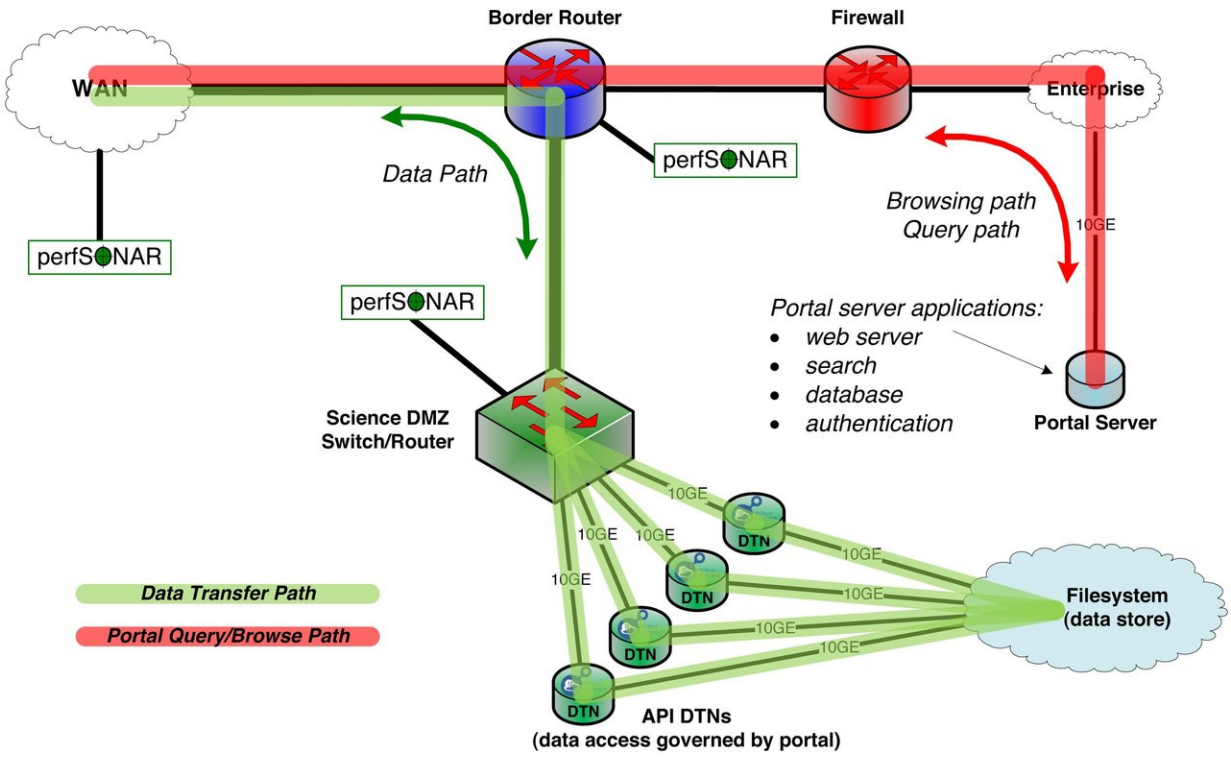
Networking, data experts design a better portal for scientific discovery

January 29 2018, by Jon Bashor



The Science DMZ includes multiple DTNs that provide for high-speed transfer between network and storage. Portal functions run on a portal server, located on the institution's enterprise network. The DTNs need only speak the API of the data management service (Globus in this case). Credit: Eli Dart, ESnet

These days, it's easy to overlook the fact that the World Wide Web was created nearly 30 years ago primarily to help researchers access and

share scientific data. Over the years, the web has evolved into a tool that helps us eat, shop, travel, watch movies and even monitor our homes.

Meanwhile, scientific instruments have become much more powerful, generating massive datasets, and international collaborations have proliferated In this new era, the web has become an essential part of the scientific process, but the most common method of sharing research data remains firmly attached to the earliest days of the web. This can be a huge impediment to scientific discovery.

That's why a team of networking experts from the Department of Energy's Energy Sciences Network (ESnet), with the Globus team from the University of Chicago and Argonne National Laboratory, have designed a new approach that makes data sharing faster, more reliable and more secure. In an article published Jan. 15 in *Peer J Comp Sci*, the team describes their "The Modern Research Data Portal: a design pattern for networked, data-intensive science."

"Both the size of datasets and the quantity of data objects has exploded, but the typical design of a data portal hasn't really changed," said co-author Eli Dart, a network engineer with the Department of Energy's Energy Sciences Network, or ESnet. "Our new design preserves that ease of use, but easily scales up to handle the huge amounts of data associated with today's science."

Data portals, sometimes called science gateways, are web-based interfaces for access data storage and computing systems, allowing authorized users to access data and perform shared computations. As science becomes increasingly data-driven and collaborative, data portals are advancing research in materials, physics, astrophysics, cosmology, climate science and other fields.

The "legacy" portal is driven by a web server that is connected to a

storage system and a database and processes users' requests for data. While this simple design was straightforward to develop 25 years ago, it has increasingly become an obstacle to performance, usability and security.

"The problem with using old technology is that these portals don't provide fast access to the data and they aren't very flexible," said lead author Ian Foster, who is the Arthur Holly Compton Professor at the University of Chicago and Director of the Data Science and Learning Division at Argonne National Laboratory. "Since each portal is developed as its own silo, the organization therefore must implement, and then manage and support, multiple complete software stacks to support each portal."

The new portal design is built on two approaches developed to simplify and speed up transfers of large datasets.

- The Science DMZ, which Dart developed, is a high-performance network design that connects large-scale data servers directly to high-speed networks and is increasingly used by research institutions to better manage data transfers.
- Globus is a cloud-based service to which developers of data portals and other science services can outsource responsibility for complex tasks like authentication, authorization, data movement, and data sharing. Globus can be used, in particular, to drive data transfers into and out of Science DMZs.

Kyle Chard, Foster, David Shiffett, Steven Tuecke and Jason Williams are co-authors of the paper and helped develop Globus at Argonne National Laboratory and the University of Chicago. In their paper, the authors note that the concept became feasible in 2015 as Globus and the Science DMZ became mature technologies.

"Together, Globus and the Science DMZ give researchers a powerful toolbox for conducting their research," Dart said.

One portal incorporating the new design is the Research Data Archive managed by the National Center for Atmospheric Research, which contains a large and diverse collection of meteorological and oceanographic observations, operational and reanalysis model outputs, and remote sensing datasets to support atmospheric and geosciences research.

For example, a scientist working at a university could download data from the National Center for Atmospheric Research (NCAR) in Colorado and then use it to run simulations at DOE and NSF supercomputing centers in California and Illinois, and finally move the data to her home institution for analysis. To illustrate how the design works, Dart selected a 460-gigabyte dataset at NCAR, initiated a Globus transfer to DOE's National Energy Research Scientific Computing Center at Lawrence Berkeley National Laboratory, logged in to his storage account and started the transfer. Four minutes later, the 5,141 files had been seamlessly transferred.

How the design works

The Modern Research Data Portal takes the single-server model of the traditional portal design and divides it among three distinct components.

- A portal web server handles the search for and access to the specified data, and similar tasks.
- The data servers, often called Data Transfer Nodes, are connected to high-speed networks through a specialized enclave, in this case the Science DMZ. The Science DMZ provides a dedicated, secure link to the data servers, but avoids common performance bottlenecks caused by typical designs not optimized

for high-speed transfers.
- Globus manages the authentication, data access and data transfers. Globus makes it possible for users to manage data irrespective of the location or storage system on which data reside and supports data transfer, sharing, and publication directly from those storage systems.

"The design pattern thus defines distinct roles for the web server, which manages who is allowed to do what; data servers, where authorized operations are performed on data; and external services, which orchestrate data access," the authors wrote.

Globus is already used by tens of thousands of researchers worldwide with endpoints at more than 360 sites, so many researchers are familiar with its capabilities and rely on it on a regular basis. In fact, about 80 percent of major research universities and national labs in the U.S. use Globus.

At the same time, more than 100 research universities across the country have deployed Science DMZs, thanks to funding support through the National Science Foundation's Campus Cyberinfrastructure Program.

A critical component of the system is "a little agent called Globus Connect, which is much like the Google Drive or Dropbox agents one would install on their own PCs," Chard said. Globus Connect allows the Globus service to move data to and from the computer using high performance protocols and also HTTPS for direct access. It also allows users to share data dynamically with their peers.

According to Chard, the design provides research organizations with easy-to-use technology tools similar to those used by business startups to streamline development.

"If we look to industry, startup businesses can now build upon a suite of services to simplify what they need to build and manage themselves," Chard said. "In a research setting, Globus has developed a stack of such capabilities that are needed by any research portal. Recently, we (Globus) have developed interfaces to make it trivial for developers to build upon these capabilities as a platform."

"As a result of this design, users have a platform that allows them to easily place and transfer data without having to scale up the human effort as the amount of data scales up," Dart said.

More information: Kyle Chard et al. The Modern Research Data Portal: a design pattern for networked, data-intensive science, *PeerJ Computer Science* (2018). DOI: 10.7717/peerj-cs.144

Provided by Lawrence Berkeley National Laboratory