ESnet's Petascale DTN project speeds up data transfers between leading HPC centers

December 11 2017

Operations staff monitor the network in the ESnet/NERSC control room. Credit: Marilyn Chung, Berkeley Lab

The Department of Energy's (DOE) Office of Science operates three of the world's leading supercomputing centers, where massive data sets are routinely imported, analyzed, used to create simulations and exported to other sites. Fortunately, DOE also runs a networking facility, ESnet (short for Energy Sciences Network), the world's fastest network for science, which is managed by Lawrence Berkeley National Laboratory.

Over the past two years, ESnet engineers have been working with staff at DOE labs to fine tune the specially configured systems called data transfer nodes (DTNs) that move data in and out of the National Energy Research Scientific Computing Center (NERSC) at Lawrence Berkeley National Laboratory and the leadership computing facilities at Argonne National Laboratory in Illinois and Oak Ridge National Laboratory in Tennessee. All three of the computing centers and ESnet are DOE Office of Science User Facilities used by thousands of researchers across the country.

The collaboration, named the Petascale DTN project, also includes the National Center for Supercomputing Applications (NCSA) at the University of Illinois in Urbana-Champaign, a leading center funded by the National Science Foundation (NSF). Together, the collaboration aims to achieve regular disk-to-disk, end-to-end transfer rates of one petabyte per week between major facilities, which translates to achievable throughput rates of about 15 Gbps on real world science data sets.

Research projects such as cosmology and climate have very large (multi-petabyte) datasets and scientists typically compute at multiple HPC centers, moving data between facilities in order to take full advantage of the computing and storage allocations available at different sites.

Since data transfers traverse multiple networks, the slowest link determines the overall speed. Tuning the data transfer nodes and the border router where a center's internal network connects to ESnet can smooth out virtual speedbumps. Because transfers over the wide area network have high latency between sender and receiver, getting the highest speed requires careful configuration of all the devices along the data path, not just the core network..

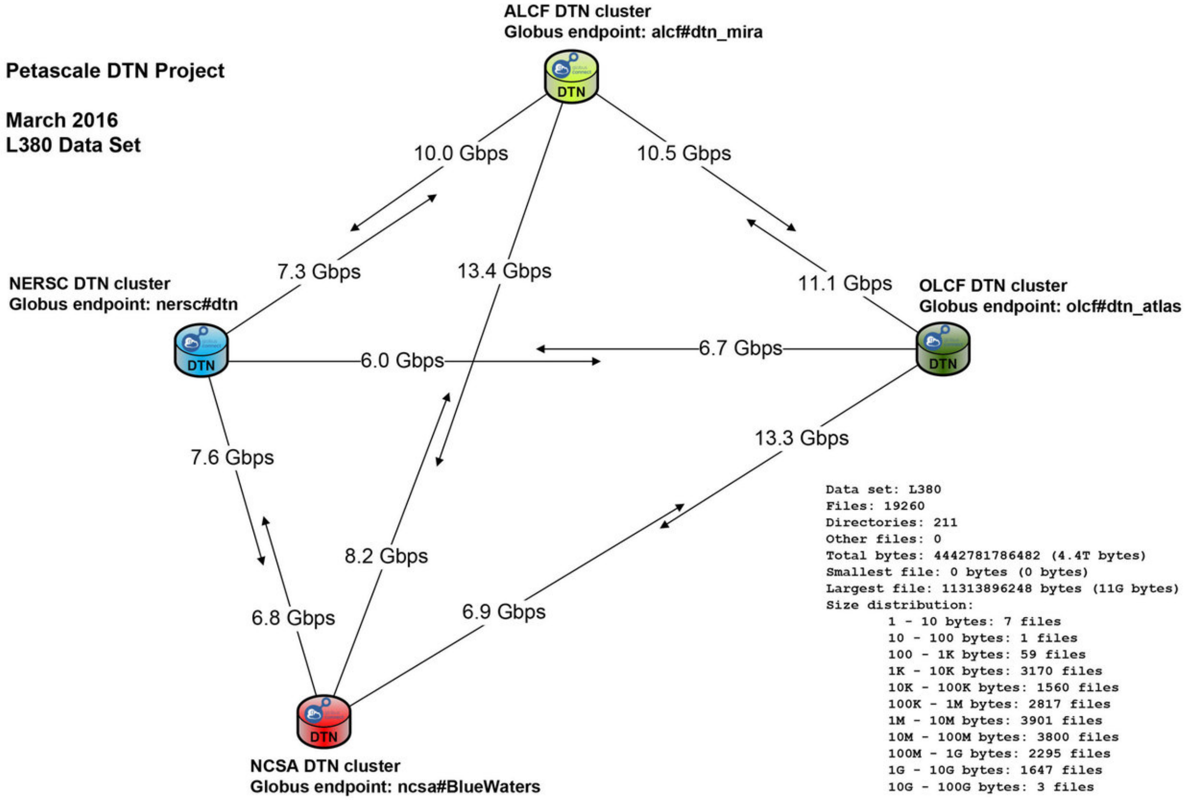
In the past few weeks, the project has shown sustained data transfers at well over the target rate of 1 petabyte per week. The number of sites with this base capability is also expanding, with Brookhaven National Laboratory in New York now testing its transfer capabilities with encouraging results. Future plans including bringing the NSF-funded San Diego Supercomputer Center and other big data sites into the mix.

"This increase in data transfer capability benefits projects across the DOE mission science portfolio" said Eli Dart, an ESnet network engineer and leader of the project. "HPC facilities are central to many collaborations, and they are becoming more important to more scientists as data rates and volumes increase. The ability to move data in and out of HPC facilities at scale is critical to the success of an ever-growing set of projects."

When it comes to moving data, there are many factors to consider, including the number of transfer nodes and their speeds, their utilization, the file systems connected to these transfer nodes on both sides, and the network path between them, according to Daniel Pelfrey, a high performance computing network administrator at the Oak Ridge Leadership Computing Facility.

The actual improvements being made range from updating software on the DTNs to changing the configuration of existing DTNs to adding new nodes at the centers.

"Transfer node operating systems and applications need to be configured to allow for WAN transfer," Pelfrey said. "The connection is only going to be as fast as the slowest point in the path allows. A heavily utilized server, or a misconfigured server, or a heavily utilized network, or heavily utilized file system can degrade the transfer and make it take much longer."



```
Data set: L380
Files: 19260
Directories: 211
Other files: 0
Total bytes: 4442781786482 (4.4T bytes)
Smallest file: 0 bytes (0 bytes)
Largest file: 11313896248 bytes (11G bytes)
Size distribution:
        1 - 10 bytes: 7 files
        10 - 100 bytes: 1 files
        100 - 1K bytes: 59 files
        1K - 10K bytes: 3170 files
        10K - 100K bytes: 1560 files
        100K - 1M bytes: 2817 files
        1M - 10M bytes: 3901 files
        10M - 100M bytes: 3800 files
        100M - 1G bytes: 2295 files
        1G - 10G bytes: 1647 files
        10G - 100G bytes: 3 files
```

Performance data from March 2016 showing transfer rates between facilities. Credit: Eli Dart, ESnet

At NERSC, the DTN project resulted in adding eight more nodes, tripling the number, in order achieve enough internal bandwidth to meet the project's goals. "It's a fairly complicated thing to do," said Damian Hazen, head of NERSC's Storage Systems Group. "It involves adding infrastructure and tuning as we connected our border routers to internal routers to the switches connected to the DTNs. Then we needed to install the software, get rid of some bugs and tune the entire system for optimal performance."

The work spanned two months and involved NERSC's Storage Systems, Networking, and Data and Analytics Services groups, as well as ESnet, all working together, Hazen said.

At the Argonne Leadership Computing Facility, the DTNs were already in place and with minor tuning, transfer speeds were increased to the 15 Gbps.

"One of our users, Katrin Heitmann, had a ton of cosmology data to move and she saw a tremendous benefit from the project," said Bill Allcock, who was director of operations at the ALCF during the project. "The project improved the overall end-to-end transfer rates, which is especially important for our users who are either moving their data to a community archive outside the center or are using data archived elsewhere and need to pull it in to compute with it at the ALCF."

As a result of the Petascale DTN project, the OLCF now has 28 transfer nodes in production on 40-Gigabit Ethernet. The nodes are deployed under a new model—a diskless boot—which makes it easy for OLCF staff to move resources around, reallocating as needed to respond to users' needs.

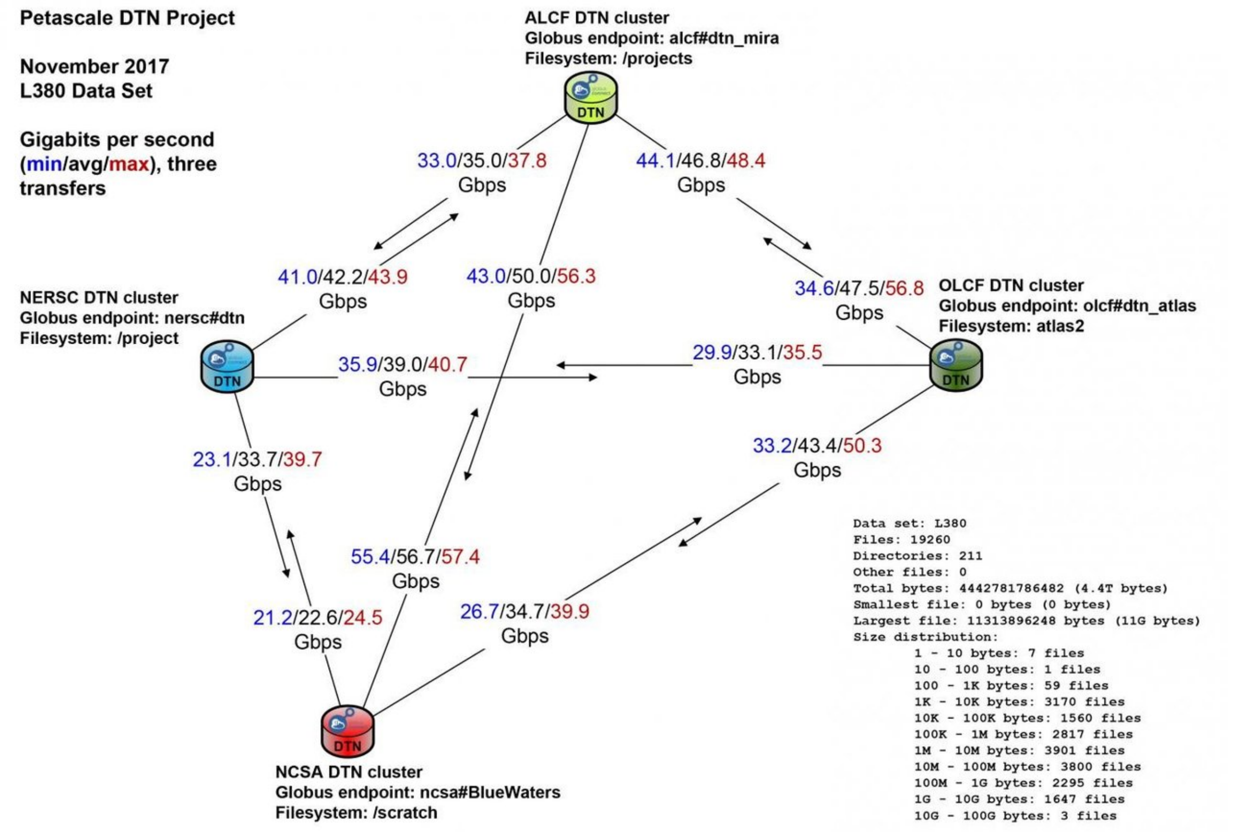
"The Petascale DTN project basically helped us increase the 'horsepower under the hood' of network services we provide and make them more resilient," said Jason Anderson, an HPC UNIX/storage systems administrator at OLCF. "For example, we recently moved 12TB of science data from OLCF to NCSA in less than 30 minutes. That's fast!"

Anderson recalled that a user at the May 2017 OLCF user meeting said that she was very pleased with how quickly and easily she was able to move her data to take advantage of the breadth of the Department of Energy's computing resources.

"When the initiative started we were in the process of implementing a Science DMZ and upgrading our network," Pelfrey said. "At the time, we could move a petabyte internally in 6-18 hours, but moving a petabyte externally would have taken just a bit over a week. With our latest upgrades, we have the ability to move a petabyte externally in about 48 hours."

The fourth site in the project is the NSF-funded NCSA in Illinois, where senior network engineer Matt Kollross said it's important for NCSA, the only non DOE participant, to collaborate with other DOE HPC sites to develop common practices and speed up adoption of new technologies.

"The participation in this project helped confirm that the design and investments in network and storage that we made when building Blue Waters five years ago were solid investments and will help in the design of future systems here and at other centers," Kollross said. "It's important that real-world benchmarks which test many aspects of an HPC system, such as storage, file systems and networking, be considered in evaluating overall performance of an HPC compute system and help set reasonable expectations for scientists and researchers."



```
Data set: L380
Files: 19260
Directories: 211
Other files: 0
Total bytes: 4442781786482 (4.4T bytes)
Smallest file: 0 bytes (0 bytes)
Largest file: 11313896248 bytes (11G bytes)
Size distribution:
        1 - 10 bytes: 7 files
        10 - 100 bytes: 1 files
        100 - 1K bytes: 59 files
        1K - 10K bytes: 3170 files
        10K - 100K bytes: 1560 files
        100K - 1M bytes: 2817 files
        1M - 10M bytes: 3901 files
        10M - 100M bytes: 3800 files
        100M - 1G bytes: 2295 files
        1G - 10G bytes: 1647 files
        10G - 100G bytes: 3 files
```

Performance measurements from November 2017 at the end of the Petascale DTN project. All of the sites met or exceed project goals. Credit: Eli Dart, ESnet

Origins of the project

The project grew out of a Cross-Connects Workshop on "Improving Data Mobility & Management for International Cosmology," held at Berkeley Lab in February 2015 and co-sponsored by ESnet and Internet2.

Salman Habib, who leads the Computational Cosmology Group at Argonne National Laboratory, gave a talk at the workshop, noting that large-scale simulations are critical for understanding observational data and that the size and scale of simulation datasets far exceed those of observational data. "To be able to observe accurately, we need to create accurate simulations," he said.

During the workshop, Habib and other attendees spoke about the need to routinely move these large data sets between computing centers and agreed that it would be important to be able to move at least a terabyte a week. As the Argonne lead for DOE's High Energy Physics Center for Computational Excellence project, Habib had been working with ESnet and other labs on data transfer issues.

To get the project moving, Katrin Heitmann, who works in cosmology at Argonne, created a data package of small and medium files totaling about 4.4 terabytes. The data would then be used to test network links between the leadership computing facilities at Argonne and Oak Ridge national labs, the National Energy Research Scientific Computing Center (NERSC) at Lawrence Berkeley National Laboratory, and the National Center for Supercomputing Applications (NCSA) at the University of Illinois in Urbana-Champaign, a leading center funded by the National Science Foundation.

"The idea was to use the data as a test, to send it over and over and over between the centers," Habib said. "We wanted to establish a performance baseline, then see if we could improve the performance by eliminating any choke points."

Habib admitted that moving a petabyte in a week would only use a fraction of ESnet's total bandwidth, but the goal was to automate the transfers using Globus Online, a primary tool for researchers accessing high performance networks like ESnet for rapidly sharing data or to use remote computing facilities.

"For our research, it's very important that we have the ability to transfer large amounts of data," Habib said. "For example, we may run a simulation at one of the large DOE computing centers, but often where we run the simulation is not where we want to do the analysis. Each center has different capabilities and we have various accounts at the centers, so the data gets moved around to take advantage of this. It happens all the time."

Although the project's roots are in cosmology, the Petascale DTN project will help all DOE scientists who have a need to transfer data to, from, or between the DOE computing facilities to take advantage of rapidly advancing data analytics techniques. In addition, the increase in data transfer capability at the HPC facilities will improve the performance of data portals, such as the Research Data Archive at the National Center for Atmospheric Research, that use Globus to transfer data from their storage systems.

"As the scientists deal with data deluge and more research disciplines depend on high-performance computing, data movement between computing centers needs to be a no-brainer for scientists so they can take advantage of the compute cycles at all DOE Office of Science user facilities and the extreme heterogeneity of systems in the future" said ESnet Director Inder Monga.

Provided by Lawrence Berkeley National Laboratory

Citation: ESnet's Petascale DTN project speeds up data transfers between leading HPC centers (2017, December 11) retrieved 19 April 2024 from https://phys.org/news/2017-12-esnet-petascale-dtn-hpc-centers.html