume, complexity, and diversity of the data generated and used in science. The result is increased pressure on investigators who need not only data storage, but full-service data lifecycle management processes, encompassing data collection, storage, sharing, metadata, search, archiving, provenance, assignment of DOIs, security, privacy, etc. Establishing and efficiently executing such processes would demand substantial time and resources that most researchers do not have, and cannot easily acquire.

I believe that the solution to this problem is not simply to define "best practices"—nor to provide researchers with software. Once defined, best practices must still be implemented. software still must be installed, operated, and maintained. Those implementation, installation, and operations steps are precisely where many investigators run into problems.

Instead, I argue that we should aim to outsource major elements of the lifecycle management process to third-party **Research Data Lifecycle Management services**. Ideally, such services will encompass discipline-specific practices and methods, so that the individual researcher can connect their lab and then have many of their problems taken care of—much as many outsource their email to Google today.

My team at Argonne National Laboratory and the University of Chicago has been exploring such an approach for the past three years, with some success as I describe here. Our first attack on the problem is Globus Online (www.globusonline.org). This hosted (software as a service: SaaS) system allows researchers to outsource research data transfer, synchronization, and sharing tasks to a third-party service. Web 2.0 interfaces provide for convenient Web browser, REST, and command line interfaces. The ability to configure the details of personal, campus, and national resources makes interoperating among different storage systems straightforward. Even this limited capability is proving to be of significant value to researchers, as evidenced by a steady flow of new users, who as of April 2013 have used it to move more than 16 PB and close to one billion files. Many campuses, supercomputer centers, and experimental facilities recommend it to their users.

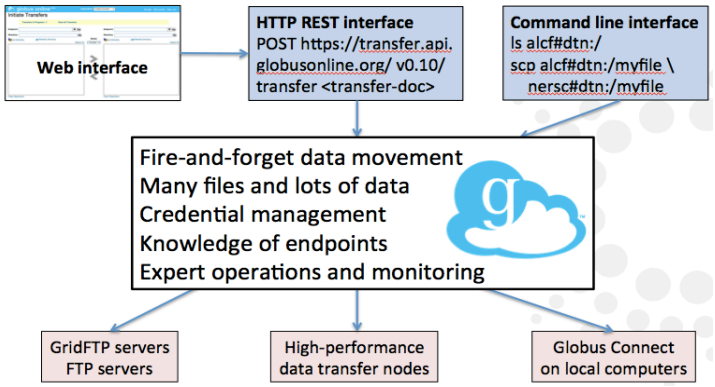
In the following, I review the current Globus Online implementation and then describe extensions that can allow it to address concerns of extreme-scale data.

Globus Online today

We started work on Globus Online in late 2009 with the goal of using SaaS capabilities to radically simplify research data transfer. The first publicly available Globus Online system, delivered in November 2010, implements methods for managing the transfer of single files, sets of files, and directories, as well as rsync-like directory synchronization. It manages security credentials, including for transfers across multiple security domains; selects transfer protocol parameters for high performance; monitors and retries transfers when there are faults; and allows users to monitor status. A Web interface allows the casual user to initiate and monitor transfers from a browser, while command line and REST interfaces

permit the frequent user to script usage within custom research workflows and integrate Globus Online into applications.

The figure presents a user view of the system. Note the scp command used to illustrate the command line interface; this command has the same syntax as the commonly used but slow secure copy, but invokes high-performance, Globus Online-optimized GridFTP transfers that can move data 20 times faster or more that regular scp in many circumstances.



Response to Globus Online has been positive from both individual, who are delighted that previously time-consuming tasks are now automated, and from resource providers, who see user productivity and resource utilization increase, and support demands decrease. Over 8,000 users have registered and hundreds use the service on a regular basis.

Extending Globus Online to extreme-scale data

We are working to extend Globus Online to encompass additional data management tasks. Many of these extensions are relevant to the demands of extreme-scale data.

We continue to **accelerate performance** achieved by data transfer operations to address the demands of extreme-scale data, ultra-high-speed networks and file systems, and hierarchical storage. We want to make it possible, for example, for a researcher to move data at wire speeds among storage and computing systems that are connected by 100 Gbps or even 1 Tbps networks; while also allowing system administrators to manage how scarce bandwidth is allocated among competing interests.

Working with the University of Southern California, we are developing a **dataset service** that will allow a researcher to create, manage, access, search, and share metadata about sets of data elements (files, directories, database rows, …)—with utilities allowing for automated metadata extraction from various data sources, and Globus Online maintaining the metadata on cloud storage. This service will allow users to track, through time and space, the evolution of the diverse data associated with a research project.

We are also integrating **computation services** with Globus Online, working initially in genomics. Our new Globus Genomics service allows users to not only manage genome sequence data but also request that computational pipelines be executed on that data.

Further reading (at www.globusonline.org)

Foster, I. Globus Online: Accelerating and democratizing science through cloud-based services. *IEEE Internet Computing*(May/June):70-73, 2011.

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. Preprint CI-PP-05-0611, Computation Institute, 2011.