Globus Online Update

Steve Tuecke

April 12, 2011


globus online
Reliable File Transfer. No IT Required.
Reliable, high-performance, secure file transfer
Move files fast. No IT required.
HOW IT WORKS
Learn more about the service
GET STARTED
Setup profile in 3 easy steps
Globus Online makes robust file transfer capabilities, traditionally available only on expensive, special-purpose software systems, accessible to everyone.
Learn more
Why Use Globus Online?
See how easy file transfer can
For HPC Resource Owners
Enable Globus Online
For Developers
Integrate with Globus Online

For Users of HPC Resources

- **Tutorials (in person & via webinar)**
  - Response to NERSC webinar on Globus Online: “OMG, that was fantastic! I have already started using the Globus Connect to transfer data, and it only took me 5 minutes to setup. I have been struggling setting up the grid server here on my machine for a week (certificates are a pain)! Thank you thank you”
- **User case studies**
  - MILC Collaboration: Lattice QCD
  - University of Colorado: Oceanography
  - STAR Experiment: Frontier Physics
- **Globus Connect**
- **Videos**

For HPC Resource Providers

- **Case studies**
- **How to enable your resource with GO**
- **GT 5.2**
- **Co-sponsored webinars to users**

Case Study: **NERSC**
Enabling Globus Online for Researchers

Seeking to provide its scientists with more effective web-based research tools, the National Energy Research Scientific Computing Center (NERSC) adopts Globus Online as a recommended method for fast, secure, reliable file transfer.

Top Features Cited

- **Fire-and-forget usage** – no complicated tools for users to learn
- **Automated performance tuning** – no need to constantly tweak the system
- **Secure enablement** – no need to implement new security processes or systems
- **Automated authentication for existing users**, with pre-create account options coming soon
- **GUI or CLI-based options** -- to meet needs and preferences of different users
- **Hosted service** -- no software installs or custom IT infrastructure required
- **Backed by expert support team** -- low-touch system for NERSC staff

"We need to provide web-based ways to accomplish computing tasks – it's what our scientists are coming to expect, and it's also what will ultimately make them more productive."

--David Skinner, Software Group Leader, NERSC

Challenge: Meeting the Demand for Web-Based Computing Services

With over 4000 users and hundreds of ongoing projects, NERSC (the National Energy Research Scientific Computing Center) at Lawrence Berkeley National Laboratory is a hotbed of computational science. The facility is known as one of the best-run scientific computing centers on the planet, stating that "what distinguishes NERSC is its success in creating an environment that makes HPC resources effective for scientific research."

"The Globus Online service is just what our scientists have been looking for."

--David Skinner, Software Group

But what was effective for one generation can be cumbersome for the next – and in the postmodern research arena, a "generation" of computational scientists can span months, not years.

For example, the method of accessing data by logging into a terminal and entering commands via a command line interface (CLI) is "in many respects going the way of the punch card where some of our scientists are concerned," says David Skinner, NERSC Software Group Leader.

NERSC's Hopper is the fifth most powerful supercomputer in existence.

"We need to provide web-based ways to accomplish computing tasks – it's what our scientists are coming to expect, and it's also what will ultimately make them more productive."

Therefore, like other top computing facilities, NERSC is working to make its data and computing services available online.

"Nowadays people go to the web for everything," observes Skinner, "and our users are looking to us to evaluate, recommend and provide access to reliable, secure services via the web. That's what led us to Globus Online."

Solution: Enable Globus Online for All NERSC Users

Globus Toolkit users at NERSC pointed Skinner to Globus Online, a fast and reliable transfer service which they had already discovered and used with success.

"placeholder for J Porter quote placeholder for J Porter quote"

--Jeff Porter, Physicist and NERSC user, Brookhaven National Lab

"It just makes sense to enable Globus Online for our users," says Skinner. "Our scientists are familiar with tools like scp and GridFTP, but with Globus Online we can offer a much simpler and faster method for moving data. Globus Online actually makes web-based data syncing an easy, nearly trivial process, so you don't have to be an IT or middleware expert to move your files."

For Developers

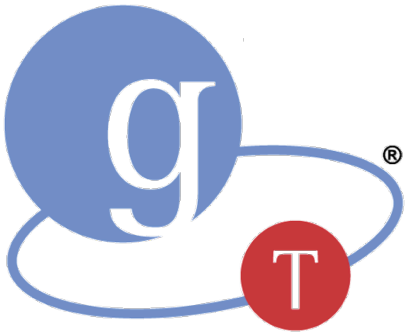
- **Application developers who want to outsource file transfer**
  - Condor
  - Science gateways
  - NERSC Web Toolkit
- **This afternoon we'll talk about additional value here…**

User Tutorial

Globus products

Globus Toolkit

Use the Grid



Reliable file transfer
Software-as-a-Service

globusonline.org

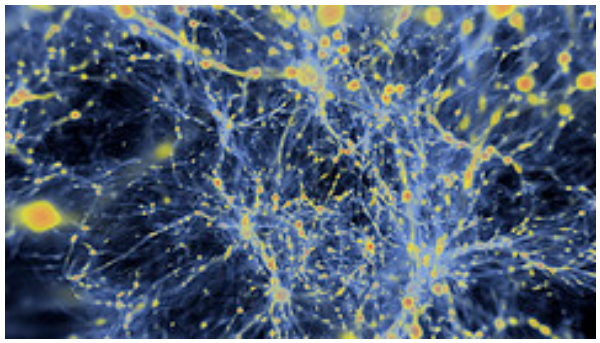
Globus Online

Build the Grid

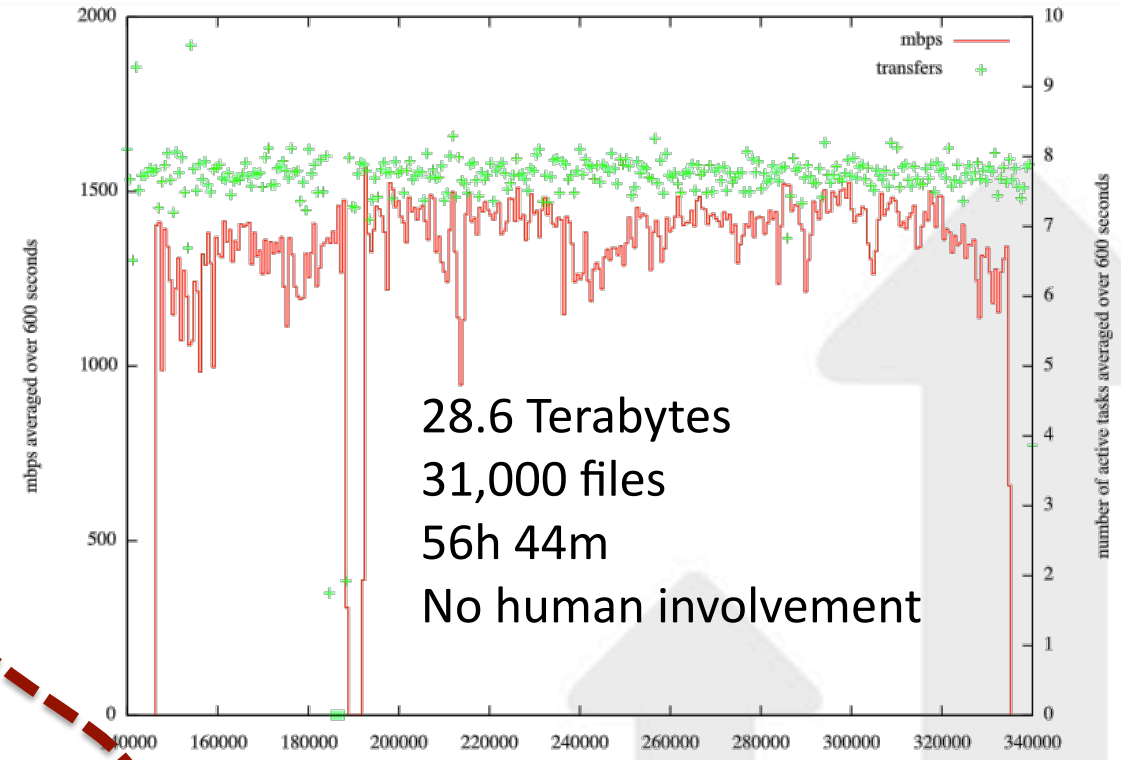
Components for building
custom grid solutions

globustoolkit.org

Globus Online In Action



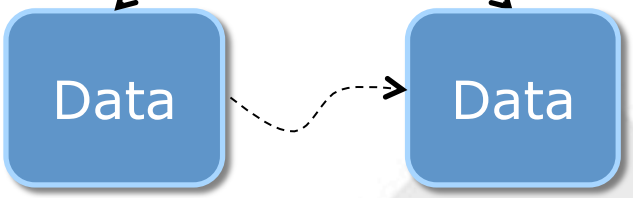
Astrophysics simulation data
generated in Tennessee,
moved to Illinois for visualization
(Enzo, UCSD; Futures Lab, Argonne)

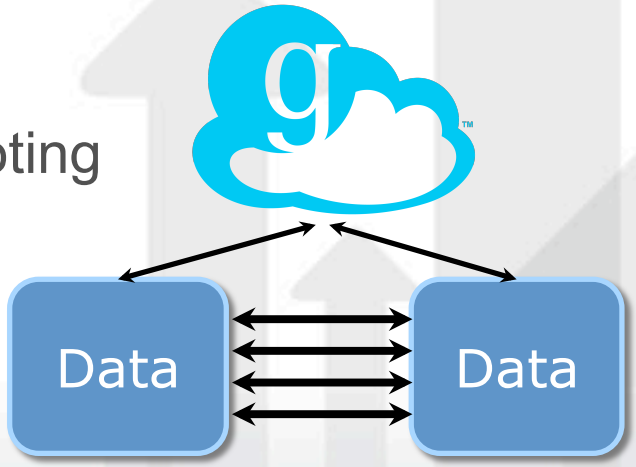


28.6 Terabytes
31,000 files
56h 44m
No human involvement

Benefits of Globus Online

- **Reliable file transfer.**
  - Easy “fire and forget” file transfers
  - Automatic fault recovery
  - High performance
  - Across multiple security domains
- **No IT required.**
  - No client software installation
  - New features automatically available
  - Consolidated support and troubleshooting
  - Works with existing GridFTP servers
  - Globus Connect solves “last mile problem”

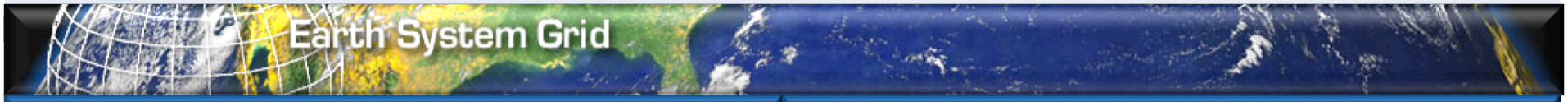




Who can benefit from Globus Online

- **Ad-hoc:** Non-programmers who need to move many files can use Web GUI
- **Scripted:** Users who want to create automated workflows can use Command Line Interface (CLI)
- **System builders:** Programmer who don't want to re-engineer file transfer solutions can use REST API

Demo



Globus Online and Earth System Grid

Rachana Ananthakrishnan & Neill Miller

Argonne National Laboratory/University of Chicago

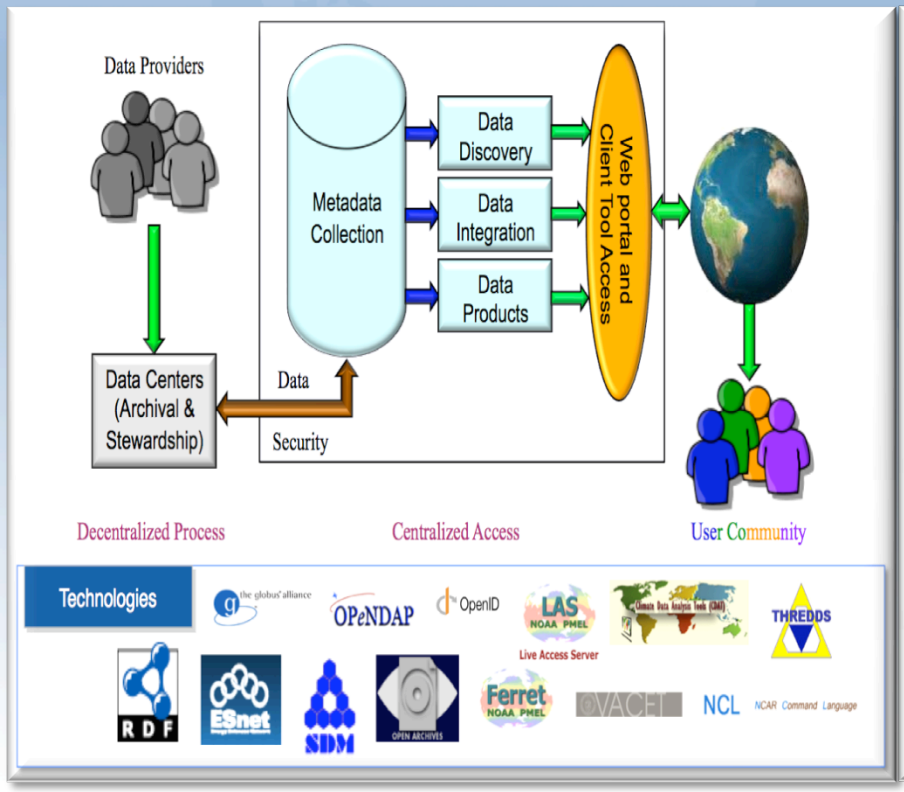
Acknowledgements: Nathan Hook, Eric Nienhouse and Nathan Wilhelmi, NCAR and Bryce Allen, Jack Kordas, and Karl Pickett, U Chicago

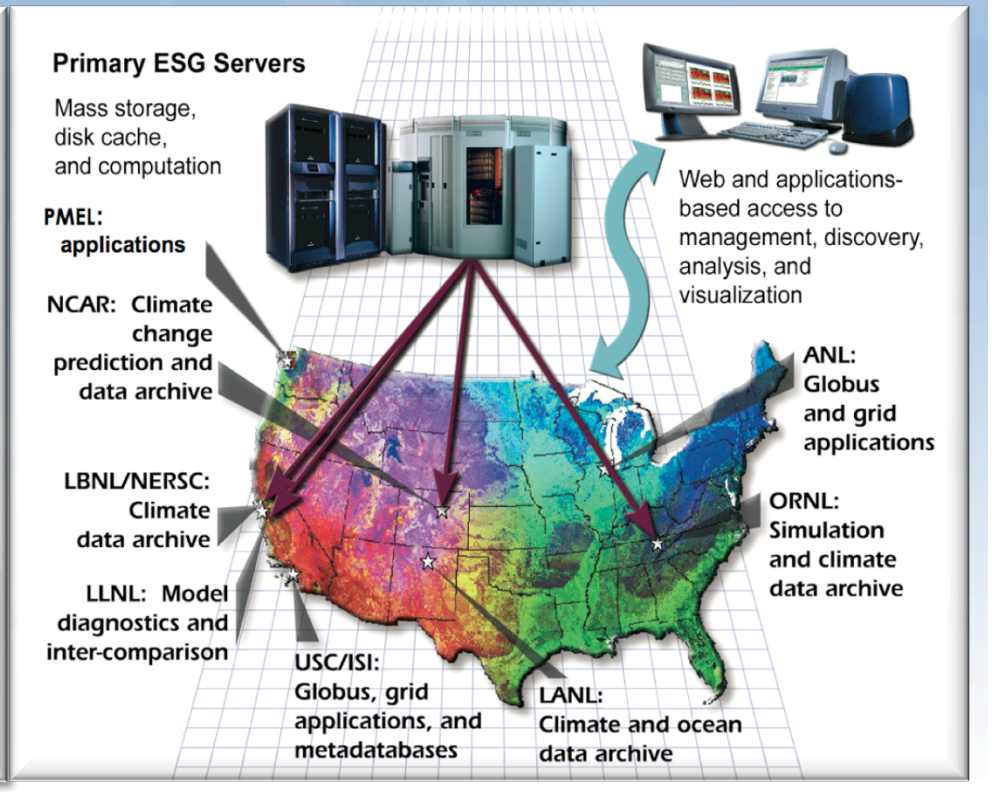
Earth System Grid Goals

Earth System Grid

To build, operate and support a global infrastructure for the management, access and analysis of climate data.

- Scope
  - Promote sharing of knowledge, software and tools among partners
  - Define APIs and protocols for interoperability among data centers
  - Collaborative development of software components
  - Deployment of common software infrastructure





ESG Federated Architecture

Earth System Grid

- Earth System Grid Federation
  - Sites across the world with ESG nodes
  - Single sign-on across the sites using OpenID
- Gateway Node
  - User registration and management
  - Data discovery and request
  - User credential management
  - Maybe customized for institutional focus
- Data Node
  - Data store and access
  - Publish metadata to Gateway
  - Download protocols: HTTP and GridFTP
- Clients
  - Browser based clients
  - Analysis tools as clients
  - Specialized download and replication clients

ESG Data Transfer Solutions

Earth System Grid

- Replication
  - Archive data transfer
  - Replication Tool with GridFTP for transfer
- Download to user machine
  - Search and select data at Gateway
  - HTTP download via wget script
  - GridFTP download via Data Mover Lite
- Transfer to external server
  - Not supported

Globus Online Solution

Earth System Grid

- Outsource data transfer to Globus Online
  - No ESG specific software
  - Public endpoints, web-based monitoring, performance tuning
- Leverage ESG infrastructure
  - Reuse ESG login mechanism and credential service
  - Existing data servers and security
- Integrate with ESG Gateway
  - User initiated transfers
  - Domain specific tools for search and select
  - Seamless and uniform experience for end-users
- Integrate with ESG Replication Tool
  - GO as a transfer option via the CLI interface

Demo

Earth System Grid

- Pre-requisites
  - ESG Account
  - Globus Online Account
- Scenarios
  - Transfer to external server
  - Download to local laptop

Demo: behind the scenes

Earth System Grid

- Globus Online Transfer API used (REST interface)
  - Java library written to support required operations
  - Library will be used for other ESG integration efforts
- ESG Portal Integration
  - Collect all required information for transfer
    - Source files and credential service information
    - Destination server and credential service information
    - GO username
  - Portal provides some of the information
    - Source files, source credential service information
    - Source login name

Demo: behind the scenes

- Custom web workflow
  - Gather destination information and passwords
  - Integrated in ESG Gateway as Spring Form Controller classes and presented with JSP page
- Auto detect the minimal information needed
  - Portal level caching of information
  - Leverage auto-activate in GO

Planned Work

Earth System Grid

- User experience improvements:
  - ESG OpenID to log into GO
  - GO account creation included in workflow
- Leverage Globus Online features:
  - Explore GO UI components
  - HTTP transfer support
- ESGF plans:
  - Integration with other climate analysis and search tools
  - Release as part of NCAR Gateway software

GridFTP Update

Raj Kettimuthu

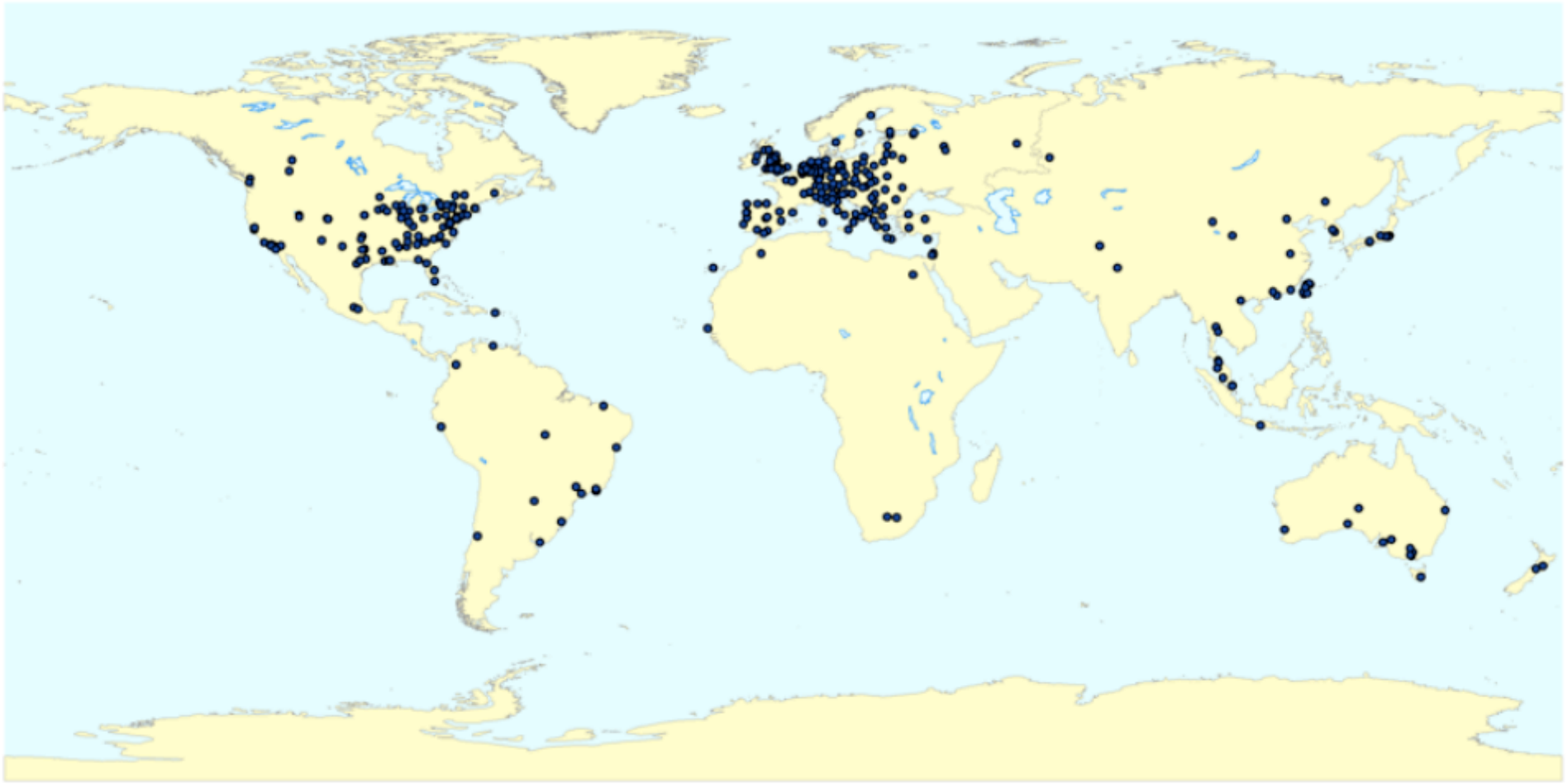
Argonne National Laboratory and

The University of Chicago

GridFTP

- **High-performance, secure data transfer protocol optimized for high-bandwidth wide-area networks**
- **Globus GridFTP**
  - Parallel TCP streams, optimal TCP buffer
  - Non TCP protocol such as UDT
  - Cluster-to-cluster data movement
  - SSH, Grid Security Infrastructure (GSI)
  - Restartable transfers

GridFTP Servers Around the World



Created by Tim Pinkawa (Northern Illinois University) using MaxMind's GeoIP technology (http://www.maxmind.com/app/ip-locate).

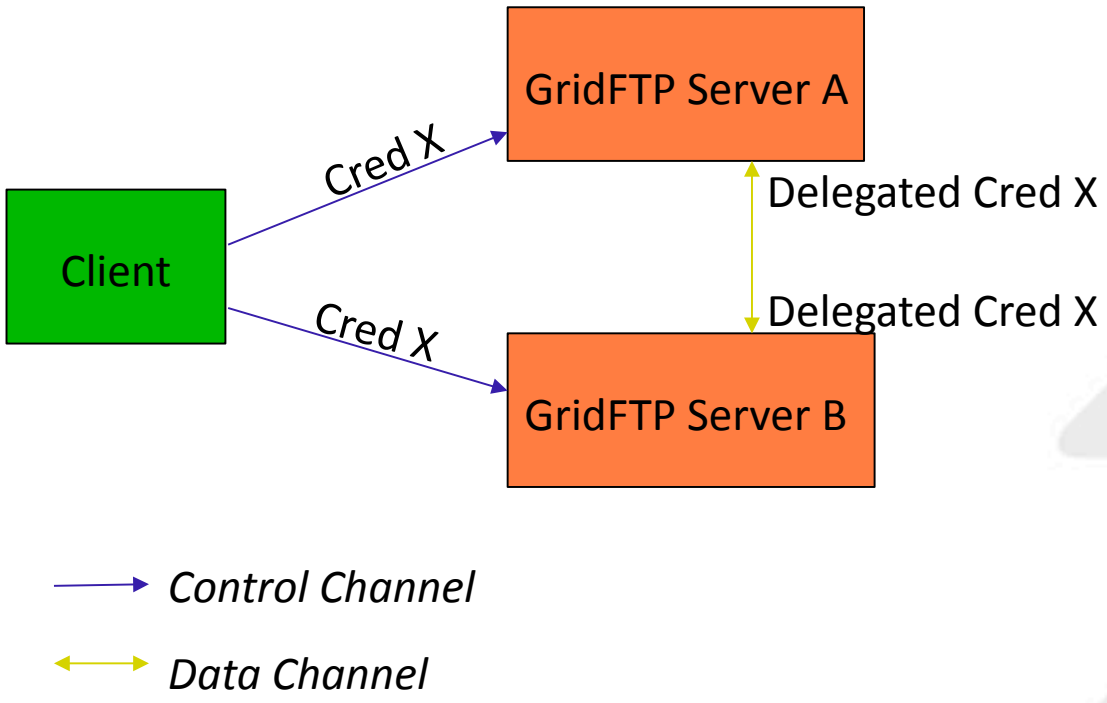
Globus GridFTP

- **Advantages**
  - Fast
  - Secure
  - Extensible
  - Standard
  - Robust
- **Globus Online + Globus Connect**
  - Easy-to-use
  - Reliable

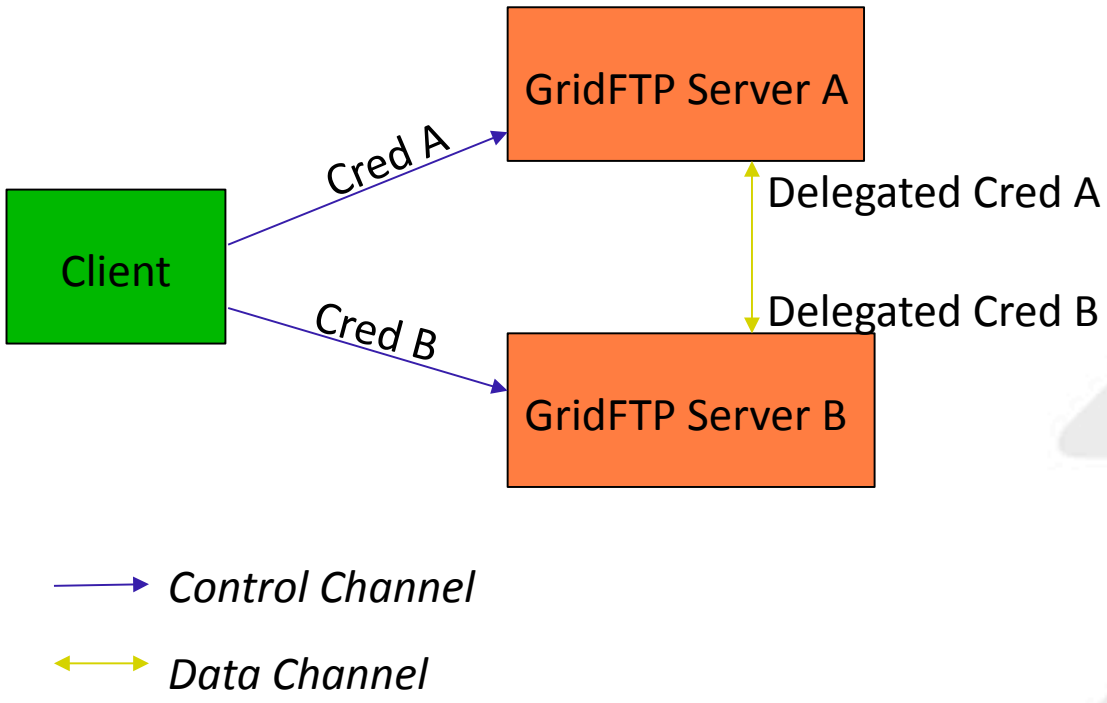
GridFTP update

- **New features added in the last year**
  - DCSC
  - Sync
  - Chroot
  - Rate limiting
- **Future plans**
  - Network reservation
  - GridFTP v3

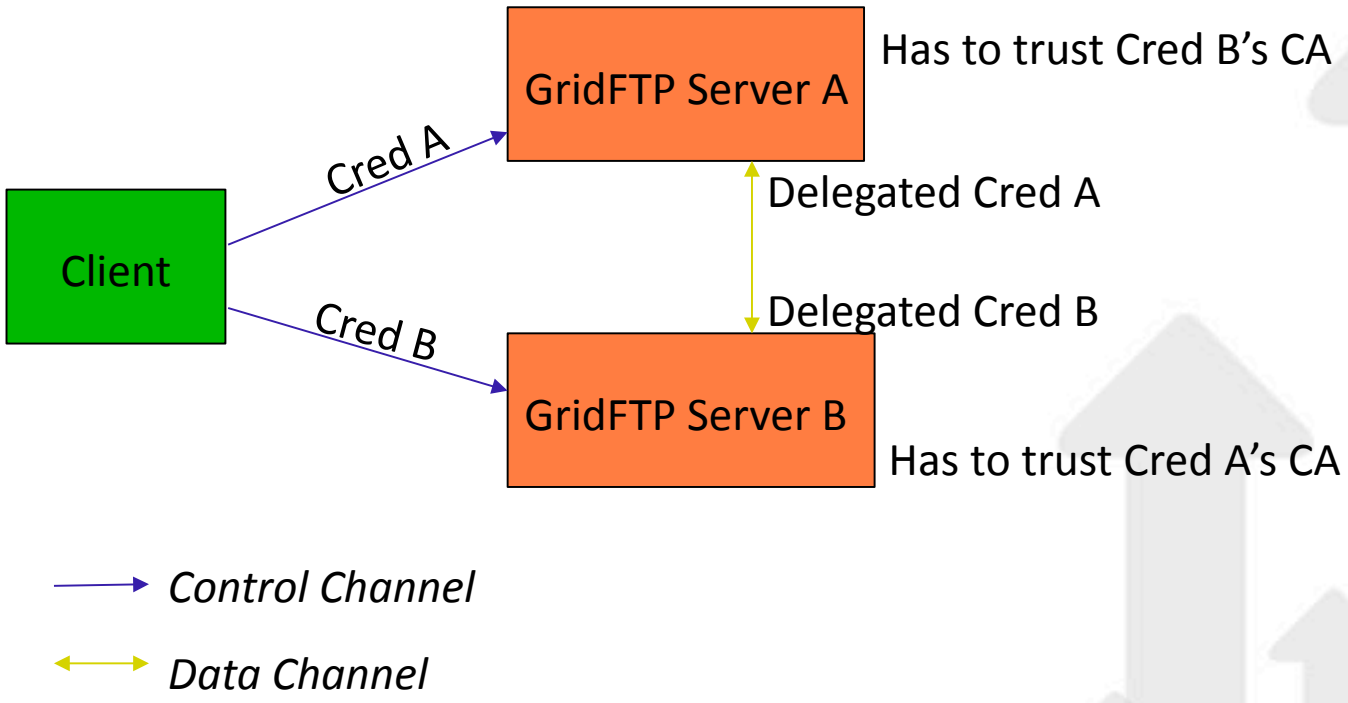
Data Channel Security Context (DCSC)



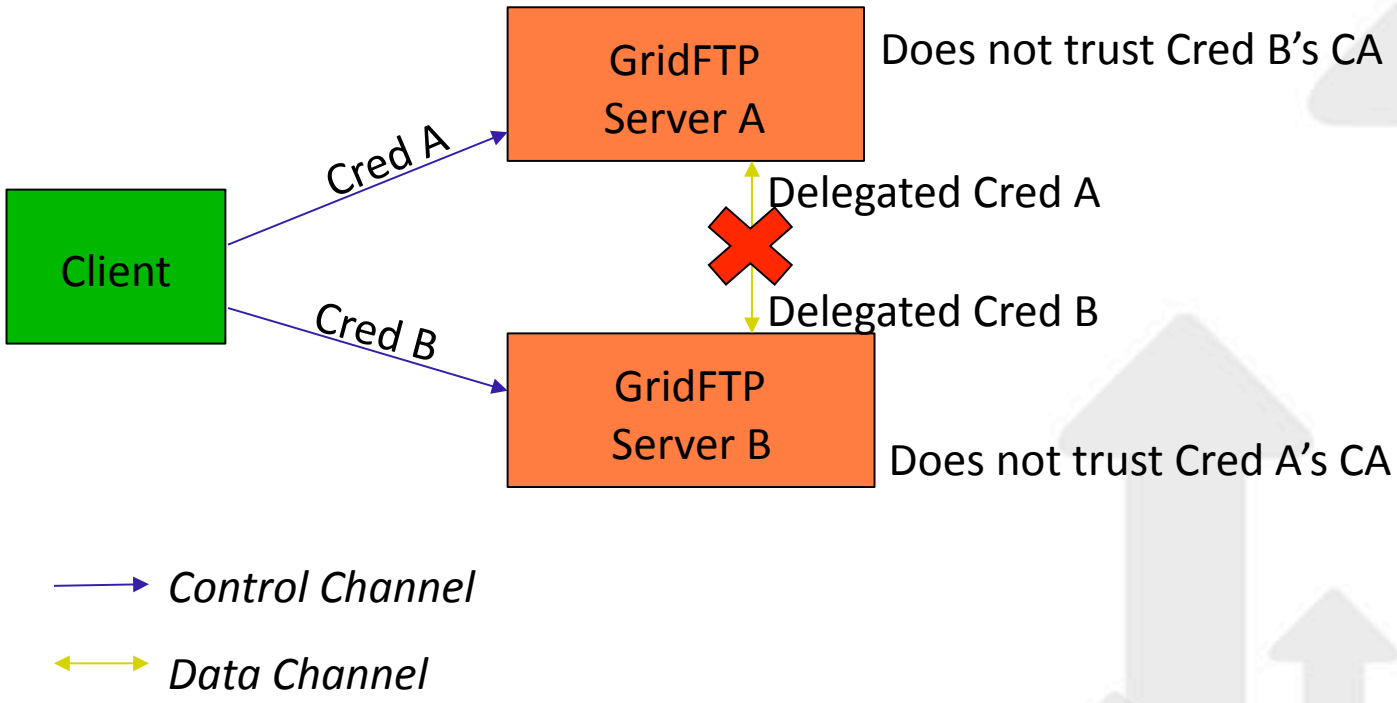
Data Channel Security Context (DCSC)



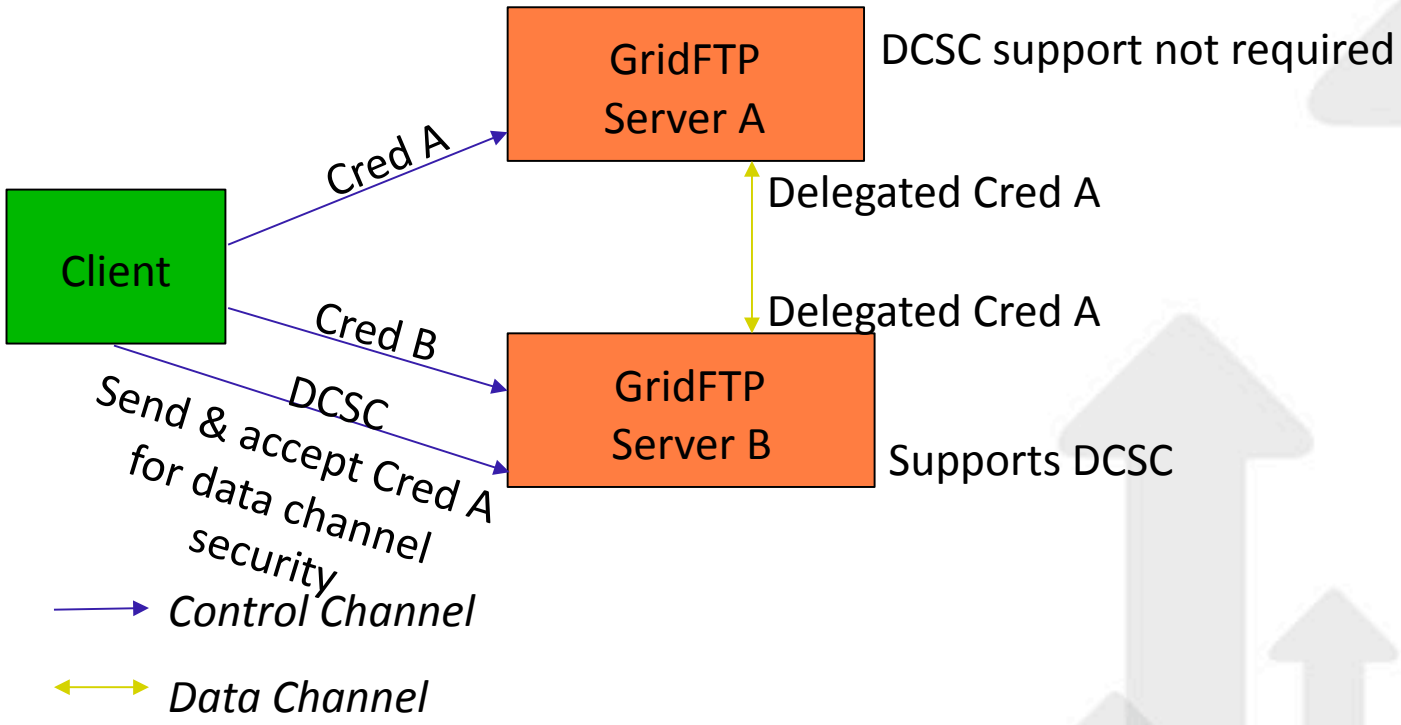
Data Channel Security Context (DCSC)



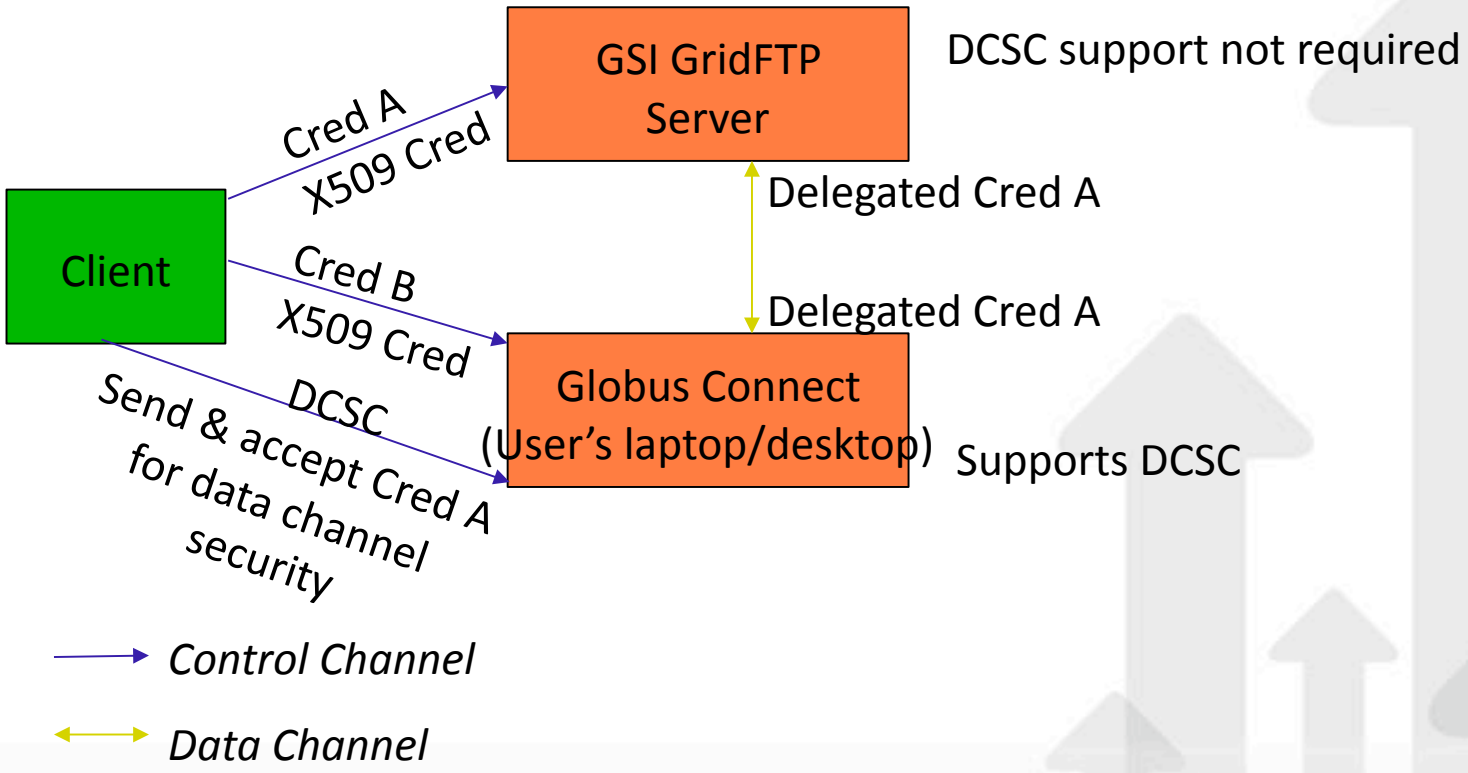
Data Channel Security Context (DCSC)



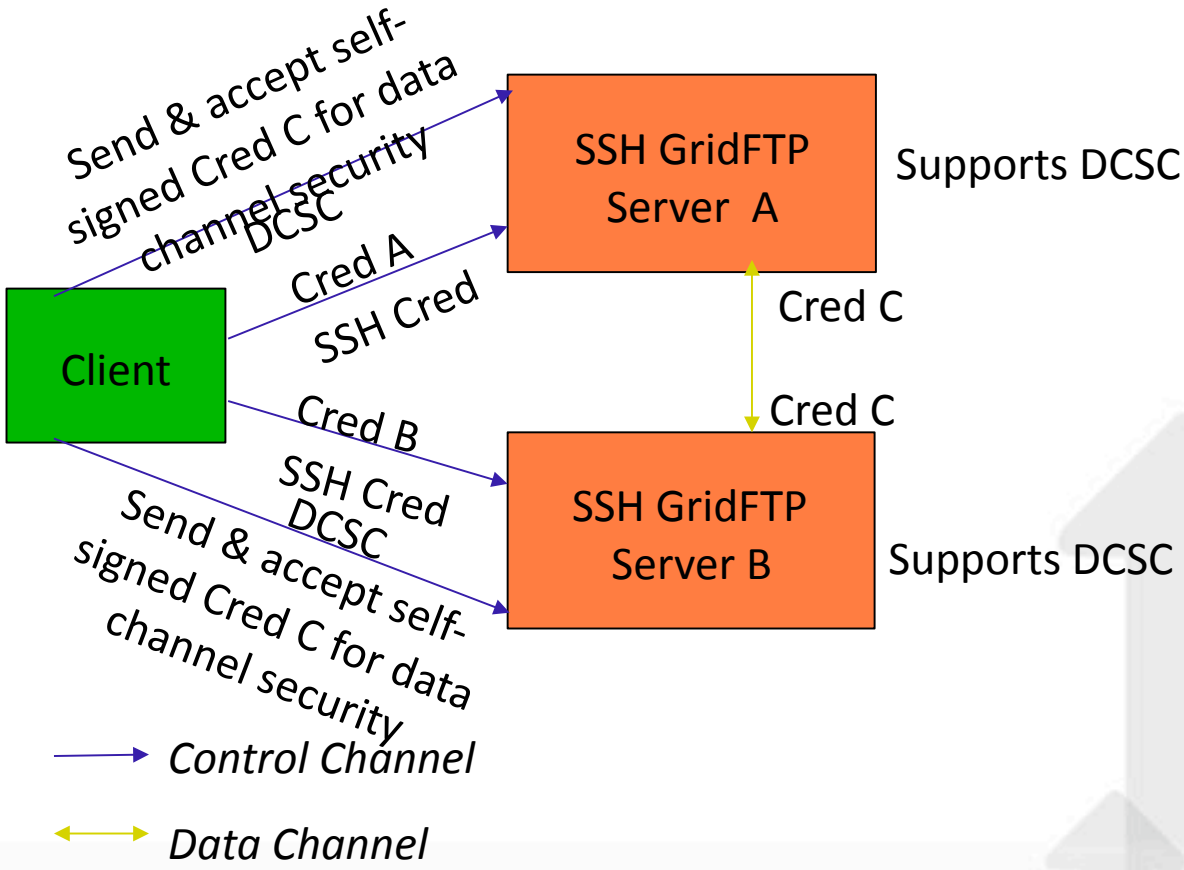
Data Channel Security Context (DCSC)



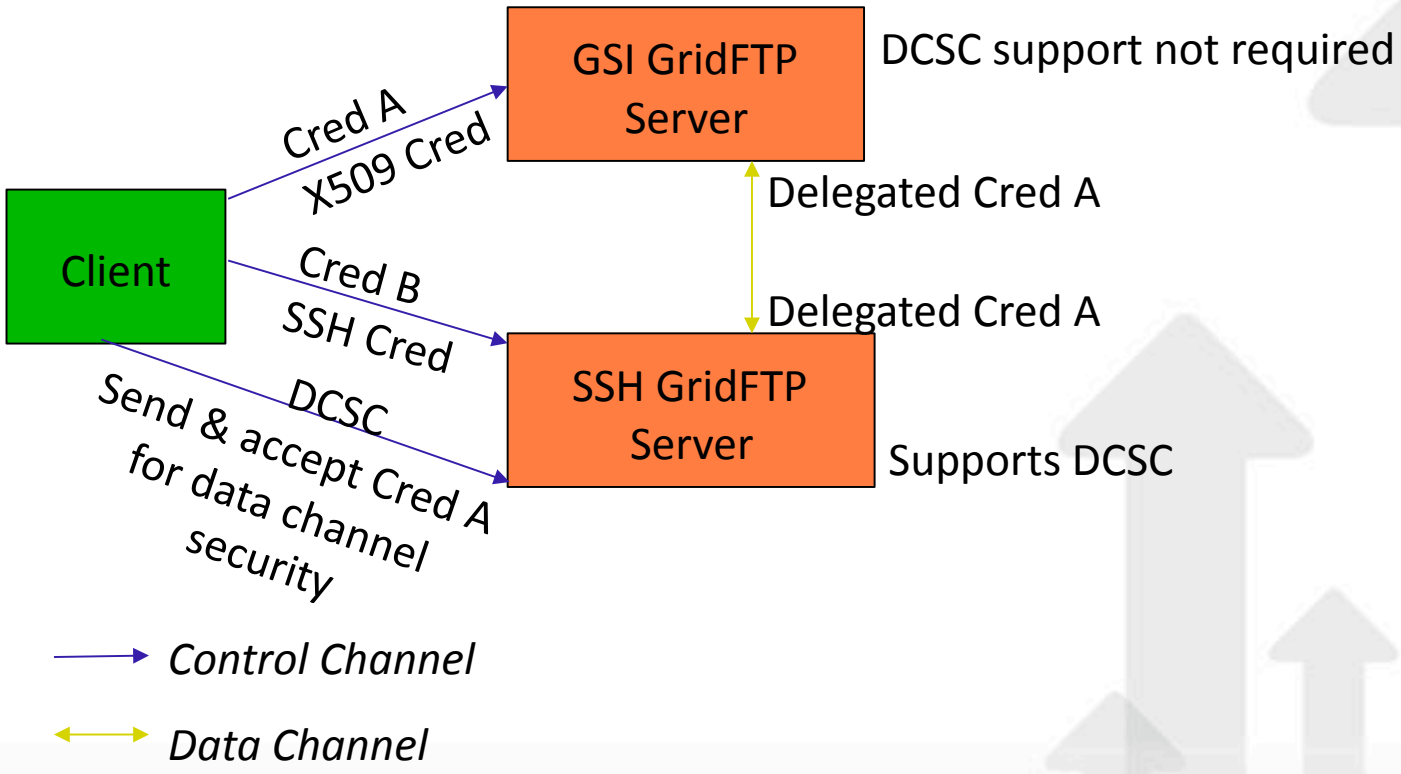
DCSC use cases



DCSC use cases



DCSC use cases



Sync feature

- **New feature to synchronize datasets**
- **Only transfer files where the destination does not exist or differs (size, timestamp, checksum) from the source**
- **'-sync-level'**
  - 0 - transfer if the destination does not exist
  - 1 - transfer if the size does not match
  - 2 - transfer if timestamp of dst is older
  - 3 - transfer if the checksums do not match.
  - The default sync level is 2.

Chrooting GridFTP server

- GridFTP server can be configured to restrict access to specific path
- Server option ‘-chroot-path <path>’
- This path must contain a valid certificate structure, /etc/passwd, and /etc/groups
- A helper script 'globus-gridftp-server-setup-chroot' available in $GLOBUS_LOCATION/sbin, can help create a suitable directory structure

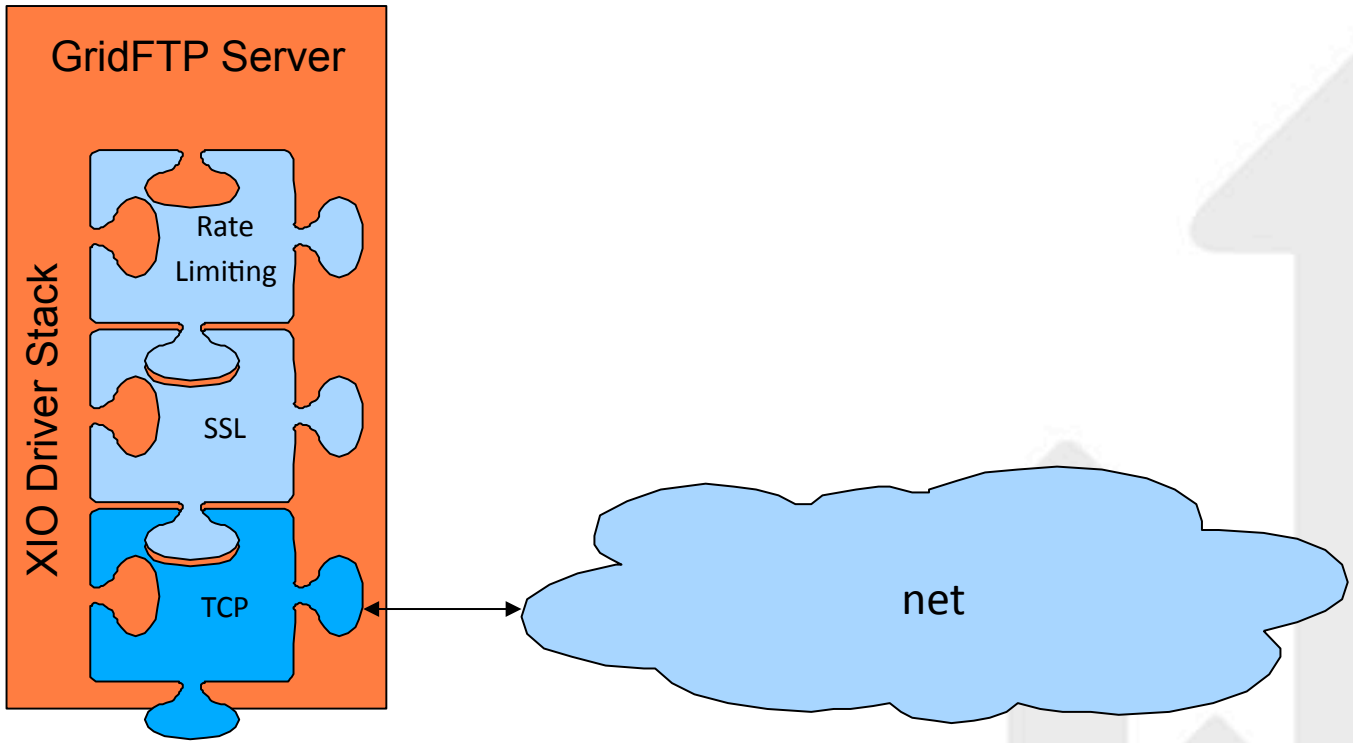
New GridFTP authz callout

- **DOE mandates the use of OTP for all access to LCF resources**
- **To ensure OTP usage for GridFTP access**
  - Authz callout
  - Users store DOEGrids credential to LCF myproxy server (once)
  - Retrieve it using OTP
  - Myproxy server configured to embed an assertion to assert that user used OTP
  - Callout verifies the assertion

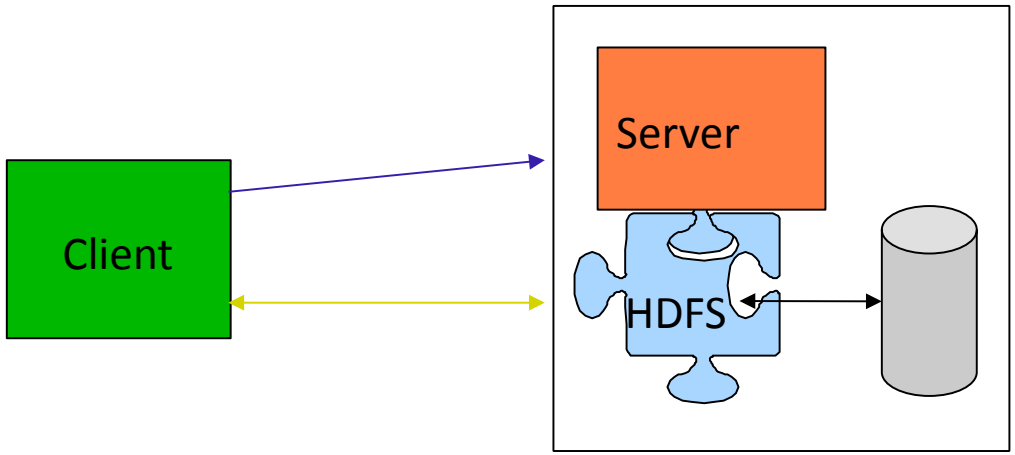
Others

- **Progress markers for stream mode**
  - Mode E directionality limitation
  - Stream mode must be used in certain scenarios.
- **'-disable-command-list <comma separated strings>' in the server has been enhanced**
  - Command strings can take arguments
  - -disable-command-list 'DCAU N' or -disable-command-list 'DCAU N', 'MODE E' can be done.

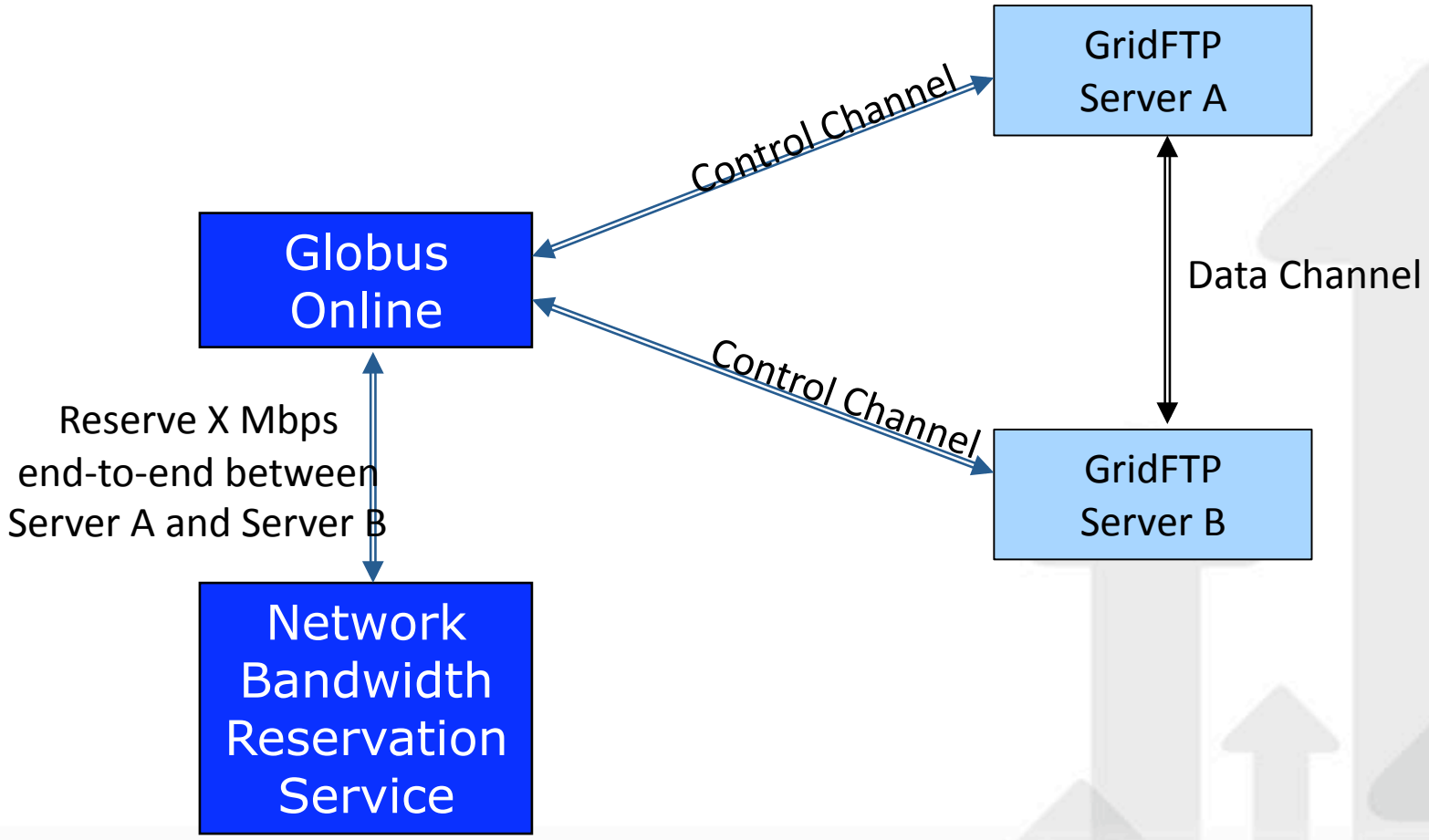
Rate Limiting



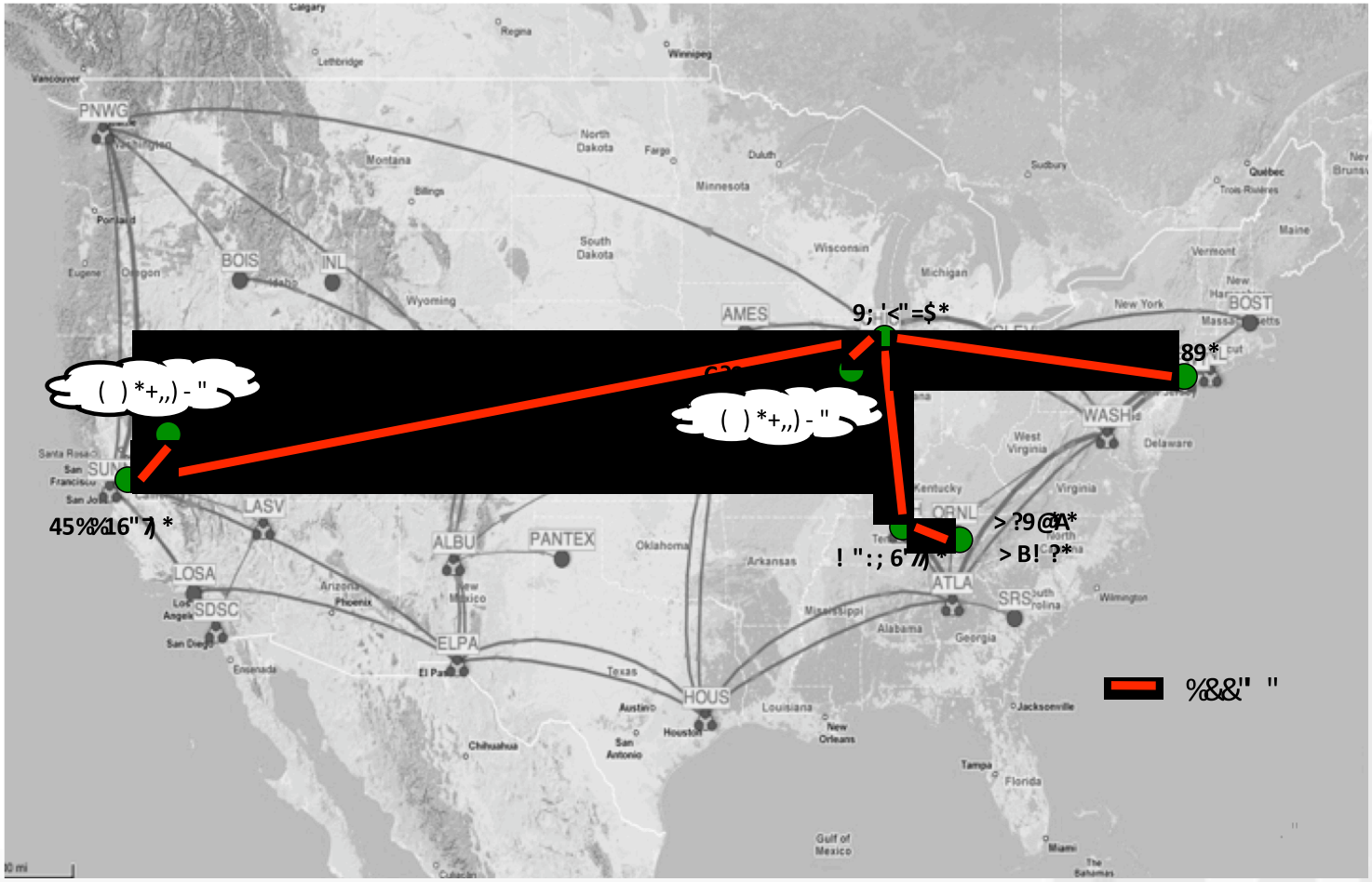
HDFS Plugin



Network Reservation



GridFTP on 100G



Terabit Networking

- Exascale computer – 2018
- Terabit Networks – 2015/2016
- Numerous challenges in getting these sort of rates end-to-end, file system-to-file system
- Need for large-scale parallelism at endpoints
- New protocol designs

GridFTP V3

- **Enable more efficient and reliable support for highly parallel end systems**
- **Enhanced protocol will enable**
  - Dynamically create and coordinate tens to hundreds of servers at end points
  - Efficiently transfer directories containing millions of files ranging from KBs to TBs
  - Overlay networks to independently optimize different network hops.

Acknowledgments

- **Mike Link**
- **Lukasz Lacinski**

Questions?