Ian Foster and the Globus Project

Charles Severance, *University of Michigan*

Ian Foster describes how the Globus project helps move large amounts of data efficiently and safely, allowing scientists to focus on their research and not on IT problems.

Before distributed utility computing was referred to as "the cloud," it was called "the grid." In the early and mid-1990s, a great deal of research focused on developing software libraries that would enable a networked "cluster" of low-priced personal computers and workstations to tackle large computing and storage problems that had previously required very expensive supercomputers to handle. We knew that if we could efficiently harness the power of hundreds or thousands of inexpensive computers, it would be less expensive than a single supercomputer.

By the end of the 1990s, the Globus Toolkit emerged as the most widely used of these software libraries that enabled distributed computing resources to be assembled into grids. I talked to Ian Foster, director of the Computation Institute and pioneer of the Globus project, about the project's early history as well as Globus's more recent work in distributed computing. You can see the full interview at www.computer.org/computingconversations.

THE NETWORK BECOMES THE COMPUTER

The National Science Foundation Network (NSFNet) was initially intended to connect researchers to supercomputers. But the side effect was that research universities became increasingly connected both within and among campuses. It didn't take long to see these always-connected computing systems as a resource to be aggregated to solve larger problems:

> I'm a bit of a technological determinist. I think that innovations often occur when certain technologies are in place that can allow them to happen. Arguably, the Web came about when our disk drives were big enough to store interesting numbers of images and the networks were fast enough to communicate those images. The work we did on the grid came about when high-speed networks were such that we could reasonably outsource computing to remote computers.

The goal was to run a standard software suite on computing hardware from many vendors so researchers could connect up and use whatever resources they needed, much like the power grid. In the early days, the challenge was the diversity of the hardware and operating systems we were running. Given that TCP/IP networking was still a relatively recent innovation in the 1990s, each operating system had its own quirks:

> We ended up spending a lot of time negotiating the vagaries of the different network protocols and operating systems that already existed. A lot of the work in the early days was working out how to connect things—how to federate computers at 10, 100, or 1,000 different locations or data stores at those different locations.

As networking, operating systems, connectivity, and hardware technologies evolved in the early 2000s, the Globus Toolkit went through several iterations. It was successful both in handling distributed computing and

Published by the IEEE Computer Society
0018-9162/14/$31.00 © 2014 IEEE

in managing data to and from many different sources. While Globus continues to be used on dedicated hardware, it's increasingly used to provide cloud-based software-as-a-service (SaaS) offerings for scientific storage and computation using commodity cloud-based infrastructures:

> What's happened over the last five years in particular has been this notion of computing as a utility. It's been effectively monetized and commercialized by the likes of Amazon and Microsoft, which is very exciting. Right now, I'm particularly interested in how we can leverage this new cloud-computing paradigm to accelerate discovery in the sciences.

While Globus still offers a set of software protocols connecting things you can download and install on your particular computer, you can also access hosted Globus services over the network just like many commercial cloud services. The cloud services offering was initially called Globus Online but is now simply called Globus. The first service that was deployed in this manner was Globus Transfer, which built on the highly successful and reliable GridFTP capabilities from the Globus Toolkit:

> We decided that if we were going to work out how to outsource activities to the cloud, we should start with something very mundane and simple that very few people would be concerned with. We decided to take on data movement because it ended up being one of the most important uses of Globus: simply managing the movement of data between the many nodes of a computing grid or a research collaboration, or perhaps moving data from a scientific instrument like a genome sequencing machine to an analytic computer and so forth.

Although this might seem to be a simple task that's suitable to services like Box or Dropbox, when you're dealing with millions of files and many Tbytes of data streaming from a genome sequencer for many weeks, the problem is significantly more complex:

> First, if you want to move data reliably, securely, and rapidly from A to B, there are complex security issues to negotiate when dealing with supercomputer and genome sequencing centers. Second, if you're trying to move 100 Gbytes or 100 Tbytes, things will inevitably fail along the way. Disk servers or the network might get overloaded, a wire could be cut, there might be a power outage—any of hundreds of things can happen. Software needs to be able to detect and respond to these problems. A lot of the logic in Globus Transfer is about ultra-reliable transfers. We have protocols that can restart transfers where they left off. We have restart markers built into the protocols so we can detect how far we've gone at any one point in time.

Ultimately, Globus is sophisticated plumbing that ensures the reliable transfer of a few Tbytes of data from one source to another. Its reliability in the face of any number of network or server problems is critical. But while building scalable and reliable data transfers is an interesting computer science problem, the real goal is to advance scientific research:

> We're interested in moving data because we want to accelerate scientific processes that depend on many things like discovering data, computing on data, and allowing people to collaborate on the processes that turn data into knowledge. We're exploring opportunities for outsourcing many of these activities to cloud computing resources. I think that's where the excitement lies in the future.

The availability of commodity networking and computing infrastructure has helped Globus move up the value chain, and has reduced the effort required to participate in distributed collaborative science. But there's still more to be done to improve the efficiency of scientific collaboration:

Right now, I'm particularly interested in how we can leverage this new cloud-computing paradigm to accelerate discovery in the sciences.

> As you start to explore how the management and movement of scientific data using cloud resources can be improved, you start to look at how the processes of science itself—the processes of discovery—change. Do we end up with a more collaborative or collective big science approach to more problems, not just in the physical and biological sciences but also in the social sciences?

As the Globus project approaches two decades of research in distributed computing, it continues to make use of the latest technologies and to focus on using technology to advance scientific discovery.

Charles Severance, *Computing Conversations column editor and* Computer*'s multimedia editor, is a clinical associate professor and teaches in the School of Information at the University of Michigan. Follow him on Twitter @drchuck or contact him at csev@umich.edu.*