Reliable high-performance data transfer via Globus Online
Accomplishments of the Center for Enabling Distributed Petascale Science

Ian T. Foster[1], Josh Boverhof[2], Ann Chervenak[3], Lisa Childers[1], Annette DeSchoen[3], Gabriele Garzoglio[5], Dan Gunter[2], Burt Holzman[5], Gopi Kandaswamy[3], Raj Kettimuthu[1], Jack Kordas[1], Miron Livny[4], Stuart Martin[1], Parag Mhashilkar[5], Zachary Miller[4], Taghrid Samak[2], Mei-Hui Su[3], Steven Tuecke[1], Vanamala Venkataswamy[3], Craig Ward[3], Cathrin Weiss[4]

[1] Argonne National Laboratory, Argonne, IL 60439, USA
[2] Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[3] Information Sciences Institute, U. Southern California, Marina del Ray, CA 90292, USA
[4] University of Wisconsin, Madison, WI 53706, USA
[5] FermiLab, Batavia, IL 60510, USA

Abstract: The SciDAC Center for Enabling Distributed Petascale Science (CEDPS) seeks to accelerate DOE research by eliminating barriers to reliable and performant wide area data movement. A major focus of CEDPS R&D has been the development of Globus Online, a hosted data movement service to which users can hand off responsibility for a range of data movement tasks. Here we introduce the Globus Online system and summarize its impact to date since its launch in November 2010, which include more than 500 terabytes moved, 1,000 registered users, and adoption by DOE leadership computing facilities. We also review work aimed at further expanding Globus Online's utility by integration with the Condor high-throughput computing system, generalization of the endpoint concept to virtual endpoints with specialized properties such as replication, and performance monitoring as a step toward automated diagnosis of performance problems.

1. Introduction

Effective use of high-speed networks for DOE research requires that we overcome the usability barriers that impede network use by nonexpert users. To this end, the Center for Enabling Distributed Petascale Science (CEDPS) project launched the Globus Online project in 2009 to enable reliable high-performance research networking for the masses. The key idea is to outsource complex file movement tasks (e.g., "copy directory D from system A to system B" or "synchronize directory E and F") to a specialized software-as-a-service (SaaS) provider, Globus Online [2, 5, 6], that then takes responsibility for managing the end-to-end process. Data transfers are performed via GridFTP [4], which supports high-speed transport but has previously required tedious manual configuration. The Globus Online service incorporates knowledge about endpoint and network protocol configurations to reduce the expertise (and software installation) required of users. Intuitive Web 2.0 interfaces and a one-click install data movement client further simplify the user experience. The service also incorporates logic designed to optimize end-to-end performance and to recover from transient failures. The SaaS approach also permits rapid (and transparent) software upgrades and expert operator diagnosis of persistent failures.

We describe four distinct activities relating to this project. First, we describe the Globus Online design, with a particular focus on its Web 2.0 user interfaces and their intended usage modalities. We also summarize initial experiences with its production deployment, which has so far involved 1000 registered users, the high-speed movement of hundreds of terabytes, and the adoption of Globus Online by NERSC as the recommended data transfer method. Second, we describe an integration of Globus Online with the Condor high-throughput computing system, an important early experiment in integrating Globus Online with a major programming tool. Third, we describe a research project that aims to generalize the Globus Online endpoint concept to encompass virtual endpoints with specialized properties such as data replication. Fourth, we describe experiments with the use of multiple sources of network performance data and logging data to automate performance analysis and problem determination for transfers.

2. Design, deployment, and application of Globus Online

CEDPS launched in late 2009 an experiment aimed at exploring the feasibility of using SaaS capabilities to achieve a radical simplification of research data transfer. Our goal was to construct a hosted service that would implement solutions to challenging data transfer tasks (e.g., transfer management, credential management, recovery from transient errors) while also providing modern Web 2.0 interfaces. The result is the Globus Online system [5, 6].

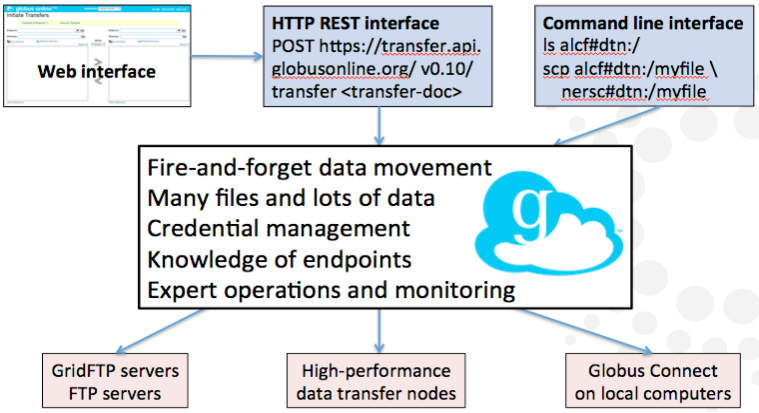

Figure 0: User view of Globus Online.

The first production-level 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 can manage security credentials, including for transfers across multiple security domains; select transfer protocol parameters for high performance; monitor and retry transfers when there are faults; and allow users to monitor status. It provides REST, Web browser, and command line interfaces, so that the casual user can initiate and monitor a transfer from a Web browser, while a frequent user can integrate Globus Online calls into an application via command-line scripting or REST messages. Figure 1 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 it invokes high-performance, Globus Online-optimized GridFTP transfers. The Globus Online implementation is hosted on Amazon Web Services cloud infrastructure to enable convenient state replication and elastic computing capacity. Note that data transfers do *not* proceed via Amazon; only the data transfer management logic executes there.

Fortuitously, our development of Globus Online coincided with the rollout of data transfer nodes (DTNs) [1] across DOE sites: dedicated Linux boxes, each with two 10 Gb/s network connections (to ESnet and an internal network), configured for high-speed GridFTP transfers to site parallel file systems. Globus Online allows users to refer to these systems (and any other endpoint) by simple symbolic names (e.g., alcf#dtn) rather than hostname, ports, and URLs. Globus Online also dynamically load balances and fails over among multiple DTNs if a site has more than one—as is the case, for example, at DOE leadership computing facilities.

The first version of Globus Online did not solve the "last mile problem," of transferring data to and from anywhere—not just between the high-end facilities that had GridFTP servers installed. We addressed this shortcoming in April 2011 with the introduction of Globus Connect (Figure 2), a special packaging and wrapping of the GridFTP server binaries for Windows, Mac OS X, and Linux that can be trivially

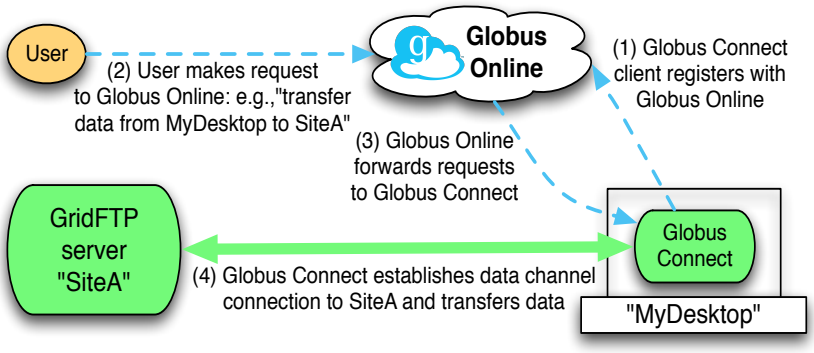

Figure 0: Schematic of the Globus Connect client.

installed by users (without administrative privileges) on their own desktops and laptop, even if they are behind a firewall or Network Address Translation device that only allows outbound connections. With a few easy steps, users can connect their computer to Globus Online, thus opening Globus Online's high-performance, easy-to-use file transfer capabilities to many more users and uses.

Globus Online has been running only six months, but we view it as a great success. More than 1,000 users have transferred over 0.5 petabytes of data, and NERSC recommends it to its users. Performance results are also encouraging. We present results in Figure 0 for transfers over ESNet between high-performance parallel storage systems at ALCF and NERSC [5]. Each data point represents the average performance achieved when transferring many files of a specified size. We give results for Globus Online in two configurations: running between a single data transfer node (DTN) at ALCF and NERSC ("go-single-ep"), and (the default configuration) using the two data transfer nodes that are supported by ALCF and NERSC ("go"). We also show results for scp and for the commonly used globus-url-copy (GUC) client, both in its default configuration and when tuned by an expert. Scp performs badly in all cases. GUC with its default configuration performs badly for all file sizes. (The default configuration needs improving.) Tuned-guc performs much better than GUC in almost all cases, but less well than Globus Online for smaller file sizes—probably because Globus Online drives GridFTP pipelining more aggressively, due to the improved pipelining support in Globus Online's fxp client to GridFTP. Tuned-guc does slightly better than Globus Online for large files; thus, there remain opportunities to tune Globus Online performance further. Note also that the Globus Online transfers to a two DTNs vs. a single DTN are not substantially different except for the largest transfer. We conclude that the bottleneck is not the DTNs, but the network or local storage.

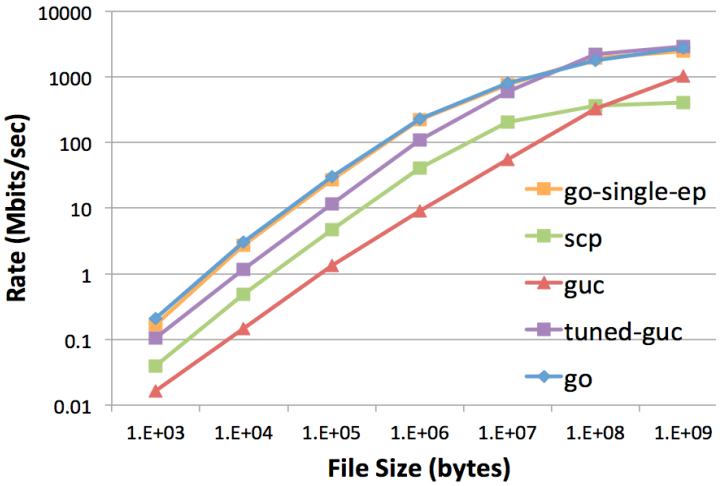

Figure 0: Globus Online performance ALCF-NERSC.

Our initial experiences also suggest that SaaS can indeed have advantages as a delivery mechanism for research data management software. At various points during Globus Online's development, we variously discovered errors in the software and received urgent requests for new features. The resulting corrections and enhancements were delivered within hours to days, rather than the weeks or even months that sometimes ensue with traditional software distribution methods. In addition, our operations team demonstrated their ability to discover persistent errors associated with specific endpoints and to notify sites of those errors in a manner that permitted their timely correction.

Inevitably we have also encountered some problems, particularly as we learned how to manage a production SaaS capability. For example, we had some brief service interruptions due to power system upgrades at the site (Argonne) used to host static (Web) content; this situation was resolved, but it emphasized to us the benefits of hosting SaaS services on commercial hosting services. (Operation of the Globus Online service itself was not interrupted by this situation.) The May 2011 Amazon service outage, which received a lot of press, led to a less-than-one-hour interruption in Globus Online service.

3. Integration with Condor in support of high-throughput computing

To facilitate the use of Globus Online by a larger audience, we have adapted the Condor [12] high-throughput computing software to allow users to specify job input files using their Globus Online account's endpoints and to transfer the job's output sandbox to a Globus Online endpoint. This support

permits a batch job to use third-party transfers both to fetch input data and to store output (significantly easing the I/O load on the submitting machine, allowing for potentially greater job throughput); the support also allows Globus Online transfers to be scheduled by the Condor software. The integration effort also provided valuable feedback for the design of the Globus Online command line interface, introducing requirements for non-interactive use. Work to integrate Globus Connect is under way to aid users in dealing with common network connectivity issues such as executing behind firewalls or NAT.

We are also investigating the use of Globus Online to drive asynchronous job data transfers in Condor. To keep CPUs busy, we want to minimize the extent to which operations are blocked on I/O. Scientific applications may run for hours and produce output data (the job's "output sandbox") that is much larger than the job's input data. This output data must be transferred back to the submitting machine from the executing machine after the job is complete. In Condor's current execution model, no other job can use a CPU while this transfer is taking place, an approach that leads to the CPU being idle.

To improve this situation, we are using Globus Online to drive asynchronous data transfers for jobs of the same user. An asynchronous data transfer overlaps a job's output sandbox transfers with the next job's input sandbox transfer, allowing the new job to start executing while the previous job's output transfer is still ongoing. Therefore, CPU idle time is limited to transfer of the new job's input data only, which, in the typical case, takes significantly less time than transferring output data. Figure 4 shows the sort of improvement that can be expected. In this preliminary experiment, performed without the use of Globus Online, we run a batch of similar jobs with an average input sandbox transfer time of 2 minutes, job execution time of 35 minutes, and output sandbox transfer time of 8 minutes. The asynchronous version achieves significant performance improvements.

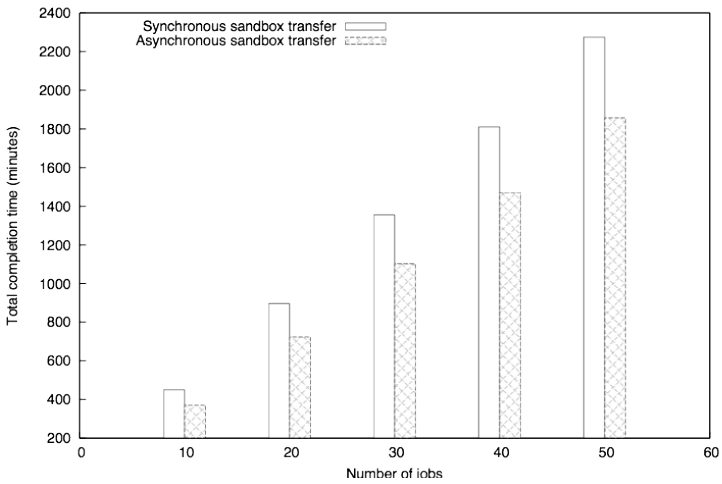

Figure 0: Synchronous vs. asynchronous staging performance.

The integration of Condor and Globus Online is also helpful within the context of the Glidein Workload Management System (Glidein WMS) [11] developed by the US Central Muon Solenoid (CMS) experiment and the Fermilab Computing Division, leveraging the work of the Condor team at UW-Madison. Glidein WMS is a service that can create, on demand, a dynamically sized overlay Condor batch system on Grid resources. Thus, for example, a user with many tasks to execute might request that Glidein WMS create a single overlay Condor system that encompasses one set of nodes on Open Science Grid [10], and another set of nodes on the Worldwide LHC Computing Grid [3]; tasks can then be dispatched to those different nodes as if they were a single system.

Well-managed access to computing resources through Glidein WMS requires an appropriately well-managed data infrastructure. Most facilities provide well-supported storage services, exposed through modern storage interfaces, such as GridFTP or the Storage Resource Manager. However, storage is only one building block of a data management infrastructure. While large groups well versed in Grid technologies, such as US CMS or ATLAS, can take advantage of data management systems built in-house, the majority of the scientific community does not have the means to commission such custom-tailored solutions.

The integration of Condor, as the base service of Glidein WMS, with Globus Online in the context of CEDPS aims at making available to all scientific communities a well-integrated computing and data management environment. Figure 5 shows an architectural view of this integration. The Glidein Factory manages the submission of glideins to remote resources, while the Condor batch system (collector + negotiator) manages the overlay pool. The VO Frontend is responsible for monitoring the user schedulers and requesting that the factory create new glideins as required. Each Factory can serve multiple glideinWMS and Corral Frontends. Glideins interface with Globus Online through Condor's Globus Online file transfer plugin to manage transfer of input and/or output sandboxes of the user job.

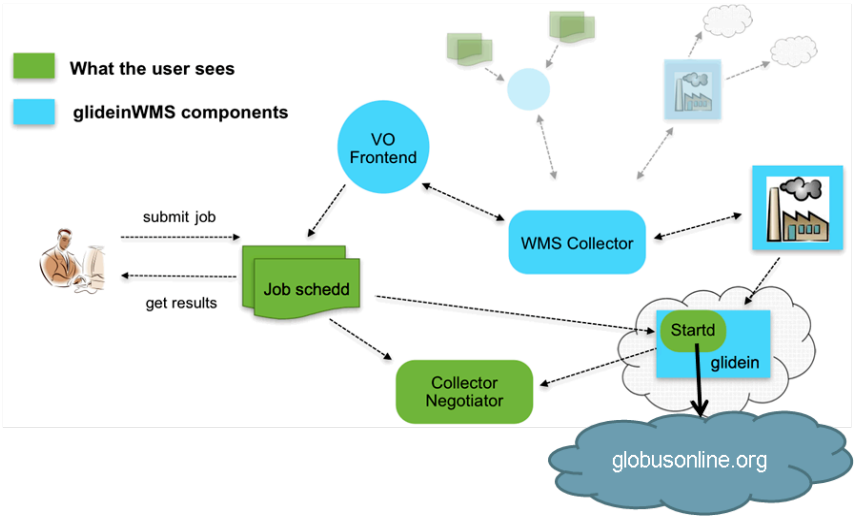

Figure 0: Glidein WMS workload management system.

4. Expanding Globus Online capabilities via virtual endpoints

Let us define a physical endpoint to be a GridFTP server. We then define a *virtual endpoint* as a set of physical endpoints with an associated set of policies. When files are transferred to a virtual endpoint, its policies govern how and what files are transferred or replicated across its constituent physical endpoints. For example, a virtual endpoint can enforce a set of data replication policies to achieve a specified level of fault tolerance and performance, in a manner that is transparent to the user. More specifically, a policy might require that every file written to a virtual endpoint be automatically replicated on three geographically distinct physical endpoints to provide a high level of availability. A more sophisticated policy might enforce different replication strategies for files of different types or sizes.

A virtual endpoint also provides users with a unified view of a shared data space, as illustrated in Figure 6. Virtual endpoint VE1 (on the right) consists of two physical endpoints, GridFTP Server 1 and GridFTP Server 2 (on the left). Let us assume that policies associated with VE1 mean that the directory *user1_data_dir*, when written to VE1, is replicated at Endpoint1://home/user1/data_dir and Endpoint2://home/user1/data_dir. Any changes made to files within a virtual endpoint automatically get reflected in all replicas of that file within that virtual endpoint. When retrieving files from a virtual endpoint, Globus Online will choose the optimal replica to minimize the data transfer time.

We have designed and prototyped support for virtual endpoints for the Globus Online system. Our design gives users the ability to create and manage virtual endpoints and to associate policies with those endpoints. It also provides additional Globus Online commands to write, read, and edit files and to manage file transfers in a virtual endpoint. When a user writes files to a virtual endpoint, those files are written to one or more physical endpoints according to the specified policies. When a user reads files from a virtual endpoint, those files can be retrieved from any physical endpoint within the virtual endpoint that holds a valid copy of the files. The system also supports a check-out operation that allows users to copy files in a virtual endpoint to a local file system, manipulate those files, and then check them back in to the virtual endpoint, where they are automatically propagated to constituent physical endpoints according to the policies of the virtual endpoint. Users can also manage file transfers to/from a virtual endpoint. The system provides commands to query transfer status and to cancel active file transfers.

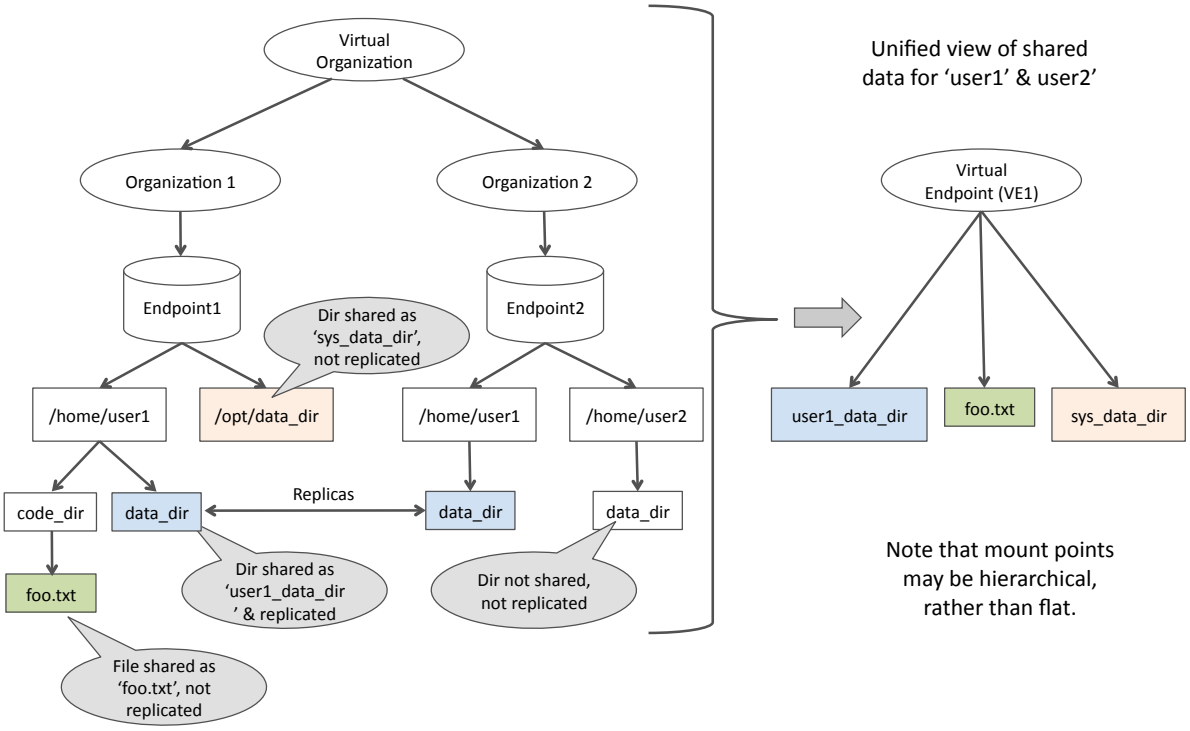

Figure 1: Illustration of the virtual endpoint design; see text for details.

5. Performance monitoring for end-to-end data transfers

CEDPS performance measurement and analysis activities have focused on two areas: collecting and normalizing Globus Online usage data into an online analysis archive, and using ongoing transfer information as signals to drive data collection and correlation with perfSONAR network measurements. For the first area, we serialized Globus Online transfer activity into the CEDPS Best Practices log format [7] and pulled this data from Globus Online to a MongoDB database running at LBNL. This data is then easily mined with analysis tools such as R or Matlab without danger of perturbing Globus Online activities. The pull frequency is roughly one minute, allowing near-real-time analysis of large transfers.

For the second area, we have developed a component called Periscope, a distributed measurement cache that combines online measurements from the application, network, and other layers of the stack with topology and service location metadata. Periscope leverages the Extensible Session Protocol (XSP) to help correlate information from multiple layers. The XSP groups related network connections, abstracting the user/network interaction away from any protocol-specific notions of a connection and into a more general concept of a *session*, which is defined in the New Oxford American Dictionary, 2nd edition, as a "period devoted to a particular activity."

In the context of Globus Online, XSP is connected to the MongoDB database on one end and Periscope on the other. XSP signals to Periscope when a GO transfer of multiple files begins and ends, and also sends per-transfer measurements. Periscope uses the transfer signals to direct its polling and caching of perfSONAR measurements, including the packets in/out of every router port along the path between the two data transfer endpoints. As a result, detailed information on a given transfer is readily available. We have integrated this information into a visual Periscope dashboard, shown in Figure 2, which displays the path for a selected transfer and provides overall bandwidth as well as interactive graphs for each monitored router on the path.

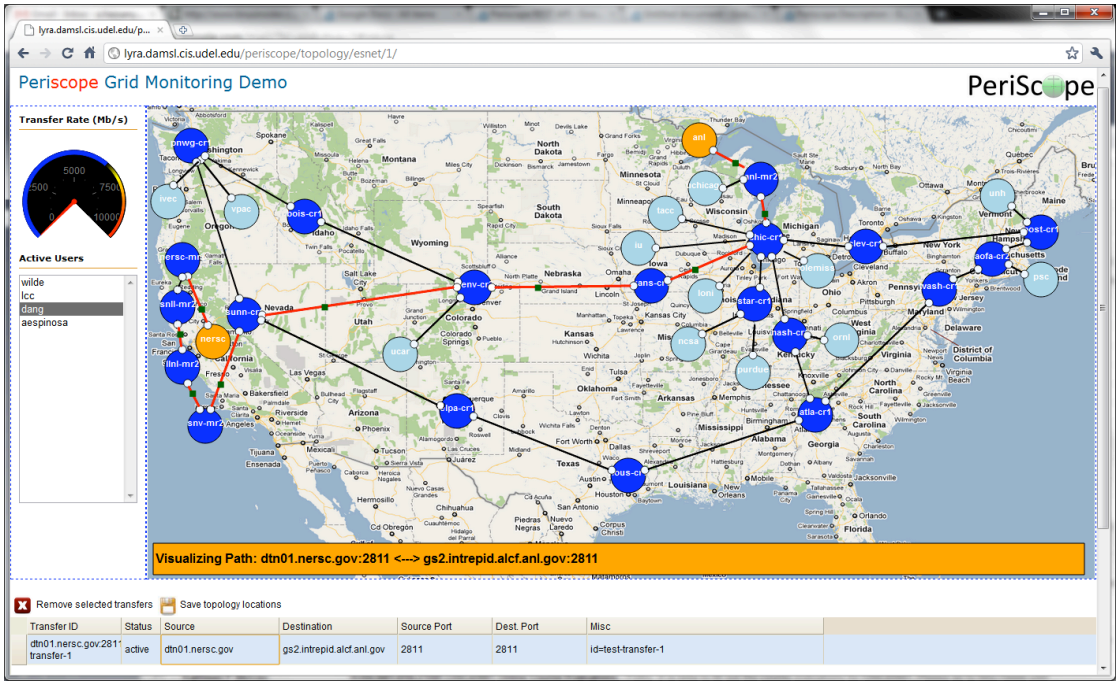

Figure 2: Periscope dashboard.

We have also integrated XSP into GridFTP's extensible I/O stack (XIO) and combined this with the NetLogger Calipers high-performance instrumentation library [8] to provide Periscope with detailed measurements of both disk and network I/O for a given transfer [9]. In this case, "sessions" can be detected as the beginning and end of a GridFTP transfer.

6. Summary and future directions

More than 10 years of DOE R&D has produced remarkable improvements in peak wide-area data transfer speed, via specialized networks, protocol optimizations, and end-user software. But unfortunately, most researchers still move data using secure copy (scp) and achieve abysmal performance. New tools are required if DOE science is to realize the promise of today's gigabit and tomorrow's terabit networks. But these tools cannot realistically require that users learn to install, configure, and operate complex software stacks. Networks need to be both faster and simpler.

Globus Online represents a first step toward a potentially promising solution to this conundrum. Its SaaS approach can move complexity from the user environment to a remote facility where specialized methods, professional operations, and economies of scale can be exploited. Initial experiences suggest that the approach has promise: users and centers are enthusiastic, performance is good, and usage is growing.

Feedback from early users suggest many directions for future research and development. From an engineering perspective, users want support for more protocols (e.g., HTTP, SRM, WebDAV) and better treatment of firewalls, and sites (e.g., supercomputer centers and experimental facilities) would like support for monitoring and managing their traffic. We also know that we can do a better job of performance tuning. Looking further forward, we want to expand the capabilities to a much wider range of research data management functions, with the ultimate goal of moving much of research data management out of the laboratory and off the researcher's desktop. The virtual endpoint concept described here is a step in that direction.

Acknowledgments

This work is supported by the U.S. Department of Energy Office of Science, Office of Advanced Scientific Computing Research, through the SciDAC program under contract DE-AC02-06CH11357.

References

1. ESnet data transfer nodes. [Accessed May 16, 2011]; Available from: http://fasterdata.es.net/fasterdata/data-transfer-node/.
2. GlideinWMS: The Glidein-based Workload Management System [Accessed June 4, 2011]; Available from: www.uscms.org/SoftwareComputing/Grid/WMS/glideinWMS/.
3. Worldwide LHC Computing Grid. [Accessed June 8, 2011]; Available from: http://lcg.web.cern.ch/lcg/.
4. Allcock, B., Bresnahan, J., Kettimuthu, R., Link, M., Dumitrescu, C., Raicu, I. and Foster, I., The Globus Striped GridFTP Framework and Server. SC'2005, 2005.
5. 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-5-0611, Computation Institute, The University of Chicago, 2011.
6. Foster, I. Globus Online: Accelerating and democratizing science through cloud-based services. *IEEE Internet Computing*(May/June):70-73, 2011.
7. Gunter, D. Logging Best Practices Guide. [Accessed June 2, 2011]; Available from: http://docs.google.com/View?id=dgtn7s3w_16fwxbfshq.
8. Gunter, D. NetLogger Calipers API. [Accessed June 8, 2011]; Available from: http://netlogger.lbl.gov/doc/netlogger-calipers.
9. Kissel, E., Gunter, D., Samak, T., El-Hassany, A., Fernandes, G. and Swany, M. An Instrumentation and Measurement Framework for End-to-End Performance Analyis. Technical Report 2011/04, University of Delaware, 2011.
10. Pordes, R., Petravick, D., Kramer, B., Olson, D., Livny, M., Roy, A., Avery, P., Blackburn, K., Wenaus, T., Würthwein, F., Foster, I., Gardner, R., Wilde, M., Blatecky, A., McGee, J. and Quick, R., The Open Science Grid. Scientific Discovery through Advanced Computing (SciDAC) Conference, 2007.
11. Sfiligoi, I., Bradley, D.C., Holzman, B., Mhashilkar, P., Padhi, S. and Wurthwein, F., The Pilot Way to Grid Resources Using glideinWMS. WRI World Congress on Computer Science and Information Engineering, Los Angeles, California, USA, 2009, 428-432.
12. Thain, D., Tannenbaum, T. and Livny, M. Condor and the Grid. Berman, F., Fox, G. and Hey, A. eds. Grid Computing: Making The Global Infrastructure a Reality, John Wiley, 2003.