Experience Report on the Supercomputing 2009 Bandwidth Challenge from the Perspective of the Earth System Grid

December 31, 2009

Prepared by
Dean N. Williams (Lawrence Livermore National Laboratory) and Alex Sim (Lawrence Berkeley National Laboratory)

With contributions from Raj Kettimuthu (Argonne National Laboratory), Eli Dart (Earth Sciences Network), Dan Gunter, Viji Natarajan (Lawrence Berkeley National Laboratory), Jeff Long (Lawrence Livermore National Laboratory), Jason Hick (National Energy Research Scientific Computing Center)

Introduction

The Supercomputing 2009 (SC09) Bandwidth Challenge entry titled, “*High Performance GridFTP Transport of Earth System Grid (ESG) Data,*” demonstrated high-performance GridFTP transport of climate data from multiple Department of Energy laboratories to the targeted SC09 showroom floor. The transferred multi-terabyte data consisted of a small portion of the multi-model Coupled Model Intercomparison Project, Phase 3 (CMIP-3) data set used in the Intergovernmental Panel on Climate Change (IPCC) Fourth Assessment Report (AR4).

The CMIP-3 data set consists of many small-size files spanning spatial and temporal time zones, making the “multiple small file” transfer optimizations in GridFTP, such as data-channel caching and pipelining, relevant to the efficiency of the data transfer protocol. The Bulk Data Mover (BDM), a higher-level data-transfer-management component, was used to manage the GridFTP transfers using optimized transfer-queue and concurrency management algorithms.

The Energy Sciences Network’s (ESnet’s) On-demand Secure Circuits and Advance Reservation System (OSCARS) provided the network bandwidth reservation capability. Dedicated circuits were used for the data transfers.

This Bandwidth Challenge activity, led by Argonne National Laboratory (ANL) with participation from the Argonne Leadership Computing Facility (ALCF), Energy Sciences Network (ESnet), Lawrence Berkeley National Laboratory (LBNL), Lawrence Livermore National Laboratory (LLNL), National Energy Research Scientific Computing Center (NERSC), and University of Utah, was particularly important for ESG and the climate community in preparing global data replication over the network for the upcoming CMIP-5, IPCC AR5 release. It demonstrated our current status in data replication over the network and indicated what needs to be done to make it better.

Hardware and Software Components

- The Green Data Oasis (GDO) at LLNL stores tens of terabytes (TB) of CMIP-3 multimodel data. Three GridFTP server nodes with Solaris 10 running ZFS on AMD-64 hardware were used with access to a 10-gigabit (Gb) ESnet network. Between LLNL and the SC09 showroom floor, ESnet reserved a 5-Gb Science Data Network (SDN) through OSCARS.
- Two NERSC Data Transfer Nodes (DTNs) were used to move data located on NERSC storage units to the SC09 showroom floor. Between NERSC and the SC09 showroom floor, ESnet reserved a 10-Gb SDN through OSCARS.
- Twenty GridFTP servers at ALCF, converted from the 20 Eureka computing nodes with a special arrangement, were used to transfer data located on the ALCF General Parallel File System (GPFS) to the SC09 showroom floor. These temporary transfer nodes had 1-Gb-per-second (Gbps) connections to the 10-Gb ESnet network. Between ALCF and the SC09 showroom floor, ESnet reserved a 10-Gb SDN through OSCARS.
- A Data Direct Networks S2A9900 high-performance storage platform was used at the SC09 showroom floor to store the transferred data and allow further processing. The Parallel Virtual File System was used initially to allow parallel access to the disk subsystem from the GridFTP servers. Intel Nehalem machines with a 10-Gb Ethernet card were used to drive the data transfers. The 10-Gb Ethernet on these machines was connected to the SC09 network through a switch.
- In collaboration with the Scientific Discovery through Advanced Computing Visualization and Analytics Center for Enabling Technologies team, a few high-quality, three-dimensional (3D) visualizations were produced using a new integration of the Climate Data Analysis Tools and the Visualization Streams for Ultimate Scalability technologies, which gave access to the ESG computational resources with advanced 3D data presentation capabilities. The visualization showed an animation spanning 200 years, from 1900 to 2100, of multi-model averaged surface temperatures and 16 levels of atmospheric temperatures.
- The NetLogger Toolkit was used to collect and analyze the monitoring information from all GridFTP servers. These data were visualized in approximate real-time.
- GridFTP servers were used from Globus Toolkit v4.0.8 (NERSC) and v4.2.1 (ALCF, LLNL, and SC09 showroom floor).
- Globus-url-copy, a commonly used command-line scriptable client for GridFTP, was used as one of the key data-movement tools. It supports multiple techniques, such as TCP streams and concurrent transfers, to achieve high performance. The newly added reliability and load-balancing capabilities in globus-url-copy played a key role in the challenge.

- Globus.org, a hosted data-movement service, was used to mirror parts of source data from NERSC at ALCF. Globus.org focuses on providing reliable, high-performance, fire-and-forget data transfer. If problems with the endpoints prevent a transfer from completing, Globus.org will periodically retry until the endpoints are working in order to complete the transfer.
- BDM was used with the GridFTP client library from CoG-JGlobus 1.7.0. High performance was achieved using a variety of techniques, including multithreaded concurrent-transfer-connection management, transfer-queue management, and single-control-channel management for multiple data transfers. The GridFTP library supports data-channel caching and pipelining.

Setup

Data source was set up at ALCF, NERSC, and LLNL. Size of the data set was about 10 terabytes (TB), and the transfers were partitioned into three groups with 4 TB from ALCF, 4 TB from NERSC, and 2 TB from LLNL. Network bandwidth on the SC09 showroom floor was 20 Gb. SDN was reserved for the transfers through ESnet OSCARS at 10 Gb from ALCF, 10 Gb from NERSC, and 5 Gb from LLNL. The data set was partitioned in this way because 4 TB is the volume of file transfers for about 1 hour over a 10-Gb connection. When network bandwidth was fully utilized under this setup, 10 TB could be transferred in about 1.5 hours. Figure 1 shows how the network was set up and how data flowed. The SC09 demonstration achieved about 15 Gbps on average and moved about 7 TB of data in 1 hour from three data sources.

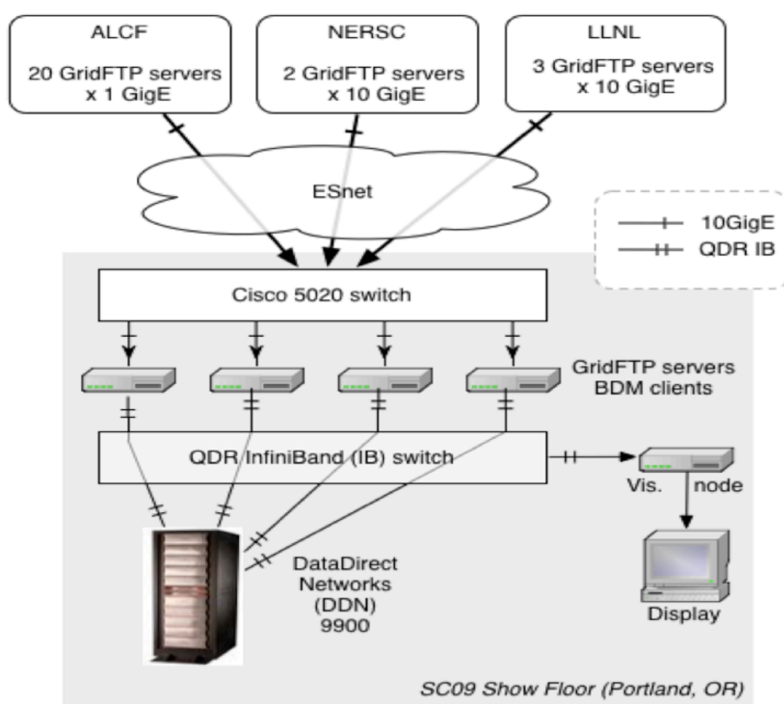


Figure 1. ***SC09 data and network setup diagram. (Courtesy of Dan Gunter, LBNL and Raj Kettimuthu, ANL).***

Experiences and To-Do Items

Network and Data Transfer Nodes with GridFTP Servers

- BDM test runs from GDO at LLNL to DTNs at NERSC showed that a single GDO GridFTP server is capable of 1.5–2.0 Gbps sustained. Dedicated testing from LLNL to NERSC showed that BDM could achieve 5 Gbps from GDO to NERSC DTNs when three separate source GridFTP servers are used.
- Raj Kettimuthu at ANL noted that the network latency between LLNL and NERSC is as low as a few milliseconds, and when the network latency is high, a few issues would be raised. From LLNL to the SC09 showroom floor, the network latency was at least 10 times higher than between LLNL and NERSC, which caused transfer rates as low as 1 Gbps from GDO at LLNL to the SC09 showroom floor.
- Jeff Long at LLNL suggested having a perfSonar server at each end of the network circuit. This configuration would make network testing and benchmarking much easier and more precise. Eli Dart at ESnet also suggested perfSonar servers at each ESG site.
- Long mentioned that the current firewall at LLNL does not support jumbo frames (MTU 9000) and imposes limits on per connection and overall throughput. A newer 10-Gb Ethernet firewall for GDO would improve the throughput.
- Dart noted that the transfer rates from LLNL averaged 30 megabits per second (Mbps) about 6 months ago and that the current rate is a big improvement. The SC09 effort resulted in a significant improvement in production infrastructure.
- Dart suggested having a Wide Area Network working group in ESG to address network data transfer and performance issues among participating ESG data replication sites.
- SUN Solaris does not work with TCP auto-tuning. Long noted that setting the GridFTP blocksize parameter in the server configuration and in the client options improved transfer performance. Dart recommends having a couple of Linux boxes in front of GDO/SUNFire for wide area transfers only, similar to the DTN servers at NERSC.
- Jason Hick at NERSC mentioned that the SC09 effort pushed NERSC DTN systems and identified several different configuration problems that have been resolved for production usage.
- /etc/xinetd.conf could limit the transfers by the following entries:
  - cps = 50 10

    Number defines the maximum of connections per second. This directive takes two integer arguments separated by white space. The first one is the maximum number of connections allowed to the service per second. The second one is the number of seconds xinetd must wait before re-enabling the service when the first one is maxed out. For example, the above numbers indicate no more than 50 connections per second are allowed to any given service. If this limit is reached, the service is retired for 10 seconds.

  - o instances = 50
    Number sets the maximum number of requests xinetd can handle at once
  - o per_source = 10
    Number defines the maximum number of instances for a service per source IP address.
- The data mirroring activity performed as a precursor to the challenge helped us identify and fix configuration problems in GridFTP servers at both ALCF and NERSC.
- ALCF DTN nodes (gs1 and gs2) had a limit of 250 Mbps on backend storage network connectivity with GPFS per GridFTP server. Loren **(need full name)** at ALCF noted that this limit is the current practical maximum (500 Mbps from both DTNs) with a bottleneck on GPFS. ALCF also has a known packet-loss issue from anyone on SDN to ALCF and from ALCF to anyone over SDN or ESnet. ALCF is working on these issues.
- At ALCF, 20 computing nodes hosted GridFTP servers, each having a 1-Gbps connection to a 10-Gb ESnet network. This configuration solved the performance limit on ALCF DTN nodes and proved that many small-capacity machines are capable of filling the large bandwidth with well-managed transfers.

NetLogger Monitoring (Contributed by Dan Gunter at LBNL)

- Data collection for the most part went smoothly. The main problem was that syslog-ng servers at the remote sites would occasionally get “stuck” and stop sending data. This problem was solved with a simple periodic “kill –HUP” of the server. We found this command was preferable to using the syslog-ng “restart” command, which would re-send the entire log from the start.
- The Web page developed specifically for SC09 was a good example of the power of simplicity. The interface for retrieving graphs of data was a simple HTTP GET of a URL (that is, a REST API) that contained descriptive parameters such as time range and graph type. This interface was very effective for fast development and allowed URLs for graphs to be easily shared. The R back-end allowed for cumulative numbers and different ways of grouping the data to be quickly added to the set of available graphs.
- The estimated cumulative bandwidth from the GridFTP measurements was somewhat crude—each transfer was assumed to be proceeding at its average rate from start to end—which resulted in a somewhat inaccurate estimate when compared with the ground-truth numbers from the SC09 network routers. This method needs to be improved. However, these results showed that rough estimates, when at least wrong in a consistent way, can still be quite useful.

GridFTP and globus-url-copy (Contributed by Kettimuthu at ANL)

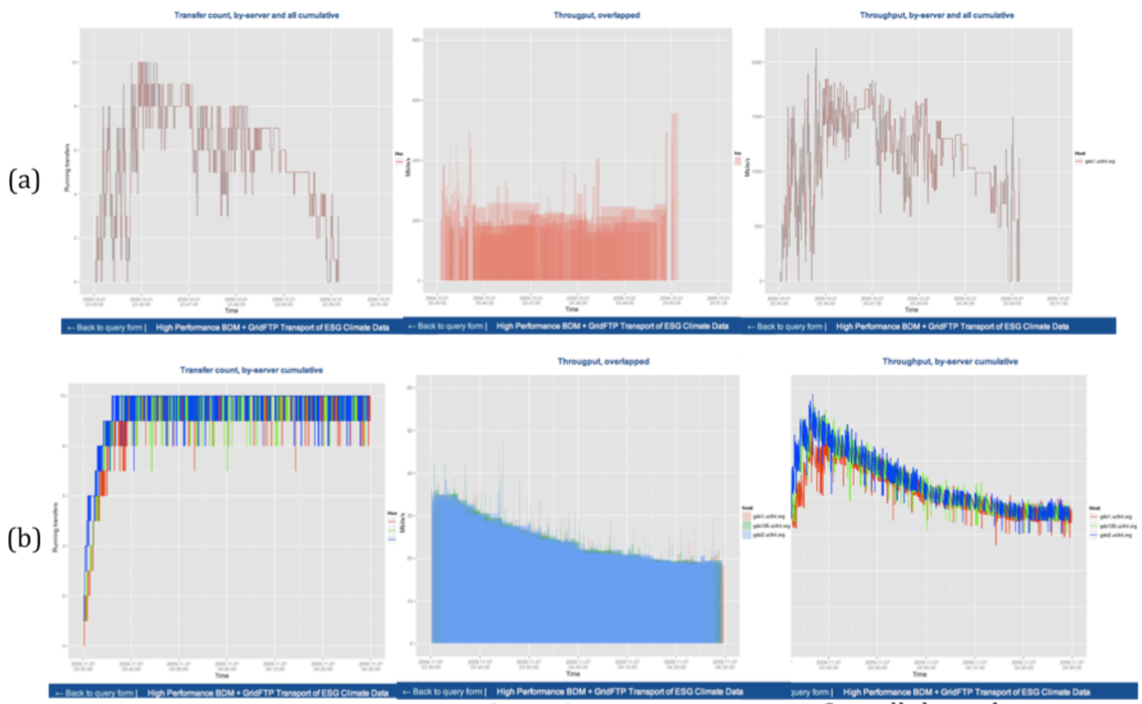
- The newly added reliability feature in globus-url-copy that stores the untransferred urls on the local file system for later restarting was instrumental in getting the data reliably moved to the showroom floor. The bandwidth challenge (BWC) tests helped with identifying and fixing a number of bugs in this new feature.
- The BWC tests prompted the development of a new load-balancing capability called client-side host aliasing, which enables the client to use multiple different hosts for concurrent transfers rather than multiple connections to the same host, without relying on the DNS round-robin functionality.
- The BWC tests and the run helped identify some bugs in the GridFTP server. These bugs will be fixed in the upcoming Globus Toolket Series 5 (GT5) release.

Globus.org (Contributed by Kettimuthu at ANL)

- The data used for the challenge was originally available only at LLNL and NERSC. We mirrored the source data at ALCF in order to use ALCF as one of the sources for the challenge. The mirroring was done using globus-url-copy and Globus.org, a hosted data-movement service. This effort helped with identifying critical bugs in globus.org and improving it a great deal. Performance issues, problems with multiuser support, and a host of other issues in globus.org were identified and fixed.

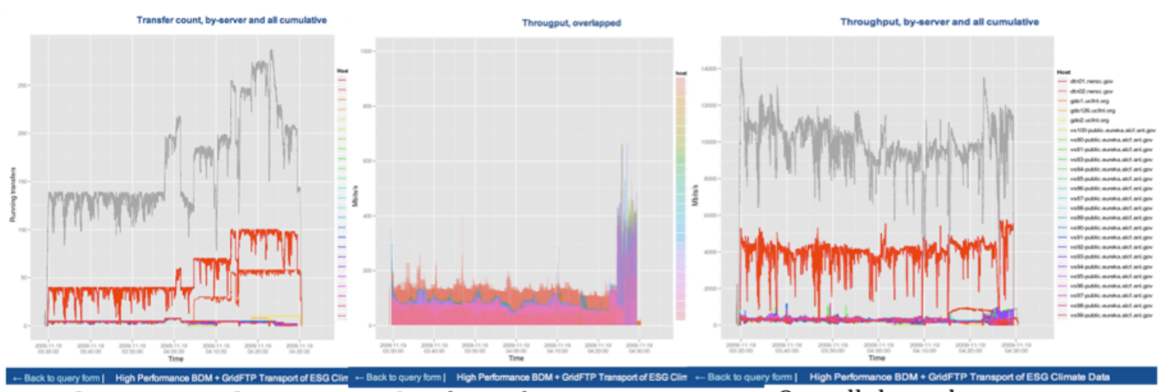
Bulk Data Mover (BDM)

- Transfer-queue and concurrency management algorithms contributed to increased transfer throughput (including both network and storage). NetLogger analyses (http://acs.lbl.gov/NetLoggerWiki/index.php/SC09_bwc_summary) provided valuable information on time-varying patterns in the overlap between multiple concurrent transfers for tuning the BDM transfer-queue and concurrency management algorithms. Figure 2 shows how concurrency algorithms change the transfer throughput. Figure 3 shows the concurrency counts and transfer throughput over time during the SC09 activity.
- Network delay should be considered in the calculation of concurrency and parallelism in BDM data transfers.
- A small discrepancy was found between the transfer rate measurement in BDM and the network throughput rate that the SC09 network showed. The calculated transfer rate in BDM corresponds to the values from NetLogger plots. The rate is determined from the transfer rate measured in BDM on a per-file basis. Per-file transfer rates are averaged to show a slightly higher or lower transfer rate than the actual network throughput flowing through the circuits. The SC09 effort showed us this discrepancy transfer rate measurement that we would not have seen otherwise. This will be adjusted in the future release of BDM.



Concurrency Count | Overlapped concurrent transfers showing each throughput | Overall throughput

Figure 2. ***Transfer-queue management and concurrency algorithms affect the throughput over time: (a) with earlier algorithms (b) after optimized data transfers from LLNL to NERSC. (Image generated by NetLogger.)***



Concurrency Count | Overlapped concurrent transfers showing each throughput | Overall throughput

Figure 3. ***SC09 concurrency counts and transfer throughput over time. (Image generated by NetLogger.)***

Summary

The SC09 Bandwidth Challenge result was successful. The SC09 effort achieved about 15 Gbps on average and moved about 7 TB of data in 1 hour from three data sources to the SC09 showroom floor. During this effort, a few issues were identified as described in the above sections. The effort demonstrated our current status in ESG data replication over the network and what needs to be done to make it better for the upcoming CMIP-5, IPCC AR5 release. The Climate 100 project and other application sciences with similar needs should benefit from these ESG data replication activities.