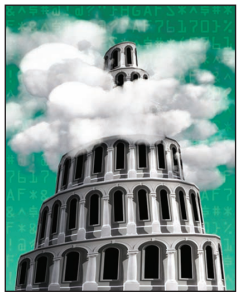
Editor: George Pallis • gpallis@cs.ucy.ac.cy



# Globus Online

## Accelerating and Democratizing Science through Cloud-Based Services

**Ian Foster** • *Argonne National Laboratory and the University of Chicago*

Many businesses today save time and money, and increase their agility, by outsourcing mundane IT tasks to cloud providers. The author argues that similar methods can be used to overcome the complexities inherent in increasingly data-intensive, computational, and collaborative scientific research. He describes Globus Online, a system that he and his colleagues are developing to realize this vision.

The scientific community today has unprecedented opportunities to effect transformational change in how individuals and teams engage in discovery. The driving force is a set of interrelated new capabilities that, when harnessed, can enable dramatic acceleration in the discovery process: greater availability of massive data, exponentially faster computers, ultra-high-speed networks, and deep interdisciplinary collaboration. The opportunity — and challenge — is to make these capabilities accessible not just to a few "big science" projects but to every researcher at every level. Here, I argue that the key to seizing this opportunity is embracing software delivery methods that haven't been widely adopted in research, notably software as a service (SaaS) — a technology that forms an important part of what people refer to as the cloud. I also describe projects in the Computation Institute at the University of Chicago and Argonne National Laboratory that aim to realize this vision, focusing initially on data movement and management.

### From Grid to Cloud

It's been more than a decade since Carl Kesselman and I posited a world in which computing is delivered on demand as a service, and virtual organizations link scientists and resources worldwide.[1] This vision is now a reality. For example, the Large Hadron Collider (LHC) Computing Grid regularly distributes tens of terabytes to hundreds of analysis sites worldwide; 25,000 people use the Earth System Grid to access climate simulation data; commercial cloud providers deliver on-demand cycles and storage at scales unachievable in academic settings; and the US InCommon trust federation allows more than 5 million people to access remote resources using local credentials. We surely didn't expect that these developments would occur so quickly and at this scale.

However, although big science projects such as those just listed can afford to create and operate dedicated grid infrastructures, smaller research teams can't. Their IT staff consists of maybe a grad student or a technician. Yet to be competitive, these teams must somehow collect, manage, move, and analyze tens of terabytes of data — just like the big guys. It isn't sufficient to give them more software because they don't have the time or expertise to install and operate it. We must find ways to deliver the IT required for research in a far more convenient and cost-effective manner. To boil it down to a couple of buzzwords, we must harness the power of the cloud to scale access to the grid.

### Research from the Coffee Shop

Creating and running a modern business is a complex, information-intensive activity. Yet amazingly, an entrepreneur today can run a business

 Published by the IEEE Computer Society 1089-7801/11/$26.00 © 2011 IEEE 

from a coffee shop, outsourcing Web hosting, email, payroll, accounting, customer-relationship management, and many other functions to third-party cloud providers such as Google and Salesforce.com. The results are both a considerable reduction in the costs associated with operating the business and accelerated innovation thanks to access to advanced IT capabilities previously available only to big companies.

Research is no less complex and information-intensive. Data from experiments is increasing rapidly — for example, the output from gene sequencing machines has grown by close to four orders of magnitude in just five years.[2] Many researchers are overwhelmed by the challenges inherent in collecting, managing, moving, analyzing, sharing, and archiving that data. Add the need to make sense of an increasingly diverse literature,[3] leverage complex simulation software, discover and access remote information sources,[4] establish and manage large distributed collaborations, and perform myriad other tasks, and it's a wonder that any work gets done at all in many laboratories. Furthermore, a recent survey estimated that US academics spend 42 percent of their research time engaged in administrative tasks.[5]

My dream is that one day soon, we'll see researchers running ambitious research programs from the same coffee shop as the entrepreneur. The key to realizing this dream will be the emergence of effective, low-cost providers of SaaS-based research tools, to which researchers can outsource time-consuming and routine aspects of the discovery process. If we succeed in this goal, we'll dramatically reduce costs and time demands and thus both accelerate discovery and allow many more people to participate in the research enterprise.

Oddly, most of the excitement within the research community around cloud computing has been focused on infrastructure as a service: on-demand computing and storage. In my opinion, that's a shortsighted and narrow view that misses the real benefit of the large-scale outsourcing and consequent economies of scale that the cloud is about. The biggest IT challenge facing research today isn't access to hardware but the complexity of the required business processes and associated IT technologies. Sure, terabytes demand new storage and computing solutions; but those resources are cheap, and, in any case, acquiring and managing computing and storage resources is only a small part of the overall problem. The complexity of managing the end-to-end research process is taking up all our time and stifling creativity. And that's where outsourcing can be transformative.

## Globus Online: Getting Data Moving

At the Computation Institute at the University of Chicago and Argonne National Laboratory, my team, along with several partners, is taking some first steps toward developing the tools and solutions needed to address this challenge. In particular, we recently launched Globus Online (www.globusonline.org), a new hosted service that lets you use powerful grid (that is, resource federation) capabilities without installing software.

Globus Online takes the task of moving large quantities of data from one place to another and packages it as a service. Using a Web browser, command line, or REST interface, you can ask Globus Online to move or synchronize files and directories — much as you might ask Amazon to ship a book. Globus Online handles the numerous tedious details of making transfers happen securely, efficiently, and reliably. It knows about commonly used data sources (much as Amazon knows about third-party vendors) and can easily be configured to know about more. You can also configure it with personal-profile information, such as which credentials to use with remote sites. (Think how Amazon can cache credit-card information.)

To authenticate with remote sites, Globus Online negotiates with the site and the user to determine a mutually satisfactory authentication mechanism and manages the retrieval and transmittal of required credentials (much as Amazon delivers credit-card information to third-party vendors). To achieve performance, it leverages specialized data transfer protocols (such as GridFTP; www.gridftp.org) and configures transfers to perform well over high-speed networks (like Amazon enables the use of different delivery methods). To achieve reliability, Globus Online, like FedEx, retries deliveries in the event of failures and gives up only when a deadline is reached.

Another important feature is the Globus Connect client, which solves the "last-mile" problem in file transfer. Although Globus GridFTP is available at most major research computing facilities, your files often need to get to or from other places: a laptop, lab server, department cluster, or scientific instrument. These computers probably don't have GridFTP installed, might be behind a firewall or NAT, and might not provide administrative privileges. Globus Connect solves this problem by requiring only outbound connections to transfer data — and thus works behind most firewalls and NATs. You can run it either temporarily or on a long-term basis, and it doesn't require administrative privileges or knowledge. (The Amazon analogies are getting a bit forced, but you might think of Globus Connect as one-click purchasing.) I've started using it to back up my laptop.

I joke that the target audience for Globus Online is anyone whose data is in the wrong place. In our initial work, we're particularly targeting users of high-performance

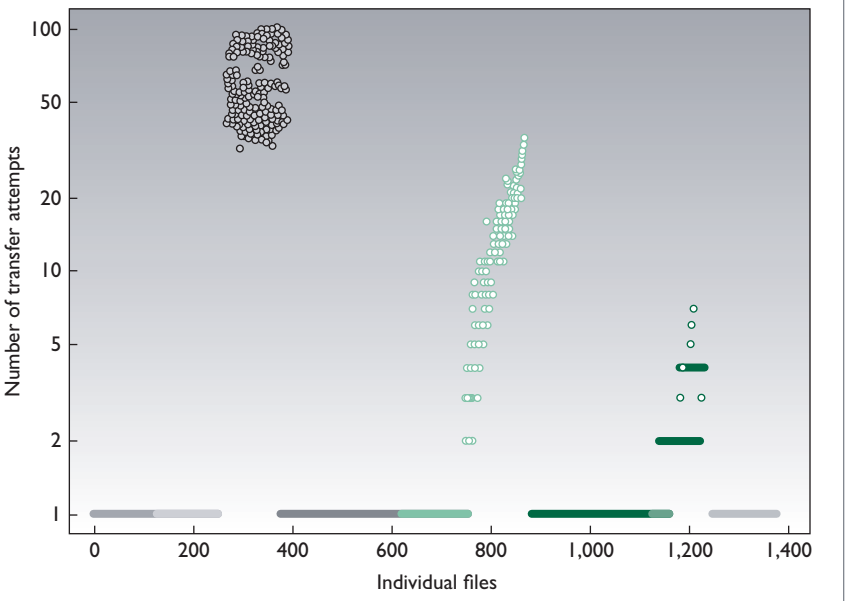



*Figure 1. A usage report from a Globus Online transfer. The report involved 125 files sent from one source to each of 11 destinations. The x-axis shows file numbers and the y-axis the number of transfer attempts required for success. (Note the log scale.) Different sites are color coded. The third, seventh, and tenth destinations had difficulties.*

computing facilities who routinely need to move data between centers and their local machines, and also between different centers. For example, an Indiana University physics researcher recently used Globus Online to move 730 Gbytes of simulation output from a supercomputer in Tennessee (Kraken) to one in Texas (Longhorn) — in just 1.5 hours. Additionally, we used Globus Online to move 300,000 files totaling 586 terabytes from Argonne to the National Energy Research Scientific Computing center (NERSC) in California and Oak Ridge National Laboratory in Tennessee in a few weeks, with no user involvement. This scale of data movement is unusual even for big science projects; for ordinary users to do it without installing any software is unprecedented.

## The Initial Use Case: Why Data Movement?

You might complain that data movement is a mundane task. It is indeed. But, in a very real sense, this is the point. Many people have data to move, but who wants to become a data movement specialist? Yet moving terabytes reliably and efficiently can be surprisingly complicated. You need to discover end points, determine available protocols, negotiate firewalls, configure software, manage space, negotiate authentication, configure protocols, detect and respond to failures, determine expected and actual performance, identify, diagnose, and correct network misconfigurations, integrate with file systems, and a host of other things. These are all tasks that can be automated to a significant extent, via software that we can operate far more easily and cost-effectively than can individual researchers.

We also find that the SaaS approach has value for troubleshooting. For example, one early user employed Globus Online to move data to 11 sites across the US. All completed successfully, but our monitoring showed that three sites had high transfer-retry rates (see Figure 1). Further investigation revealed a misconfigured firewall at one site and old GridFTP servers at two others. These were problems that had gone unnoticed for months. We got them fixed within days.

In defining Globus Online, we chose not to focus on hosting scientific application software. Compelling examples demonstrate how valuable such services can be: see, for example, the Network Enabled Optimization Service (NEOS),[6] nanoHUB[7] and other HUBs, and the various science gateways the TeraGrid supports.[8] These services provide a valuable function, but for now we address the more tedious task of data management.

Globus Online complements, rather than competes with, many other cloud services. For example, I frequently use Dropbox and YouSendIt to share files. But neither system helps me move terabytes at gigabit-per-second speeds. I'm also a fan of Amazon Web Services, and indeed we host Globus Online on Amazon, using EC2 to host transfer management and monitoring logic, S3 to backup Globus Online's state, and the Elastic Load Balancer for load balancing and fail-over between server instances running in multiple Amazon data centers. I also use Amazon for computing and storage. But again, not all my data and computing is on Amazon, so Amazon alone can't meet all of my needs.

## Looking to the Future

Embracing the elimination of tedium as our mission, we're rapidly expanding Globus Online's capabilities. As an example, one time-consuming activity frequently associated with data movement is controlling who can access files and datasets. Thus, we're expanding Globus Online to support data-sharing functions. If I share one file with a group of people, I'll often end up sharing other files with them. So, we're integrating the important (and tedious) function

of group management via Internet2's Grouper (www.internet2.edu/grouper). With personal profiles and group management in place, it becomes easy to interface to other software that researchers often struggle to configure properly, such as wikis and mail list managers. Thus, Globus Online is rapidly taking on important characteristics of the research enablement platform that we ultimately aspire to produce.

Looking forward, I believe that we must also look hard at yet more mundane tasks that occupy so much researcher time. In the biomedical field, for example, researchers know that applying for Institutional Review Board (IRB) approval for new studies and keeping track of who is a member of an IRB study can occupy a great deal of time. Surely, SaaS providers can automate and manage the purely bureaucratic aspects of those tasks, and do the same for grant reporting as well.

As we move further down this path, we start to approach our goal of a coffee shop research lab. Let me close by imagining the (hopefully not too distant) future: A researcher is enjoying a cup of coffee and pondering new methods for responding to cholera outbreaks. She has an idea for a new drug. Logging on to her virtual lab, she launches a computational workflow aimed at screening small molecules that might have the desired effect. She also cranks up a set of automated experiments designed to test promising candidates as they're identified. She locates an epidemiological simulator to study the impact of reducing bacterial survival rates on epidemic intensity. Simultaneously, she's sharing results and collecting inputs from collaborators — and screening the enormous literature on her subject for related results. These computations and experiments aren't performed in her own laboratory but at low-cost, high-throughput third-party facilities. With services and tools emerging to make this image a reality, it starts to seem quite feasible for first-class science to be performed far more cheaply, and by many more people, than today.


### Acknowledgments

Globus Online is the work of many people in the Computation Institute and partner institutions: see www.globusonline.org for details. This work is supported by the University of Chicago; the US Department of Energy, under contract number DE-AC02-06CH11357; and the US National Science Foundation, under contract OCI-534113.


### References


1. I. Foster and C. Kesselman, "Computational Grids," *The Grid: Blueprint for a New Computing Infrastructure*, I. Foster and C. Kesselman, eds., Morgan Kaufmann, 1999, pp. 2–48.
2. S.D. Kahn, "On the Future of Genomic Data," *Science*, vol. 331, no. 6018, 2011, pp. 728–729.
3. J.A. Evans and J.G. Foster, "Metaknowledge," *Science*, vol. 331, no. 6018, 2011, pp. 721–725.
4. I. Foster, "Service-Oriented Science," *Science*, vol. 308, no. 5723, 2005, pp. 814–817.
5. S. Rockwell, "The FDP Faculty Burden Survey," *Research Management Rev.*, vol. 16, no. 2, 2000, pp. 28–41.
6. J. More, J. Czyzyj, and M. Mesnier, "The NEOS Server," *IEEE J. Computational Science and Eng.*, vol. 5, 1998, pp. 68–75.
7. G. Klimeck et al., "nanoHUB.org: Advancing Education and Research in Nanotechnology," *Computing in Science and Eng.*, vol. 10, no. 5, 2008, pp. 17–23.
8. N. Wilkins-Diehr et al., "TeraGrid Science Gateways and Their Impact on Science," *Computer*, vol. 41, no. 11, 2008, pp. 32–41.



**Ian Foster** is the Arthur Holly Compton Distinguished Service Professor of computer science at the University of Chicago and a distinguished fellow at Argonne National Laboratory. His research interests include distributed, parallel, and data-intensive computing, and their applications in the sciences. Foster has a PhD in computer science from Imperial College, London. In 2011, he received the IEEE's Tsutomu Kanai award. Contact him at foster@anl.gov.