Data Movement in HPC using GridFTP and Globus

Fundamentals of Data Management

September 2014

Albert Heyrovsky
Applications Developer, EPCC
a.heyrovsky@epcc.ed.ac.uk

Course outline

- Why talk about GridFTP?
- What is GridFTP?
- Main features of GridFTP
- GridFTP clients
- Globus File Transfer
- Who uses GridFTP?
- Summary

- After completing this lesson, you should know:
  - What GridFTP is
  - What it can be used for
  - How it can be used - clients

Why talk about GridFTP?

- It is widely used by many HPC centres (including EPCC)
- It is a standard protocol for data transfers
- It is reliable, high performance
- It can transfer very large files over Wide Area Networks (in the gigabytes to terabytes range)
- It is free, free clients available

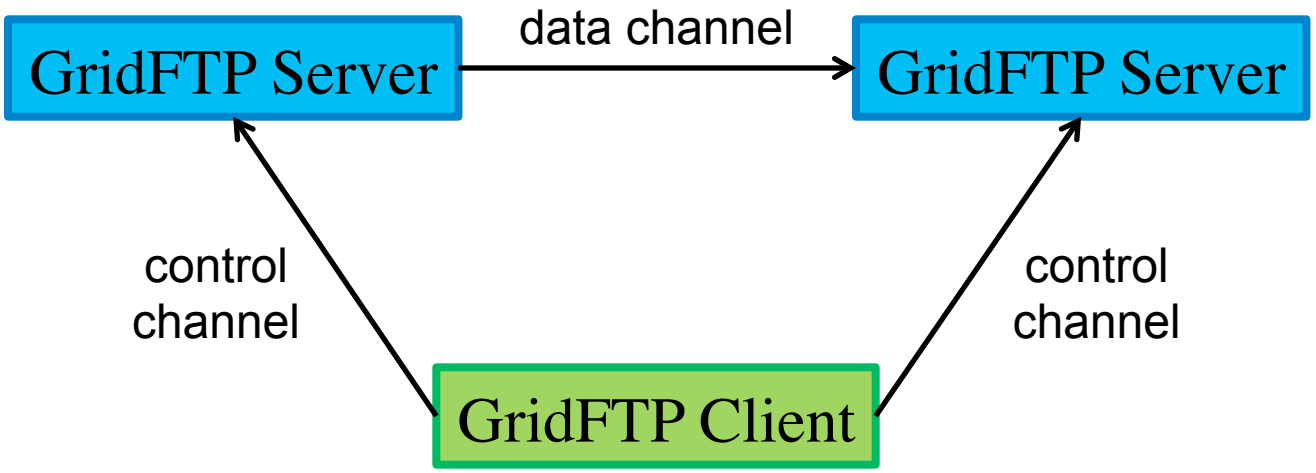
What is GridFTP?

As per Wikipedia (http://en.wikipedia.org/wiki/GridFTP):

- GridFTP is an extension of the standard File Transfer Protocol (FTP)
- It is used for high-speed, reliable and secure data transfer
- It has been defined within the GridFTP working group of the Open Grid Forum

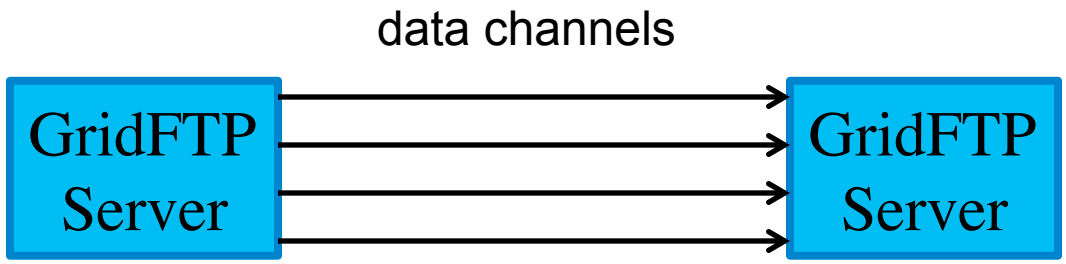
Main features of GridFTP

- Security with GSI
  - GSI - Grid Security Infrastructure provides authentication and encryption to file transfers using X.509 digital certificates
- Third party transfers
  - GridFTP allows remote transfers between servers initiated by a local client



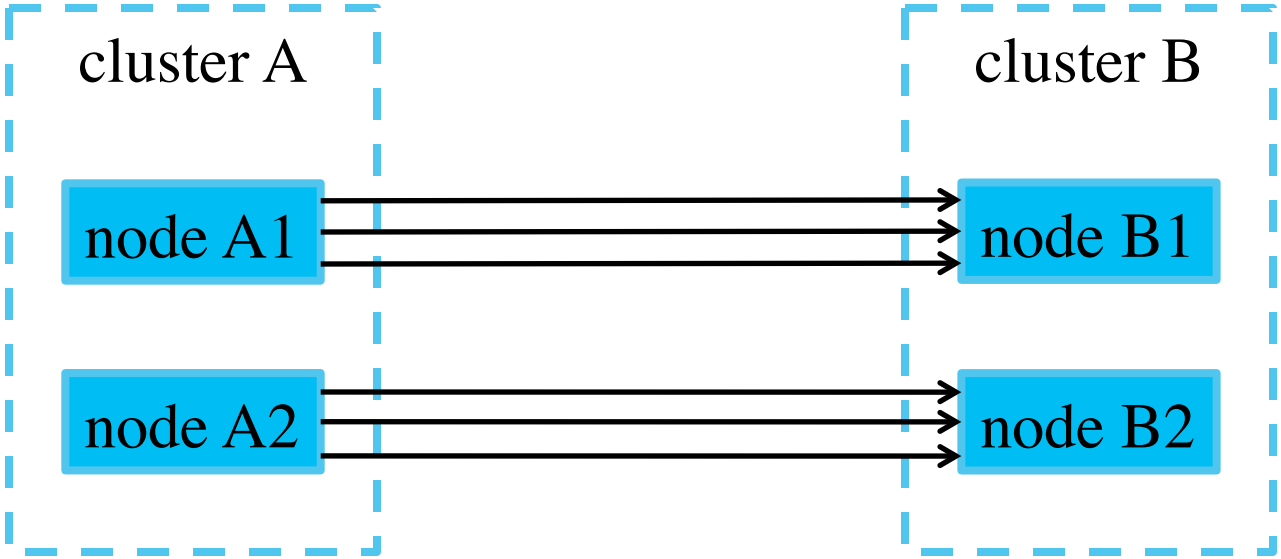
Main features of GridFTP

- Parallel TCP data streams
  - GridFTP achieves much greater use of bandwidth by allowing multiple simultaneous TCP streams



Main features of GridFTP

- Striped data transfers
  - Used for transfers of files between clusters with parallel shared file systems
  - Each node in the cluster reads a section of the file and sends it over the network
  - Striping and parallelism may be used together



Main features of GridFTP

- Partial file transfers
  - Allows transfers of sections of a file by specifying an offset and the length of the block desired
  - This feature is useful when only a small section of a very large file is required for processing
- Fault tolerance and resuming transfers
  - GridFTP handles file transfer failures due to network unavailability and server problems
  - GridFTP servers can resume transfers based on where they left off before a failure

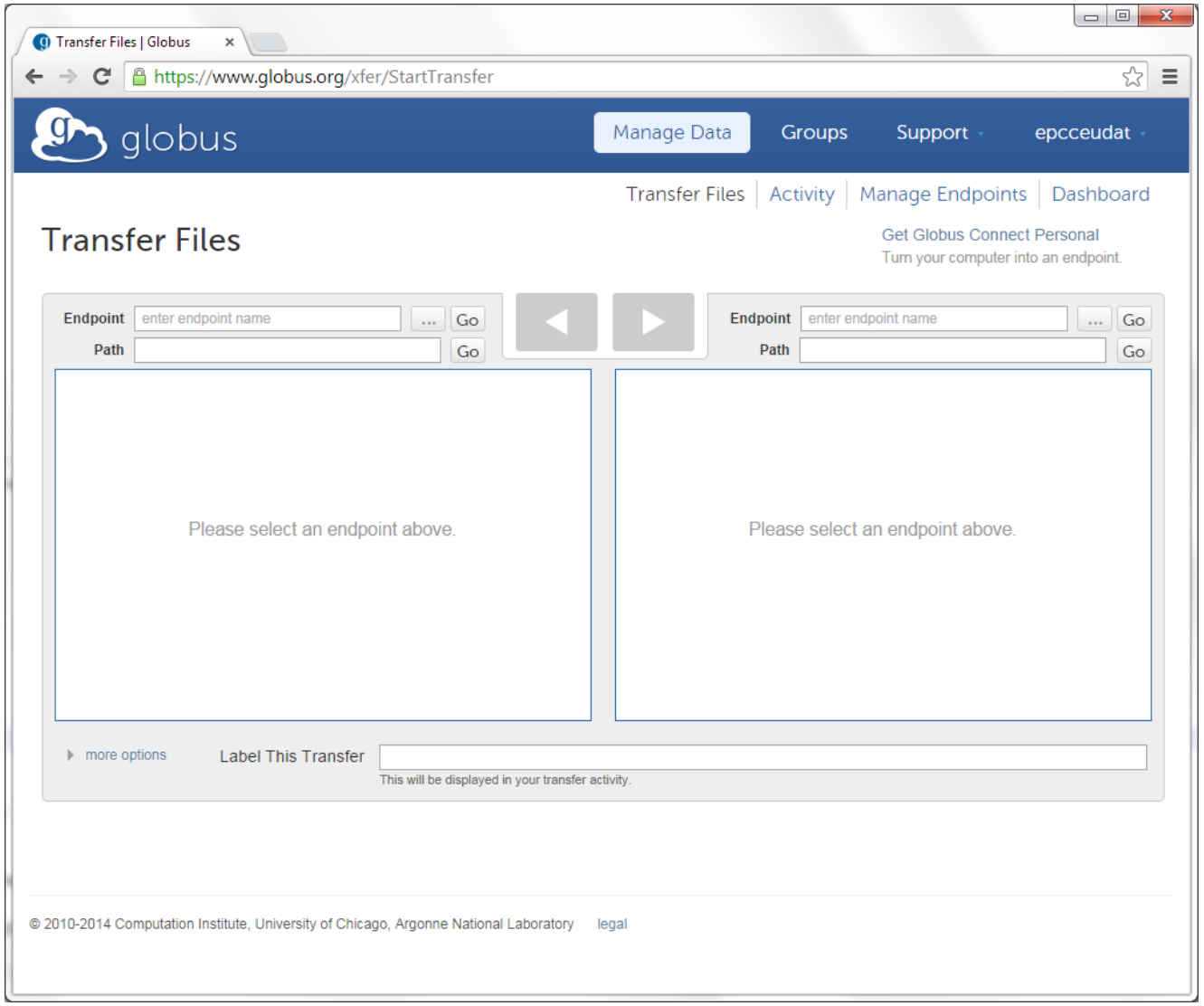
GridFTP Clients

- Command line clients
  - The Globus Alliance provides a Globus Toolkit which contains a command line client *globus-url-copy*
  - UberFTP – an interactive GridFTP client
- Web browser client
  - Globus File Transfer – a GridFTP client in the cloud

Globus File Transfer

- Globus File Transfer is one of the Globus services
- Globus is a project which has its beginnings in the development of Grid Computing
- Globus services can be accessed at https://www.globus.org/
- The technology underlying this service is GridFTP
- It is software as a service (SaaS), enabled by the cloud (hosted by Amazon AWS)

Globus File Transfer



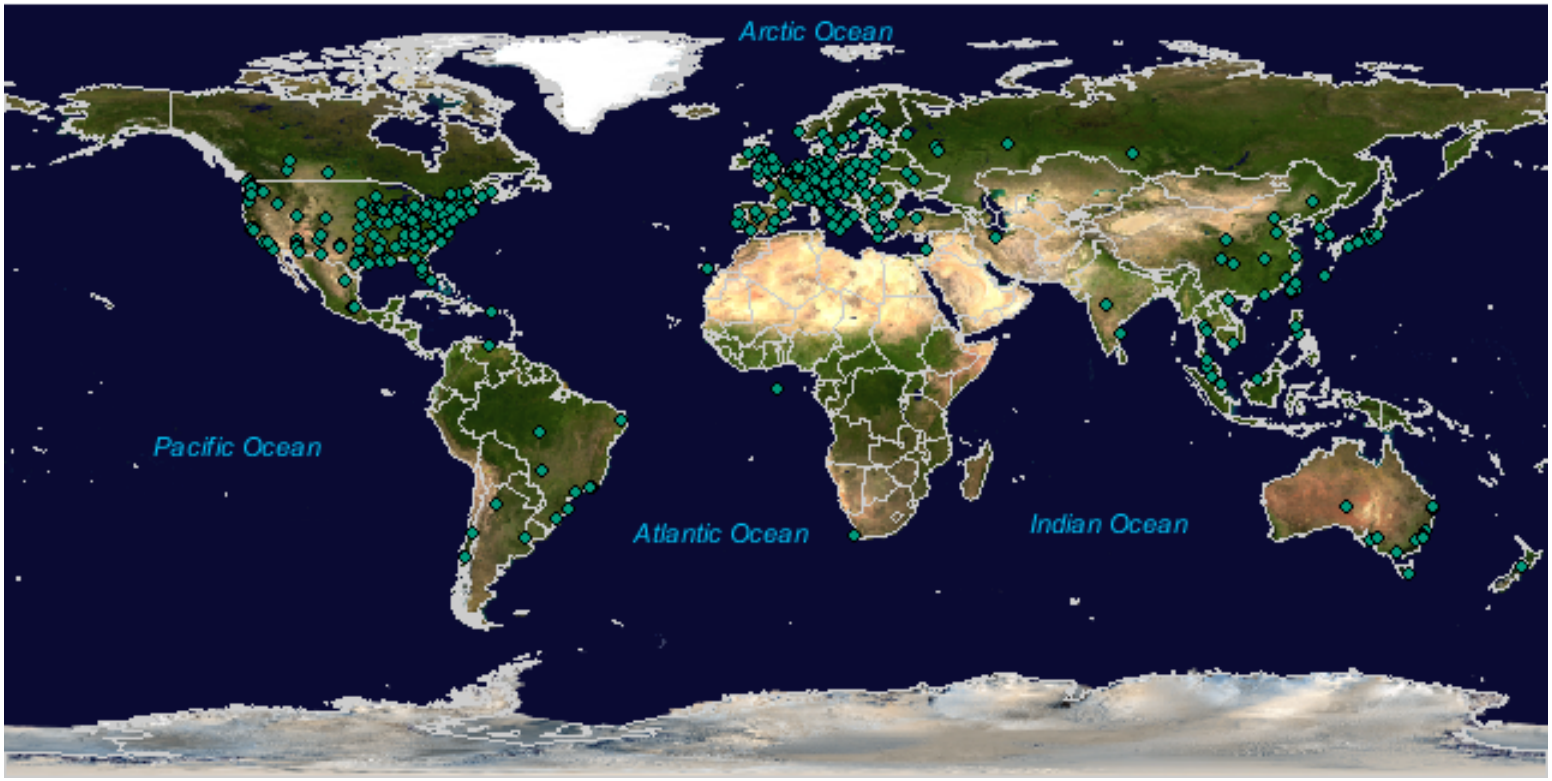
Globus File Transfer

- Data is transferred between Globus “endpoints”
- An endpoint is a logical address for a GridFTP server
- One can turn a laptop or a personal computer into a Globus endpoint using Globus Connect Personal
  - In this way one can transfer files to and from a local computer (desktop computer, laptop)
- It provides an easy to use “fire and forget” data transfer mechanism which can be accessed from a web browser
- It can also be accessed via a command line interface
- There are also APIs which allow file transfers to be integrated into custom applications

Who uses GridFTP?

- Many HPC sites / supercomputing centres around the world, e.g.
  - EPCC
  - Barcelona Supercomputing Centre (BSC)
  - Consortium of Italian Universities CINECA
  - National Center for Supercomputing Applications (NCSA)
  - San Diego Supercomputer Center (SDSC)
- Many scientific facilities, e.g.
  - Argonne National Laboratory (ANL) – the developers of GridFTP
  - Large Hadron Collider (LHC) Computing Grid
  - Laser Interferometer Gravitational-Wave Observatory (LIGO)
  - Southern California Earthquake Center (SCEC)
  - European Space Agency (ESA)
- Other organizations, e.g. BBC

GridFTP Servers Around the World



Created by Lydia Prieto, G. Zarrate, Anda Imanitchi (Florida State University) using MaxMind's GeoIP technology (http://www.maxmind.com/app/ip-locate).

Summary

- GridFTP is a file transfer protocol originating in Grid Computing, it is an extension of FTP
- Its main features are
  - Security, reliability
  - Third party transfers
  - Parallel file transfers
  - Striped transfers
  - Partial file transfers
  - Fault tolerance and recovery
- The main GridFTP clients are
  - Command-line globus-url-copy and UberFTP
  - Web based Globus File Transfer service

Acknowledgements

- Parts of this presentation have been taken from Wikipedia
- Thanks to staff from the Argonne National Laboratory for providing some of the lecture materials