



Towards the Development of a Pipelined Data Movement with Globus Online

Mary Thomas¹, Christopher Paolini¹, Rajkumar Kettimuthu^{2,3}

¹San Diego State University, ²Argonne National Laboratory, ³University of Chicago



Abstract

Big data has become an inevitable part of science:

- Data generated from experiments and simulations must be transferred to other locations for analysis, visualization and archival purposes.
- The time required to transfer this data is a function of approach, transfer mechanisms used, file size and the bandwidth available on the network.
- Data needs to be shared with collaborators and communities.
- Software tools needed to simplify the task of transferring and managing data.

In this research, we investigate a new mechanism to improve Globus Online transfers by pipelining data transfer through an intermediate host. We also describe the applicability of this solution in the context of a real world application – a coastal ocean model simulation run on XSEDE resources.

Motivation

Users of XSEDE resources encounter significant performance reductions while transferring datasets and often choose an alternative path with a lower packet loss rate that traverses different Internet2 connector sites. Based on the users environment, alternative paths are defined by specifying a mix of intermediary XSEDE systems based on characteristics including:

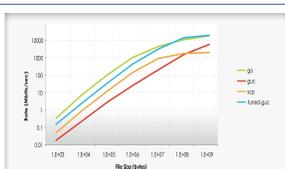
- Type of network (TCP/IP, DMZ, Lambda network, XSEDE network): optimal choice is a high-speed, low-latency dedicated connections to a particular XSEDE resource. Reality: TCP/IP.
- Availability of tools: FTP, SCP, GridFTP, GlobusConnect.
- Geographical distance between user and nearest XSEDE resource.
- Data file sizes and archival resources and time Firewalls force users to stage data between different hosts.

Typical approach: transfer from source to intermediate node, followed by a transfer from intermediate node to destination. Issues: multiple transfers are serial additional disk write and disk read involved.

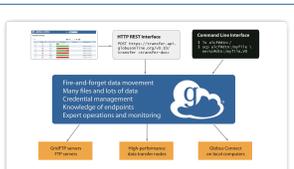
Globus Online

Globus Online makes the task of data transfer easier for users by allowing them to outsource the task of data transfer to the hosted Globus Online service. Features include:

- Uses GridFTP to get high performance, monitors the transfers and restarts the transfers automatically in case of any transient failures.
- Integrated into all XSEDE resources, includes authentication.
- Simple client & server tools, including a Web interface.
- Transfers limited by bottleneck of network bandwidth between the source and the destination.
- Where standard network path between the source and destination is slow, it is possible to find a better alternative path using an intermediate host.

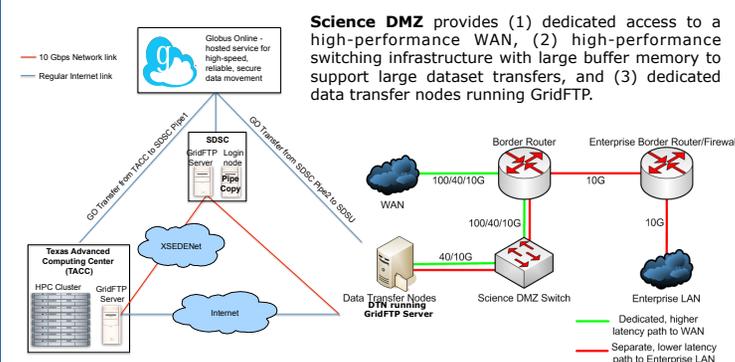


Comparison of Globus Online performance with other transfer mechanism. Globus Online's autotuning performs better than or similar to the expert tuned traditional GridFTP client.



~20 Petabytes of data moved, More than 10,000 Registered users, 99.9% availability

Our Approach



Science DMZ provides (1) dedicated access to a high-performance WAN, (2) high-performance switching infrastructure with large buffer memory to support large dataset transfers, and (3) dedicated data transfer nodes running GridFTP.

A new approach: use the concept of Unix pipes to support parallel transfers and avoid file copy//IO on remote host → improve performance for transfers through an intermediate node.

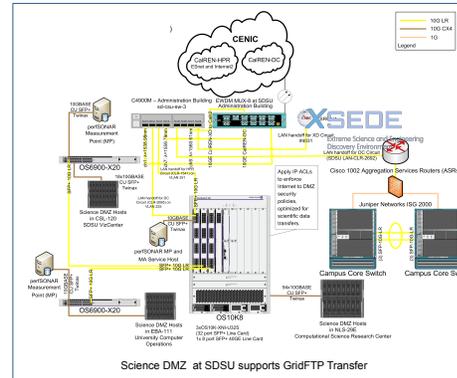
SDSU Science DMZ

Objectives:

- Facilitate high-performance data transfer for scientific applications using Globus Online GridFTP
- Provide uncumbered, high-speed access to online scientific applications and data generated at SDSU
- External access to science resources not impacted by regular "enterprise" or business class Internet traffic
- Focus on "BigData" Intensive Science

Capabilities:

- Alcatel-Lucent OmniSwitches
- Dedicated 10GE (maybe 40GE) uplink to Internet2 and ESnet via CENIC
- Optimized network for high-volume bulk transfer of scientific datasets
- Network performance measurement based on the PerfSONAR framework
- InCommon Federation global federated system for identity management and authentication to DMZ connected hosts and services

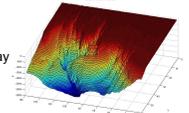


Source	Destination	Xfer Mech	Mbytes	Mbits/sec
xsedef#trestles	edwards.sdsu.edu	SCP	100 MB	9.5
xsedef#trestles	edwards.sdsu.edu	SCP	1 GB	32.9
xsedef#trestles	edwards.sdsu.edu	SCP	10 MB	44.6
xsedef#trestles	mthomas@sdsuppe3	GlobusOnline	1 MB	2.667
xsedef#trestles	mthomas@sdsuppe3	GlobusOnline	10 MB	20.000
xsedef#trestles	mthomas@sdsuppe3	GlobusOnline	100 MB	50.000
xsedef#trestles	mthomas@sdsuppe3	GlobusOnline	10GB	84.211
xsedef#trestles	xsedef#onestar4	GlobusOnline	1MB	2.667
xsedef#trestles	xsedef#onestar4	GlobusOnline	10 MB	20.000
xsedef#trestles	xsedef#onestar4	GlobusOnline	100 MB	80.000
xsedef#trestles	xsedef#onestar4	GlobusOnline	10 GB	579.650
xsedef#trestles	blackbox	PipeCopy		

Application Scenario

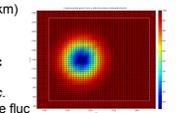
Nesting ROMS and UCOAM: A Case Study in Monterey Bay

Nesting one grid within another is an efficient way to increase the resolution locally. The embedded (child) grid gets its boundary condition data from the larger (parent) grid.



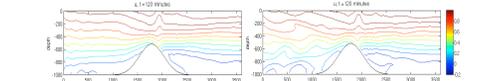
Regional Ocean Modeling System (ROMS):

- hydrostatic free-surface ocean model ideally suited to simulate medium to large-scale (10–1000 km) coastal ocean processes
- Well established ocean model.



Universal Curvilinear Ocean and Atmospheric Model (UCOAM):

- Full 3D curvilinear coordinate, non-hydrostatic.
- Large Eddie Simulation → resolve meter-scale flux
- Requires large arrays (10¹⁰) elements
- Curvilinear system requires a large number of arrays (10²)
- Communication occurs along all 3 axes
- Requires parallelization, peta-scale compute resources with large memory, and high-speed networks to manage the results.



Velocity profiles for flow over submerged seamount for fine-grained parent mesh (left) and nested child mesh (right). Profiles are qualitatively similar, and nested run was more efficient (convergence time 27000s/3200s)

Results

- **Experiments** were conducted to test the parallel pipe concept using large-scale transfers of simulated data files.
- XSEDE resources: trestles.sdsu.edu, lonestar.tacc.utexas.edu.
- SDSU DMZ: blackbox.sdsu.edu, pipeline3.acel.sdsu.edu.

Using the Service:

- The daemon is called 'pipecopyd' and it creates two UNIX FIFOs (named pipes): the source endpoint & destination endpoint. The daemon is invoked with the unix command: `pipecopyd srcfile destfile`
- Transfers are scheduled using cli-globusonline.org syntax
- **Pipecopyd** reads data from one pipe (argument #1) and writes it to another pipe (argument #2)

```
ssh trestles.sdsu.edu;
pipecopy /tmp/src_pipe /tmp/dst_pipe &; exit;
ssh cli-globusonline.org; transfer --perf-p 1 --
xsedef#lonestar4/tmp/src_file xsedef#trestles/tmp/src_pipe;
transfer --perf-p 1 -- xsedef#trestles/tmp/src_pipe \
sdsu#hpc-cluster/tmp/dst_file; exit;
```

- Results are shown in the table on the left

Future Work

- Continue to develop prototype, test basic concepts and viability of solution
- Interface to XSEDE resources
- Evaluate parallel pipecopy
- Integrate into Ocean model application
- Contribute software to appropriate open source repo



This work supported in part by the National Science Foundation (#072383, #072383, #136513), the US Dept. of Energy (DE-FC02-02ER25518), XSEDE (NSF-03-10827), the CSRC and College of Science at SDSU.



For further information, see <http://acel.sdsu.edu>
Contact: mthomas@mail.sdsu.edu

