the globus alliance
www.globus.org

Data Movement Tools for Distributed Petascale Science

Raj Kettimuthu
Argonne National Laboratory and
The University of Chicago

the globus alliance
www.globus.org

Outline

- Introduction
- Motivation
- Data Transfer Problem
- Requirements
- Reliable Data Movement Framework
- Future Directions

Today's Science Environments

- Large-scale collaborative science is becoming increasingly common

- Distributed community of users to access and analyze large amounts of data

Simulation Science

- In simulation science, the data sources are supercomputer simulations
  - For eg, DOE-funded climate modeling groups generate large reference simulations at supercomputer centers
- Combustion, fusion, computational chemistry, and astrophysics communities have similar requirements for remote and distributed data analysis

Experimental Science

- Data sources are facilities such as high energy and nuclear physics experiments and light sources.
  - For eg, LHC at CERN will produce petabytes of raw data per year for 15 years
- DOE light sources can also produce large quantities of data that must be distributed, analyzed, and visualized
- The international fusion experiment, ITER

Science Environments

- Raw simulation or observational data is just a starting point for most investigations
- Understanding comes from further analysis, reduction, visualization, and exploration

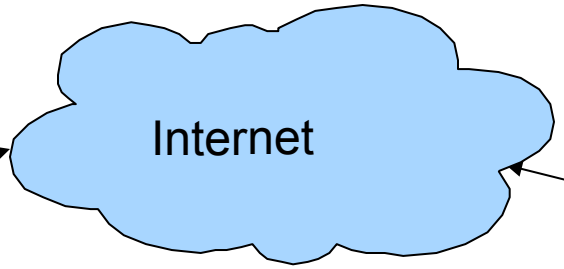
Petascale resource

Compute Cluster

Scientist's Desktop

- Furthermore the data is a community asset that must be accessible to any member of a distributed collaboration

Network Capabilities

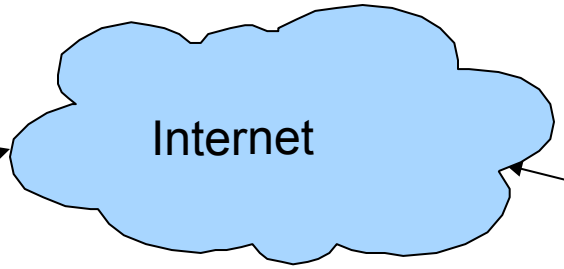


Scientist A in California

Scientist B in New York

- Scientist A wants to transfer 1 Terabyte of data to Scientist B
- What is the fastest way to transfer the data?

Network Capabilities



Scientist A in California

Scientist B in New York

- Scientist A wants to transfer 1 Terabyte of data to Scientist B
- What is the fastest way to transfer the data?

FedEx

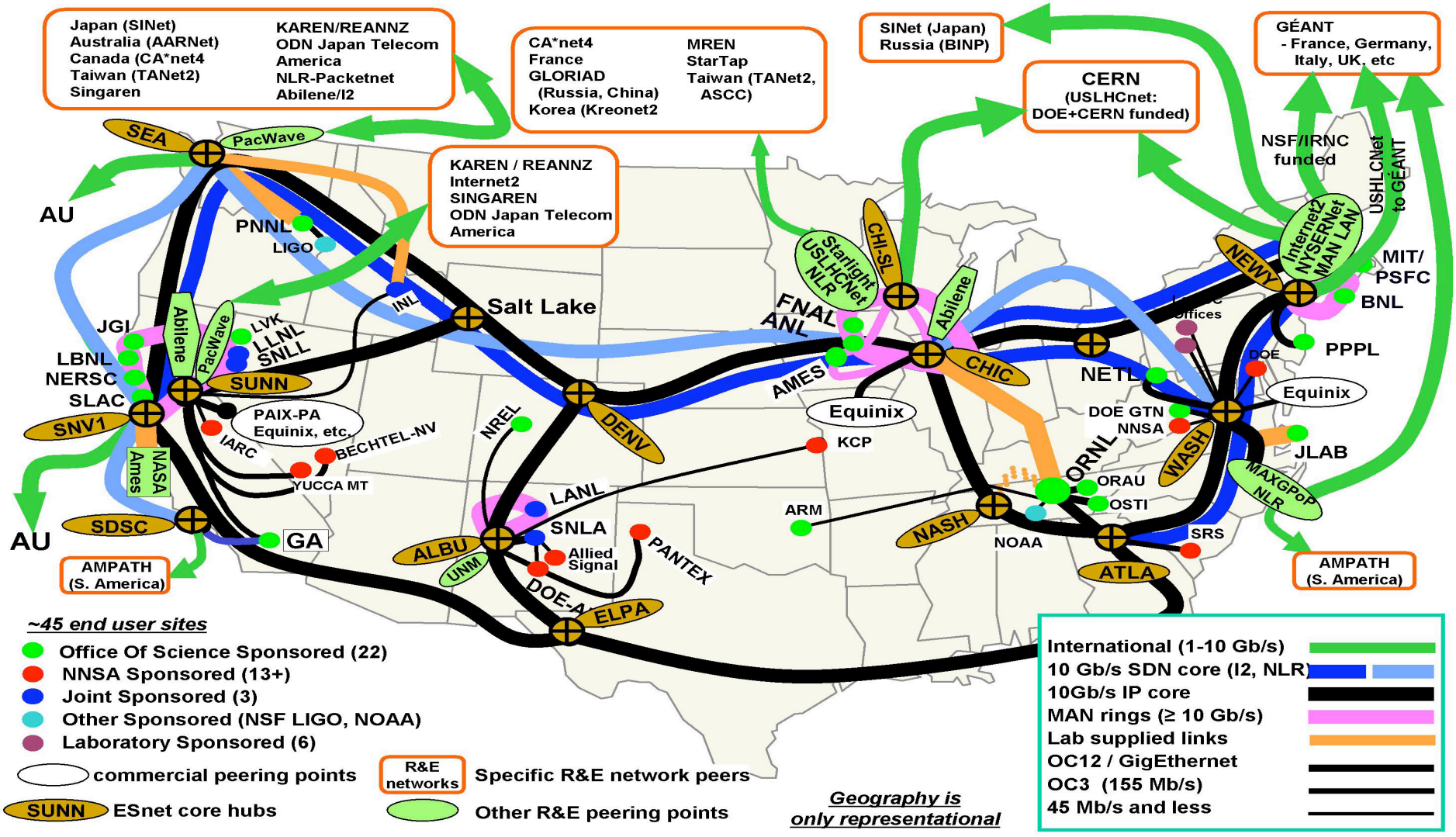
Network Capabilities

- Until a few years ago, Tri-labs (Los Alamos, Lawrence Livermore and Sandia) transferred data via tapes sent thru fedex
- To transfer 100 TB in 24 hours, need a sustained data rate > 9.5 Gbit/s
- 10 Gbit/s networks are becoming increasingly common in scientific environments
  - DOE's ESNet, UltraScience Net, Science Data Networks and Internet2 have 10Gb/s or higher links
  - Thanks to the advancement in networking technologies

the globus alliance
www.globus.org

ESNET

ESnet Provides Global High-Speed Internet Connectivity for DOE Facilities and Collaborators (12/2007)



End-to-end problem

- Now that high-speed networks are available, can we move data at network speeds on the network?
- What if the speed of airplanes had increased by the same factor as computers over the last 50 years, namely five orders of magnitude?

End-to-end problem

- Now that high-speed networks are available, can we move data at network speeds on the network?
- What if the speed of airplanes had increased by the same factor as computers over the last 50 years, namely five orders of magnitude?

We would be able to cross US in less than a second

End-to-end problem

- Now that high-speed networks are available, can we move data at network speeds on the network?
- What if the speed of airplanes had increased by the same factor as computers over the last 50 years, namely five orders of magnitude?

We would be able to cross US in less than a second

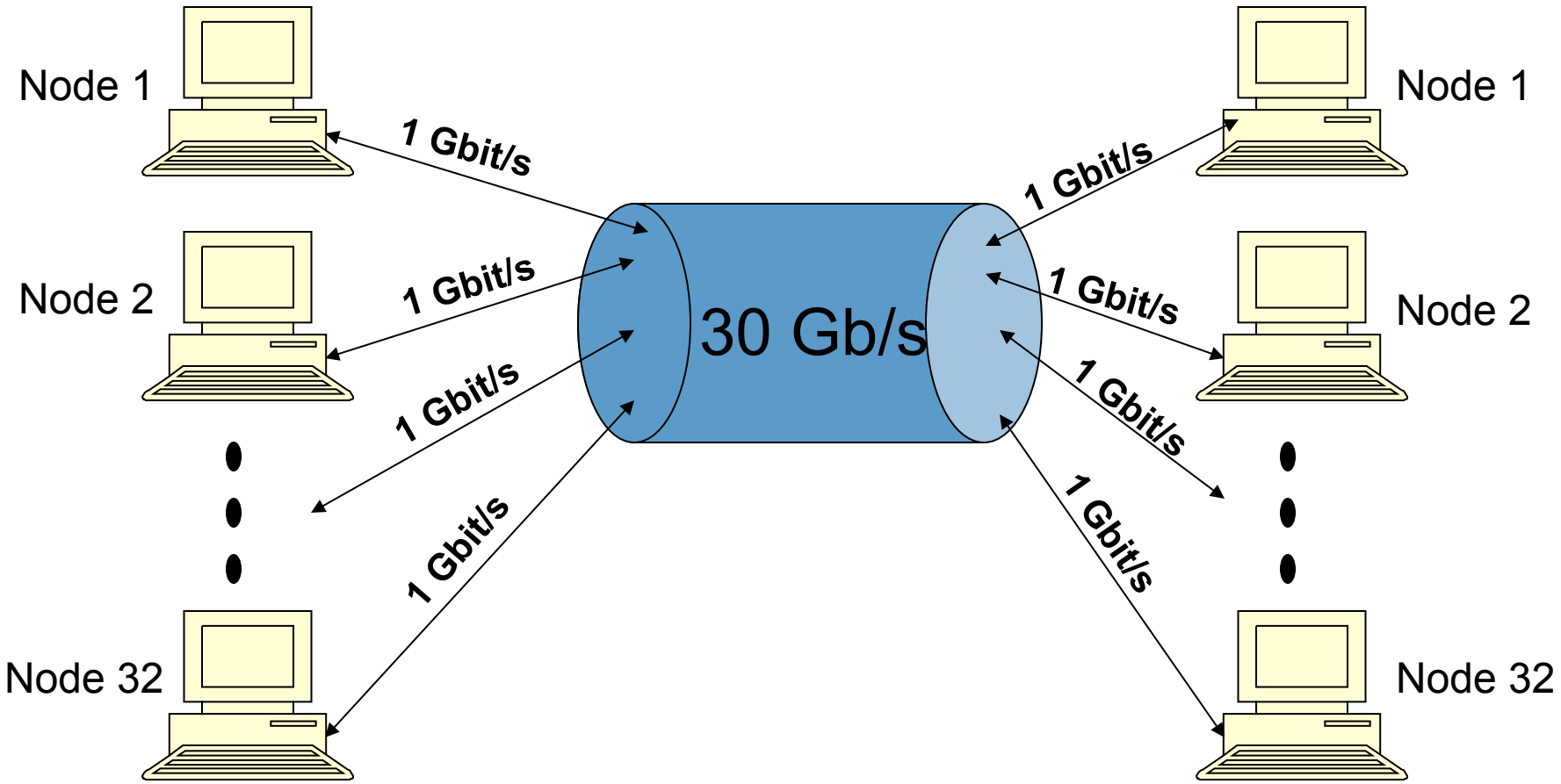
Yes. But it would still take two hours to get to downtown

End-to-end problem

- Data movement in distributed science environments is an end-to-end problem
- A 10 Gbit/s network link between the source and destination does not guarantee an end-to-end data rate of 10 Gbit/s
- Other factors such as storage system, disk, data rate supported by the end node
- Deal with failures of various sorts
  - Firewalls can cause difficulties

End-to-end data transfer

Efficient and robust wide area data transport requires the management of complex systems at multiple levels.



Requirements

- Fast
- Easy-to-use
- Secure
- Reliable
- Extensible
- Standard
- Robust

the globus alliance
www.globus.org

GridFTP

- High-performance, reliable data transfer protocol optimized for high-bandwidth wide-area networks
- Based on FTP protocol - defines extensions for high-performance operation and security
- Standardized through Open Grid Forum (OGF)
- GridFTP is the OGF recommended data movement protocol

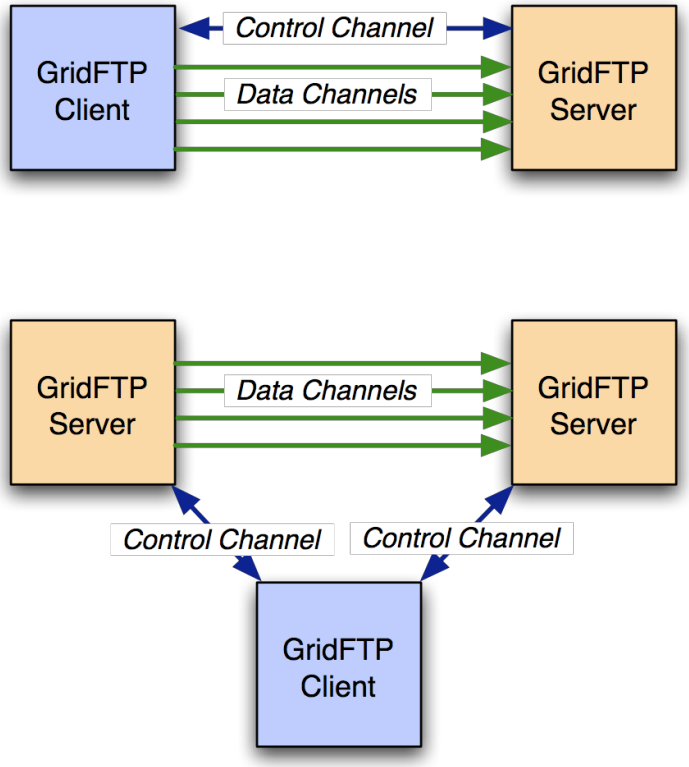
the globus alliance
www.globus.org

GridFTP

- We (Globus Alliance) supply a reference implementation:
  - Server
  - Client tools
  - Development Libraries
- Multiple independent implementations can interoperate
  - Fermi Lab and U. Virginia have home grown servers that work with ours

GridFTP

- Two channel protocol like FTP
- Control Channel
  - Communication link (TCP) over which commands and responses flow
  - Low bandwidth; encrypted and integrity protected by default
- Data Channel
  - Communication link(s) over which the actual data of interest flows
  - High Bandwidth; authenticated by default; encryption and integrity protection optional

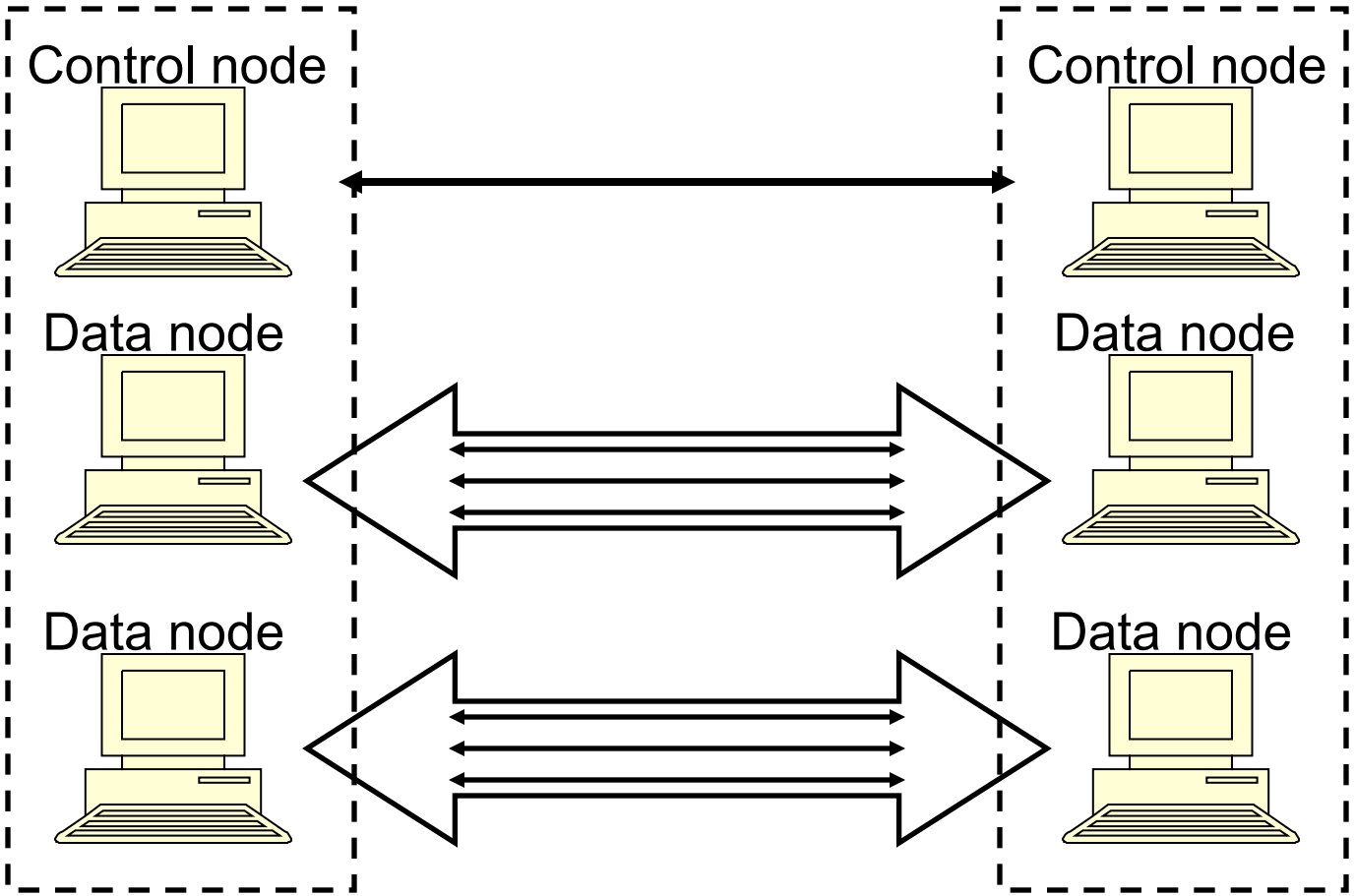


Globus GridFTP Features

- GridFTP is Fast
  - Parallel TCP streams
  - Non TCP protocol such as UDT
  - Set optimal TCP buffer sizes
  - Order of magnitude greater
- Cluster-to-cluster data movement
  - Co-ordinated data movement using multiple computers at each end
  - Another order of magnitude

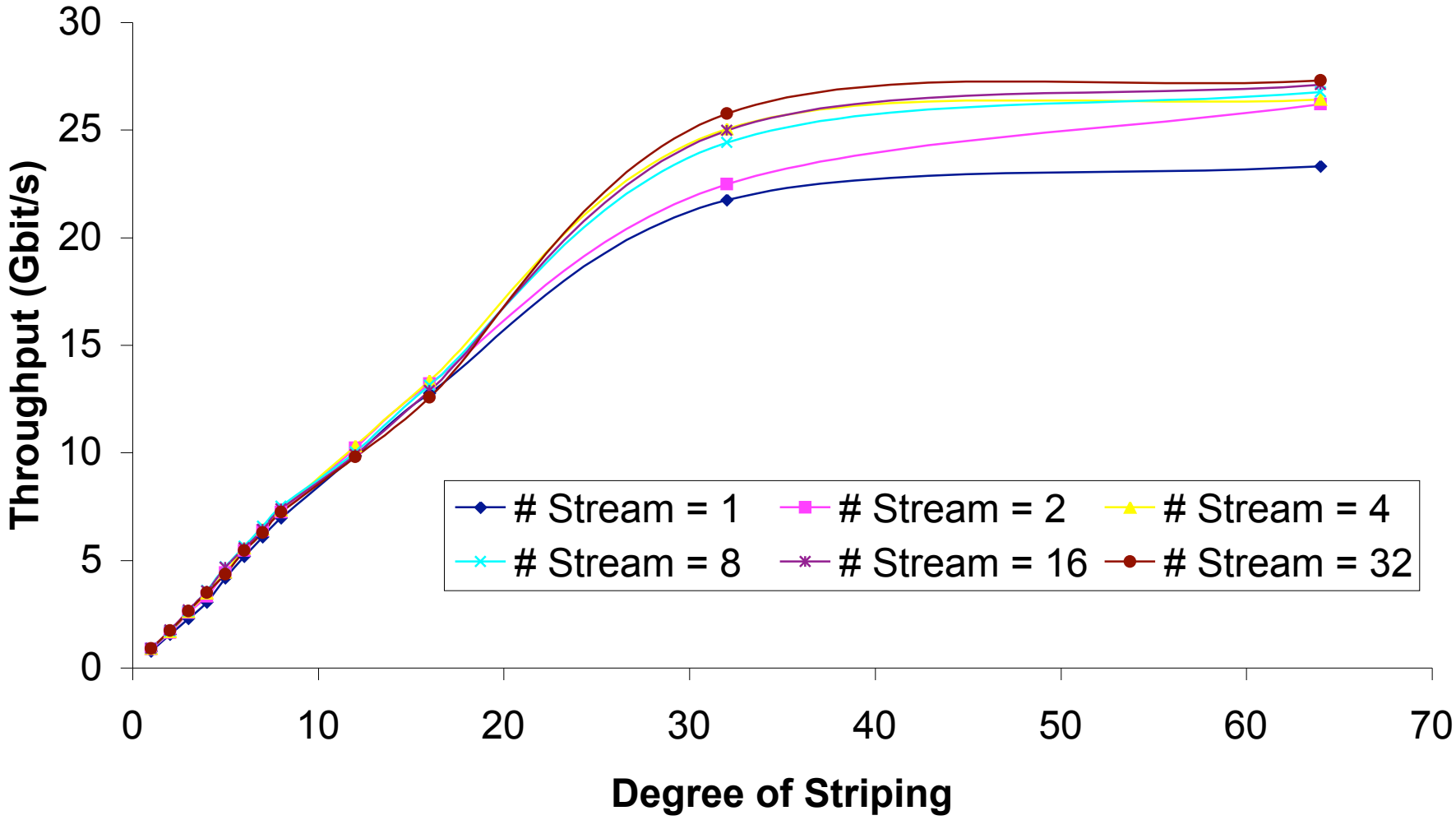
"Grid-enabled Particle Physics Event Analysis: Experiences Using a 10 Gb, High-latency Network for a High-Energy Physics Application", FGCS Journal, August 2003

Cluster-to-Cluster transfers



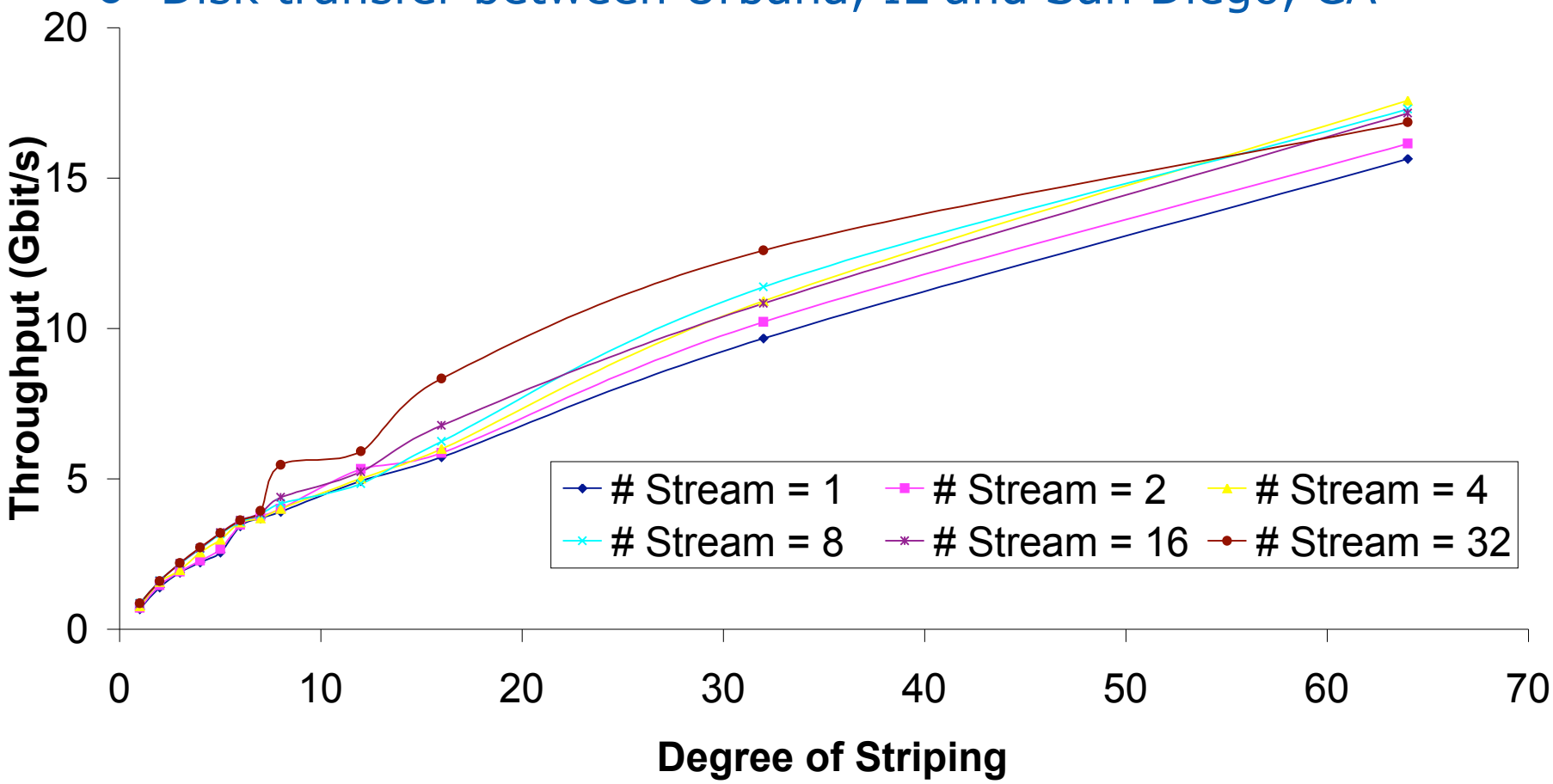
Performance

- Mem. transfer between Urbana, IL and San Diego, CA



Performance

- Disk transfer between Urbana, IL and San Diego, CA



"The Globus Striped GridFTP Framework and Server", ACM/IEEE conference on Supercomputing (SC'05)

Security

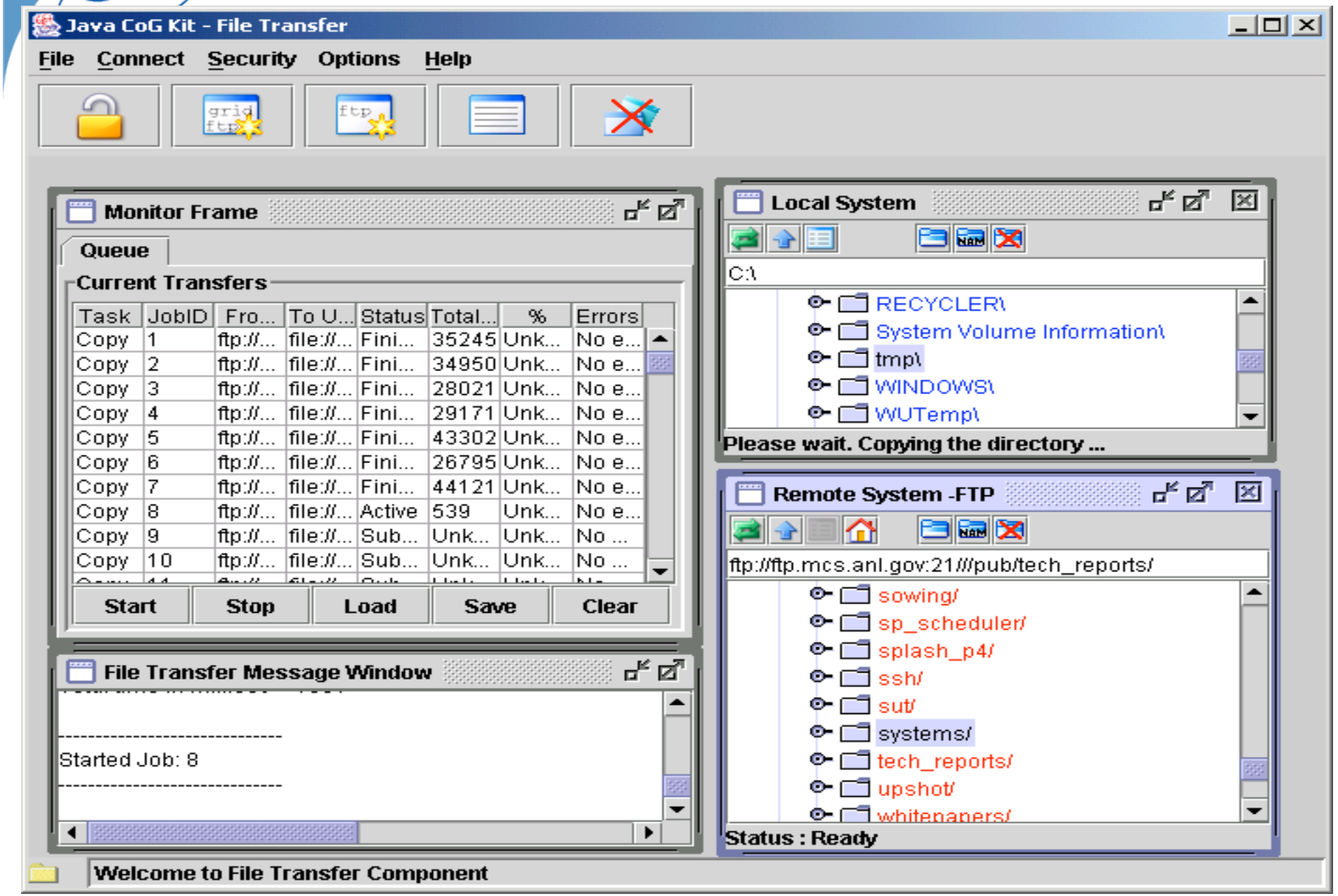
- Often there is need to authenticate clients and control access to the data
- Globus GridFTP supports multiple security mechanisms to authenticate and authorize clients
  - Anonymous access
  - Username/password
  - SSH security
  - Grid Security Infrastructure (GSI)

Easy-to-use

- Simple to install
  - Configure; make gridftp install;
  - Installs only gridftp and its dependencies
  - Binaries available for many platforms
- Various clients
  - Command-line client - globus-url-copy
  - Client libraries - well-defined API
  - Graphical User Interface

the globus alliance
www.globus.org

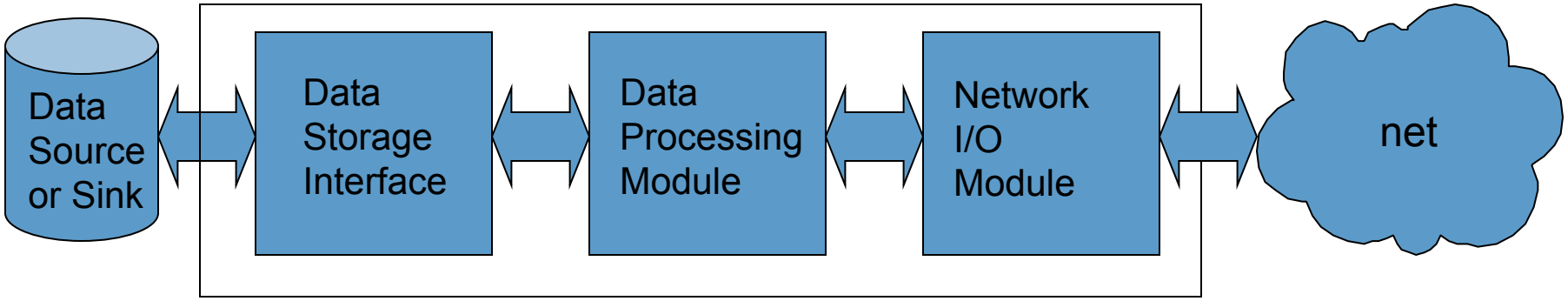
GUI Client



Requirements

- Fast ✓
- Secure ✓
- Reliable
- Extensible
- Standard ✓
- Robust
- Easy-to-use ✓

Modular



Well defined interfaces

Data Storage Interface (DSI)

- POSIX file system
- High Performance Storage System (HPSS)
- Storage Resource Broker (SRB)

"Globus Data Storage Interface (DSI) - Enabling Easy Access to Grid Datasets", Data Grids Workshop 2006

the globus alliance
www.globus.org

Modular

- Network I/O module
  - Simple Open/Close/Read/Write interface
  - Well-defined abstraction called drivers
  - Easy to plug-in external libraries
  - TCP, UDT, Phoebus
- Data processing module
  - Compression (under development)
  - Checksum

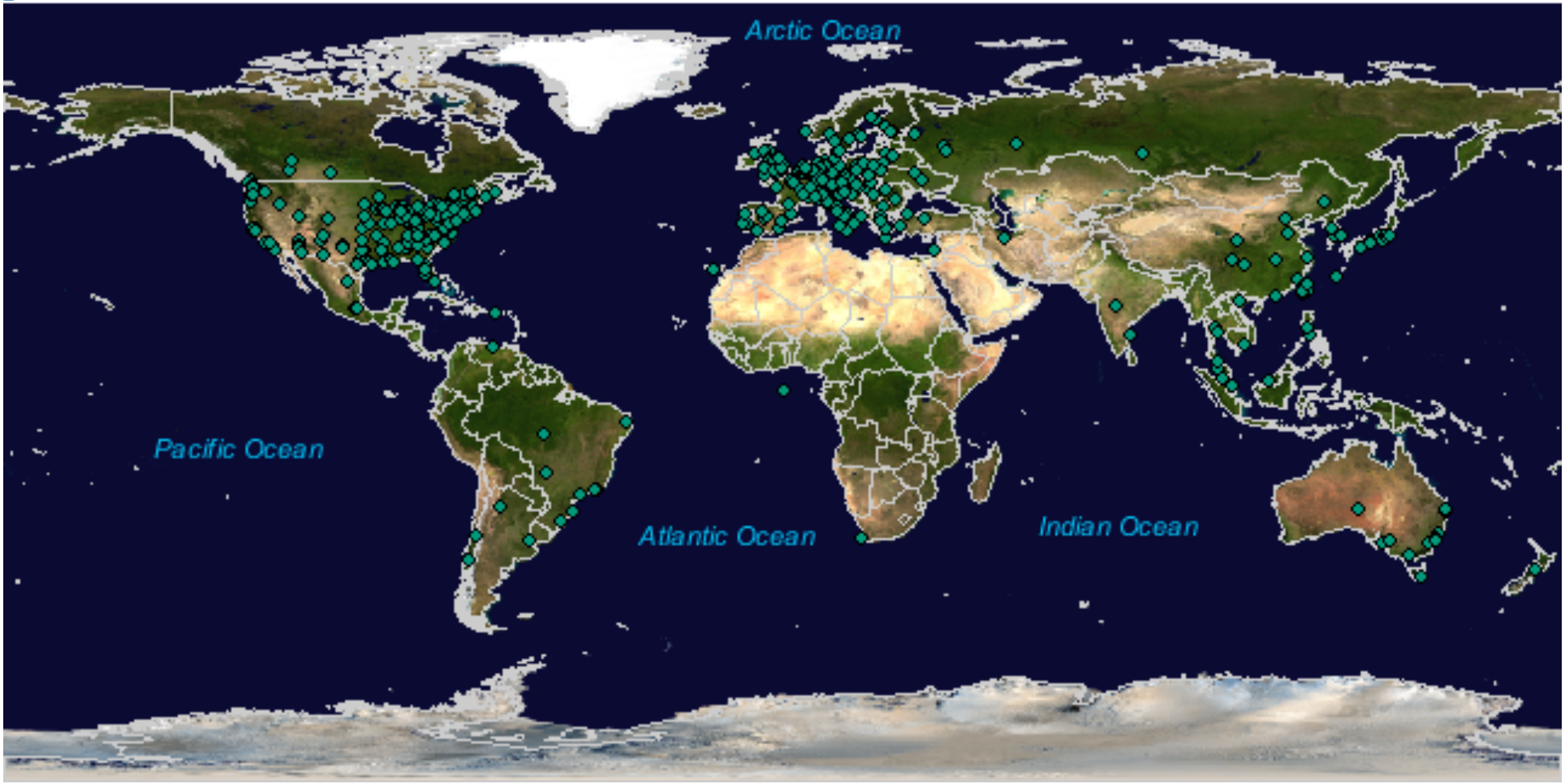
"The Globus eXtensible Input/Output System (XIO): A protocol independent IO system for the Grid", IEEE IPDPS 2005

GridFTP in production

- Many Scientific communities rely on GridFTP
  - High Energy Physics - LHC computing Grid
  - Southern California Earthquake Center (SCEC), Earth Systems Grid (ESG), Relativistic Heavy Ion Collider (RHIC), European Space Agency, BBC use GridFTP for data movement
- GridFTP facilitates an average of more than 3 million data transfers every day

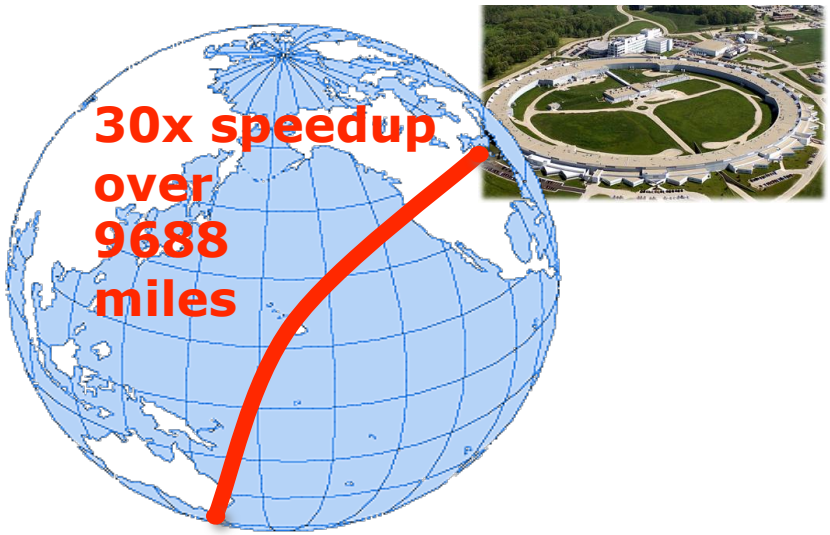
the globus alliance
www.globus.org

GridFTP Servers Around the World

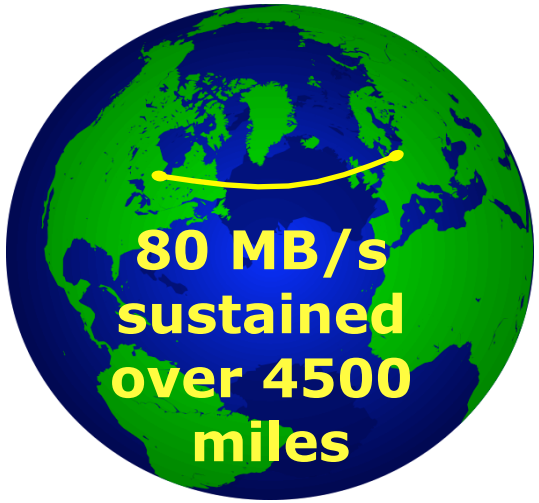


Created by Lydia Prieto ; G. Zarrate; Anda Imanitchi (Florida State University) using MaxMind's GeoIP technology (http://www.maxmind.com/app/ip-locate).

GridFTP in production



One terabyte moved from an Advanced Photon Source tomography beamline to Australia, at a rate 30x faster than standard FTP



1.5 terabyte moved from University of Wisconsin, Milwaukee to Hannover, Germany at a sustained rate of 80 megabyte/sec

Handling failures

- GridFTP server sends restart and performance markers periodically
  - Default every 5s - configurable
- Helpful if there is any failure
  - No need to transfer the entire file again
  - Use restart markers and transfer only the missing pieces
- GridFTP supports partial file transfers

Server failure

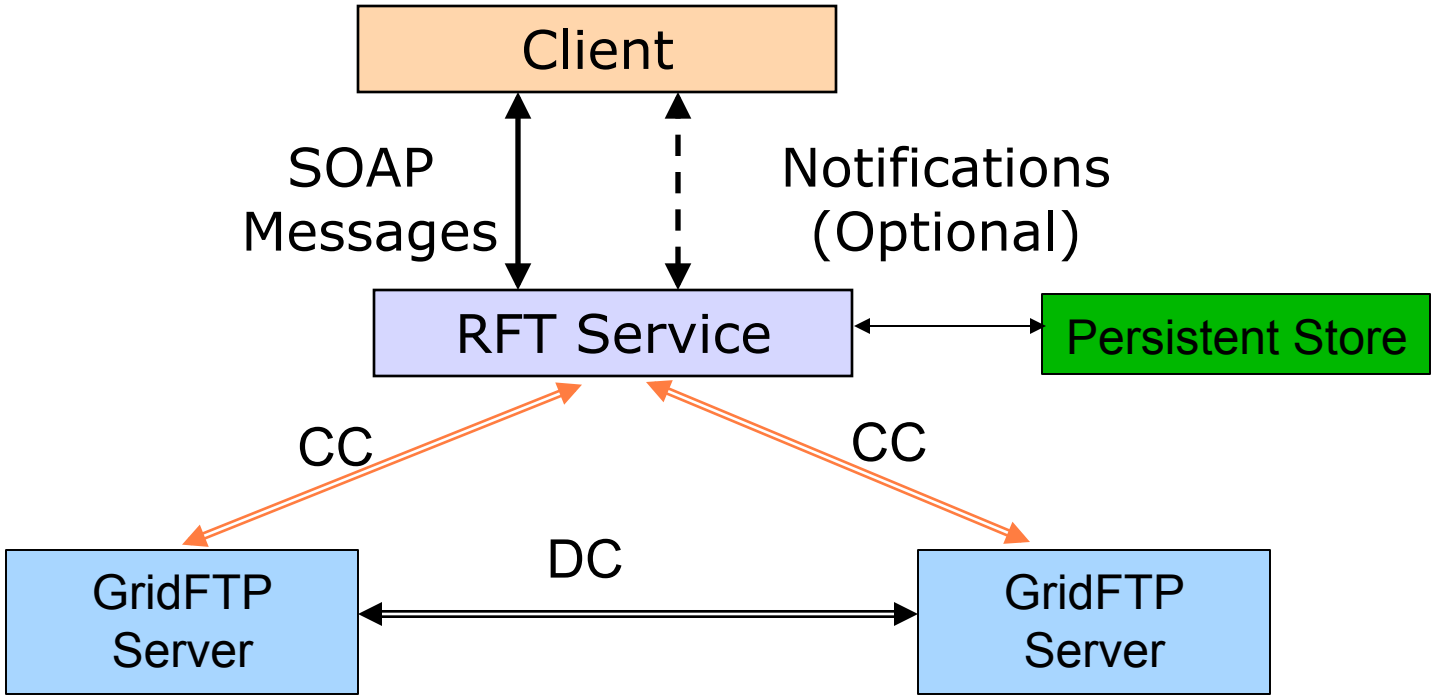
- Command-line client - globus-url-copy - support transfer retries
  - Use restart markers
- Recover from server and connection failures
- What if the client fails in the middle of a transfer?

the globus alliance
www.globus.org

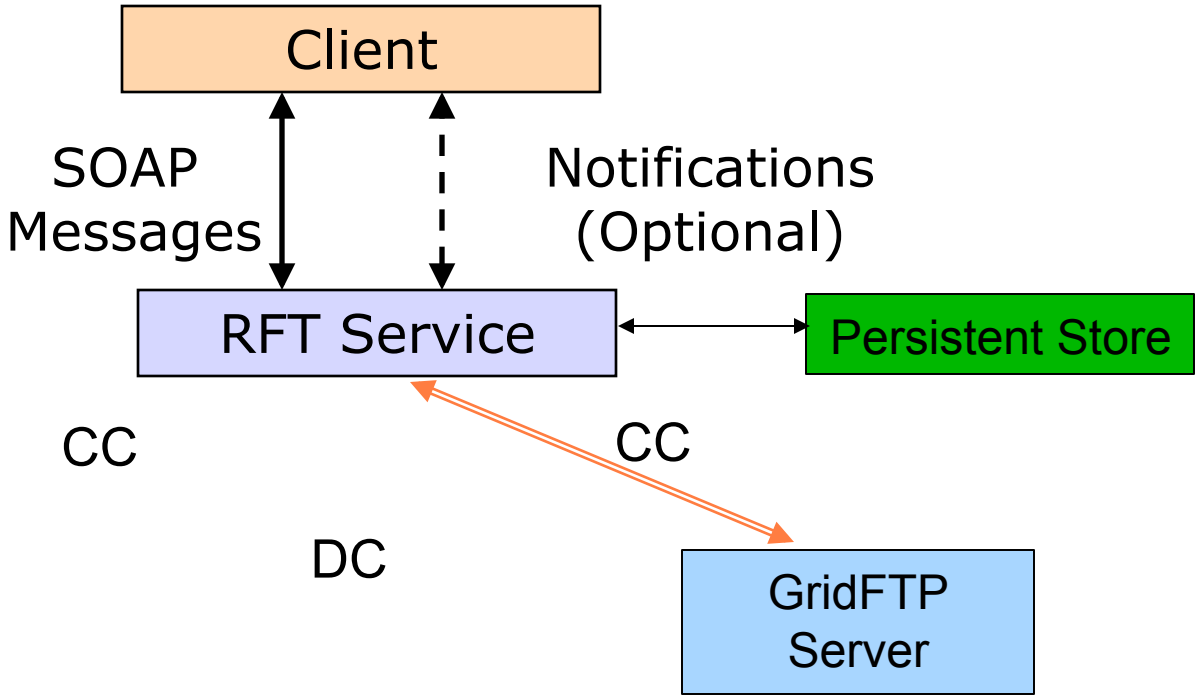
Globus Reliable File Transfer Service (RFT)

- GridFTP client that provides more reliability
- GridFTP - on demand transfer service
  - Not a queuing service
- RFT
  - Queues requests
  - Orchestrates transfers on client's behalf
  - Writes to persistent store
  - Recovers from GridFTP and RFT service failures

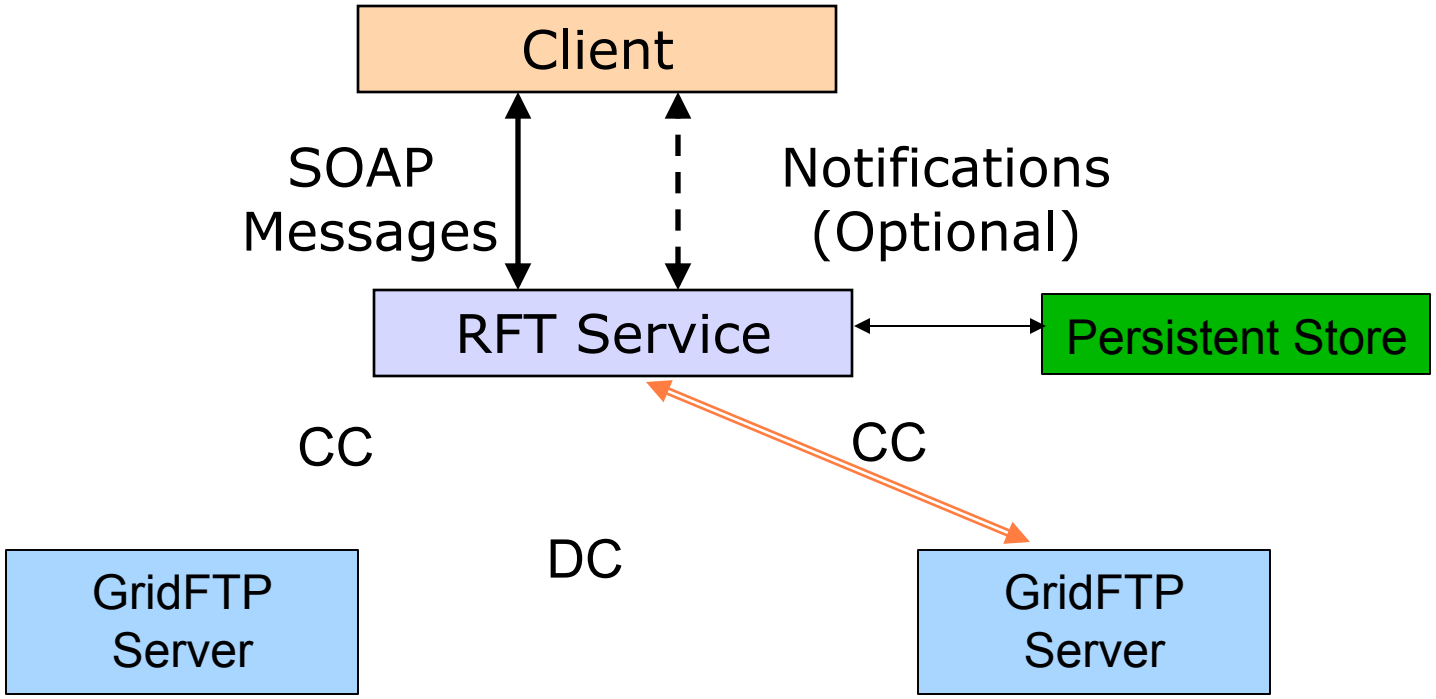
RFT


Client
SOAP
Messages
Notifications
(Optional)
RFT Service
Persistent Store
CC
CC
DC
GridFTP
Server
GridFTP
Server

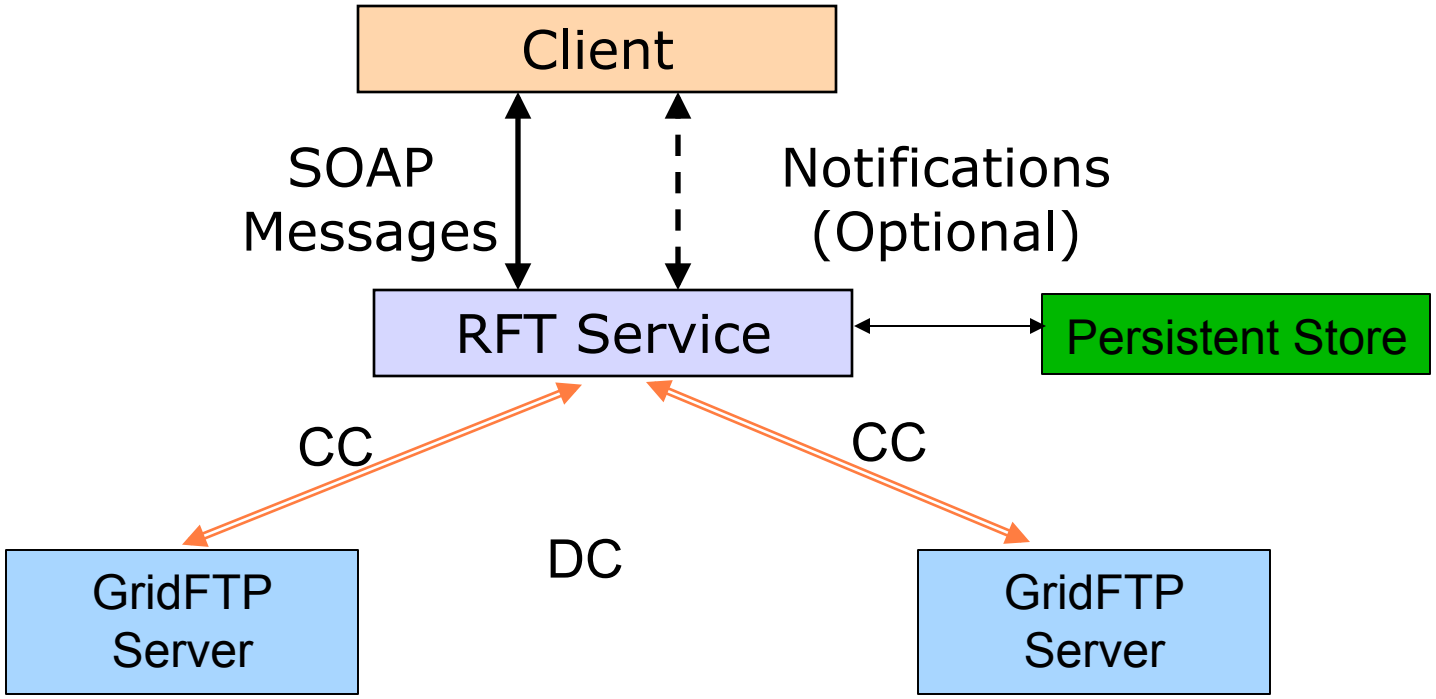
RFT


Client
SOAP
Messages
Notifications
(Optional)
RFT Service
Persistent Store
CC
CC
DC
GridFTP
Server

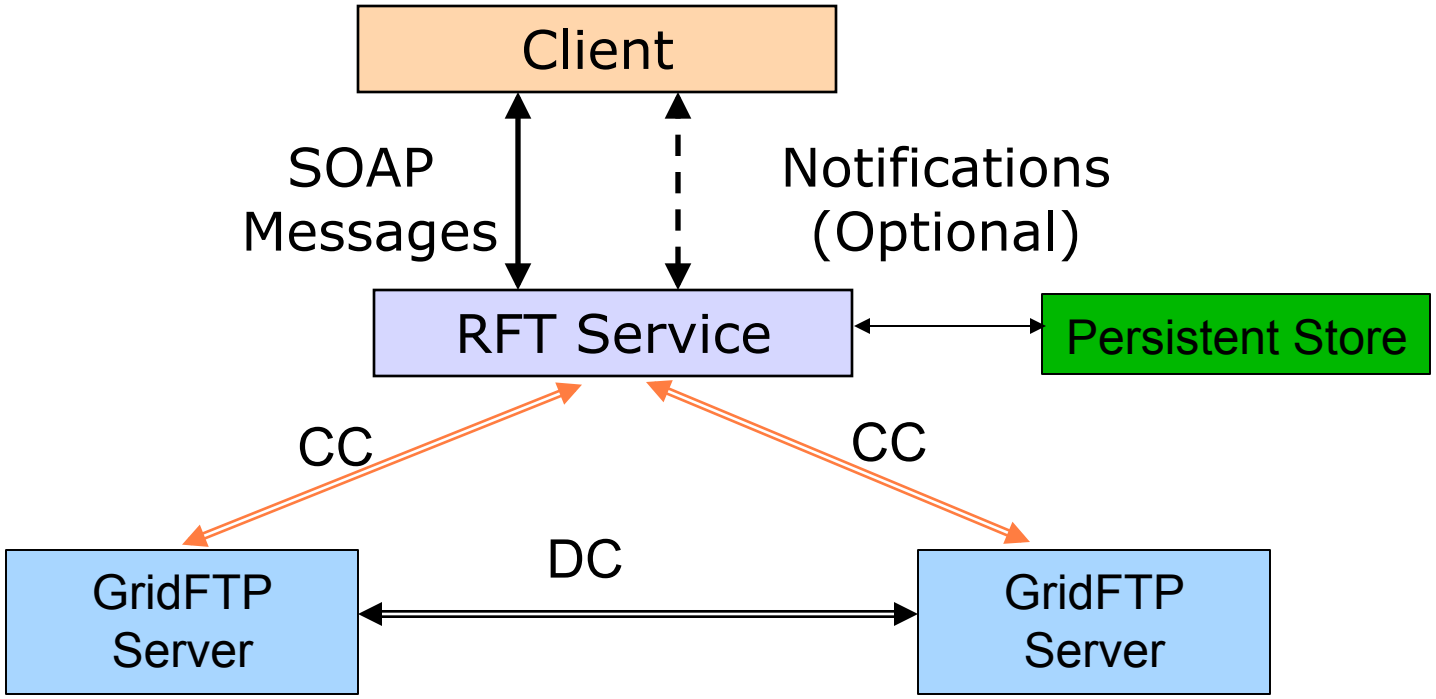
RFT


Client
SOAP
Messages
Notifications
(Optional)
RFT Service
Persistent Store
CC
CC
DC
GridFTP
Server
GridFTP
Server

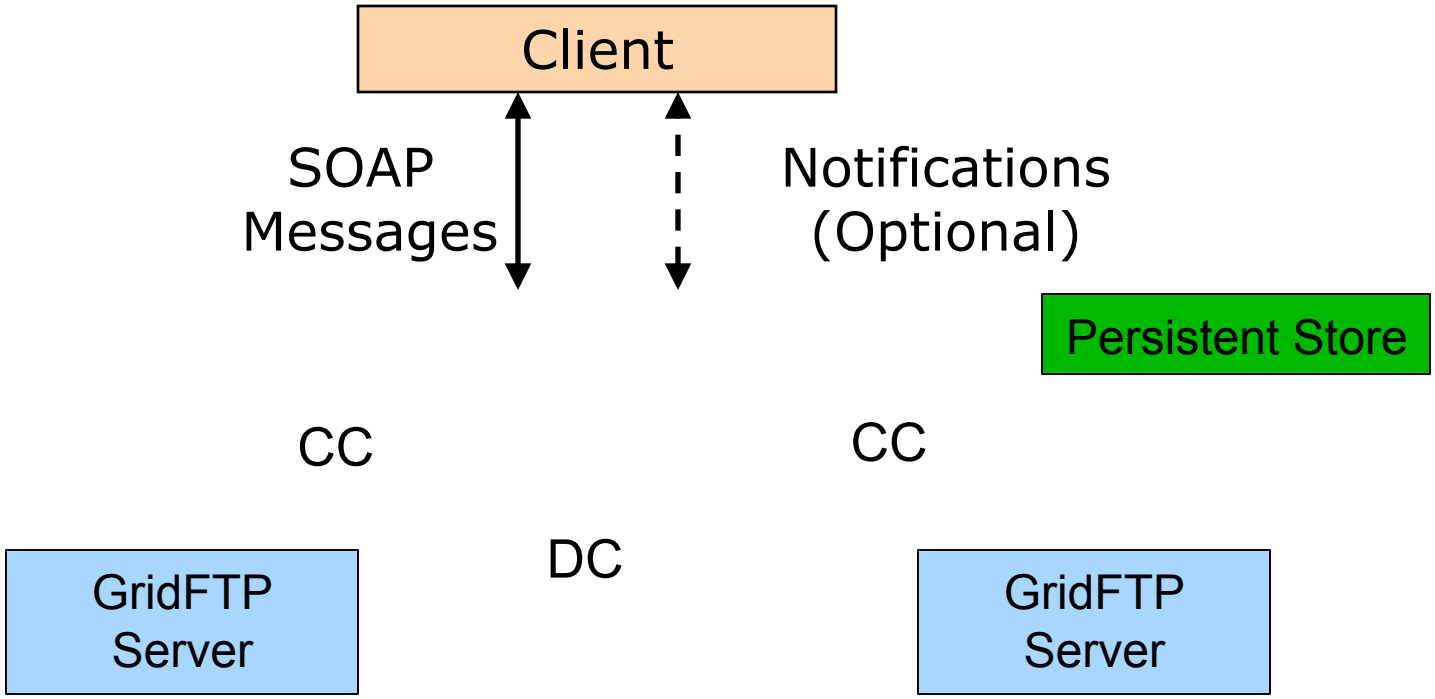
RFT


Client
SOAP
Messages
Notifications
(Optional)
RFT Service
Persistent Store
CC
CC
DC
GridFTP
Server
GridFTP
Server

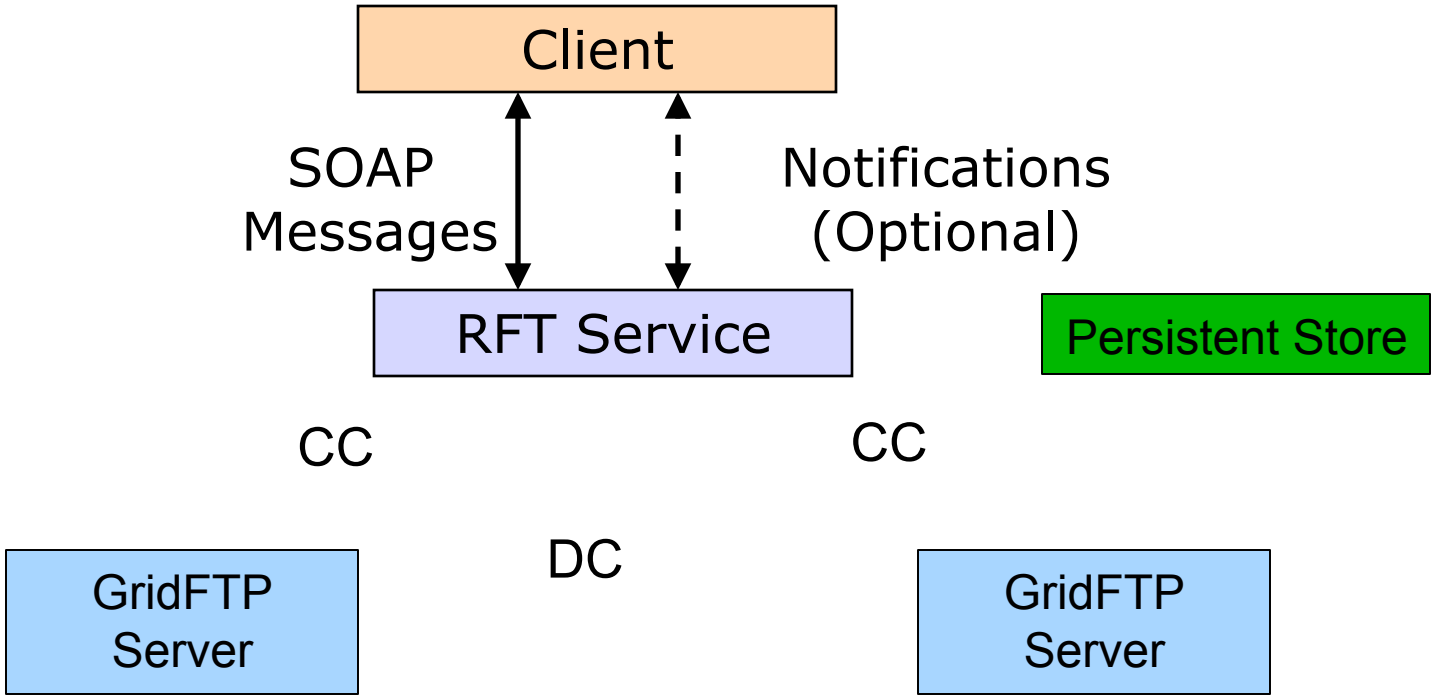
RFT


Client
SOAP
Messages
Notifications
(Optional)
RFT Service
Persistent Store
CC
CC
DC
GridFTP
Server
GridFTP
Server

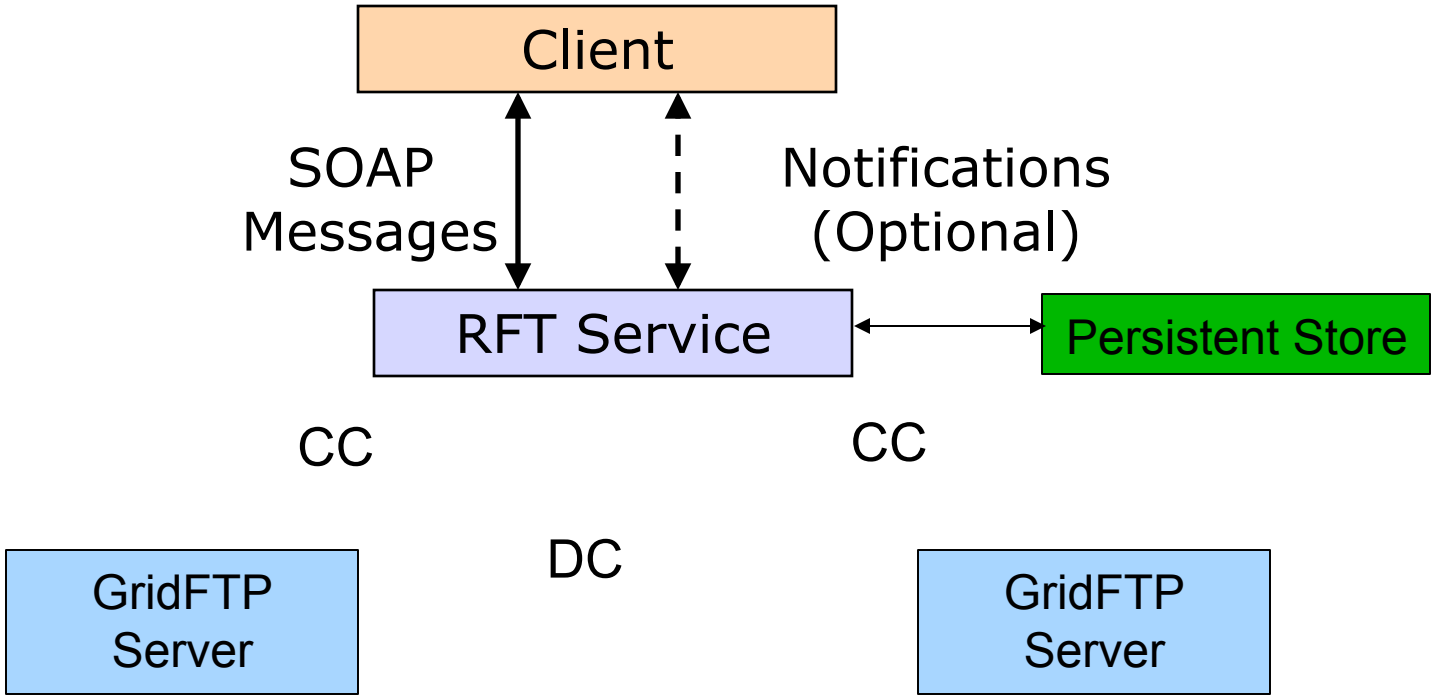
RFT


Client
SOAP
Messages
Notifications
(Optional)
Persistent Store
CC
CC
DC
GridFTP
Server
GridFTP
Server

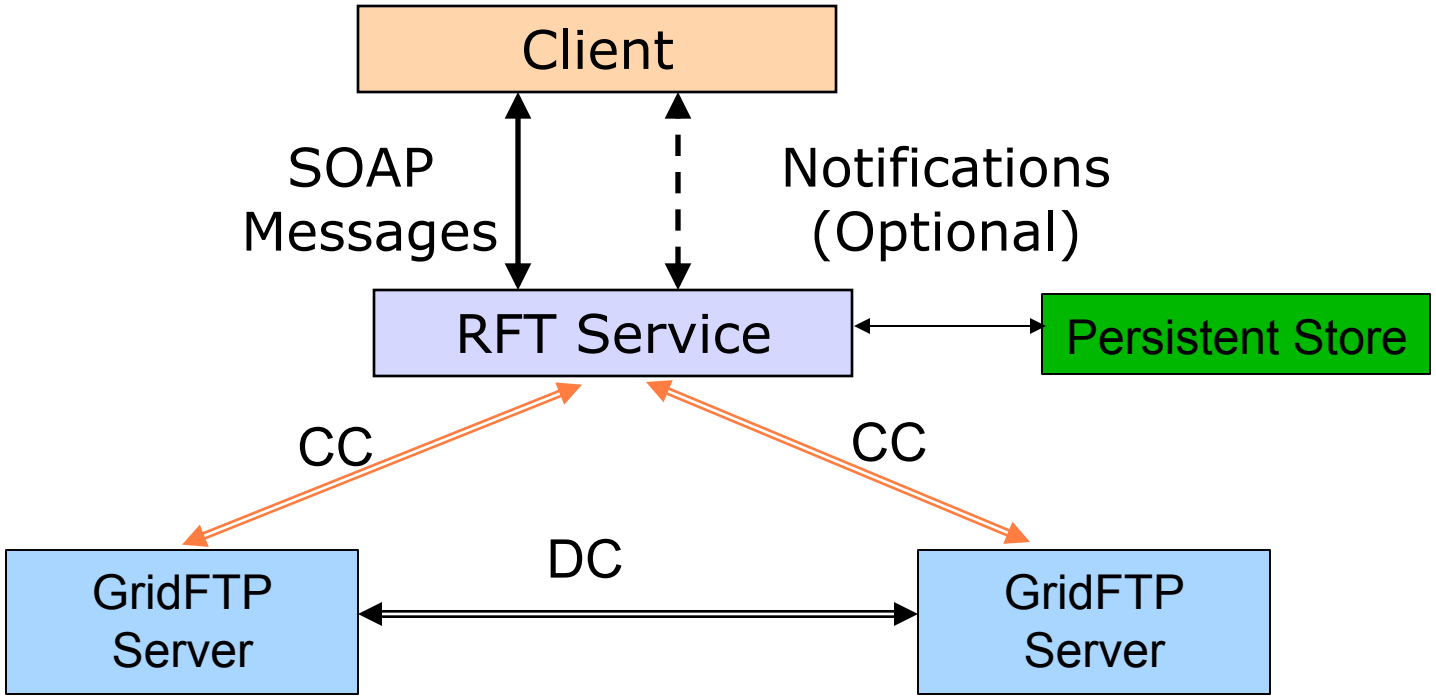
RFT


Client
SOAP
Messages
Notifications
(Optional)
RFT Service
Persistent Store
CC
CC
DC
GridFTP
Server
GridFTP
Server

RFT


Client
SOAP
Messages
Notifications
(Optional)
RFT Service
Persistent Store
CC
CC
DC
GridFTP
Server
GridFTP
Server

RFT

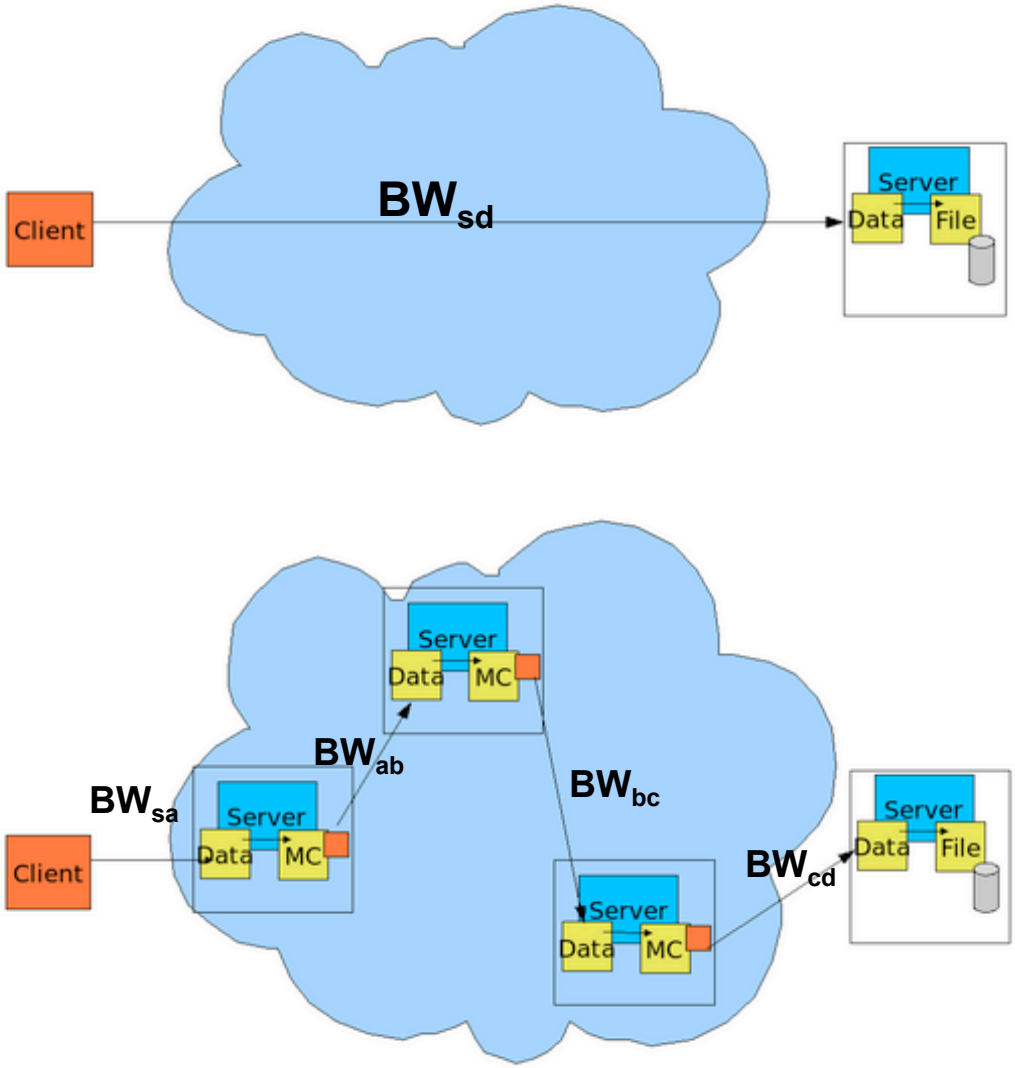

Client
SOAP
Messages
Notifications
(Optional)
RFT Service
Persistent Store
CC
CC
DC
GridFTP
Server
GridFTP
Server

Requirements

- Fast ✓
- Secure ✓
- Reliable ✓
- Extensible ✓
- Standard ✓
- Robust ✓
- Easy-to-use ✓

GridFTP

GridFTP Overlay Network



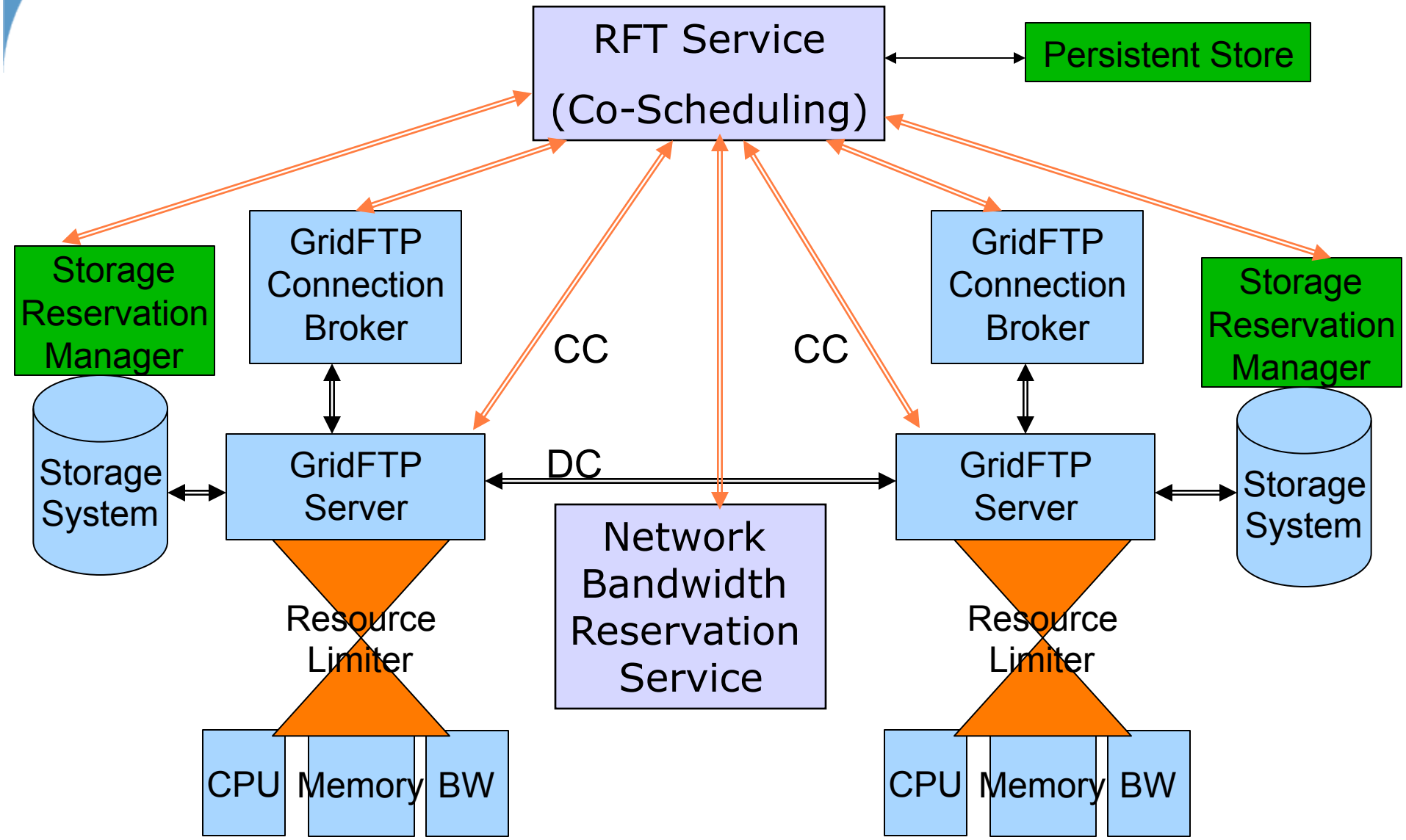
If Min(BW_{sa} , BW_{ab} , BW_{bc} , BW_{cd}) > BW_{sd},

Overlay route yields better performance

Best effort service

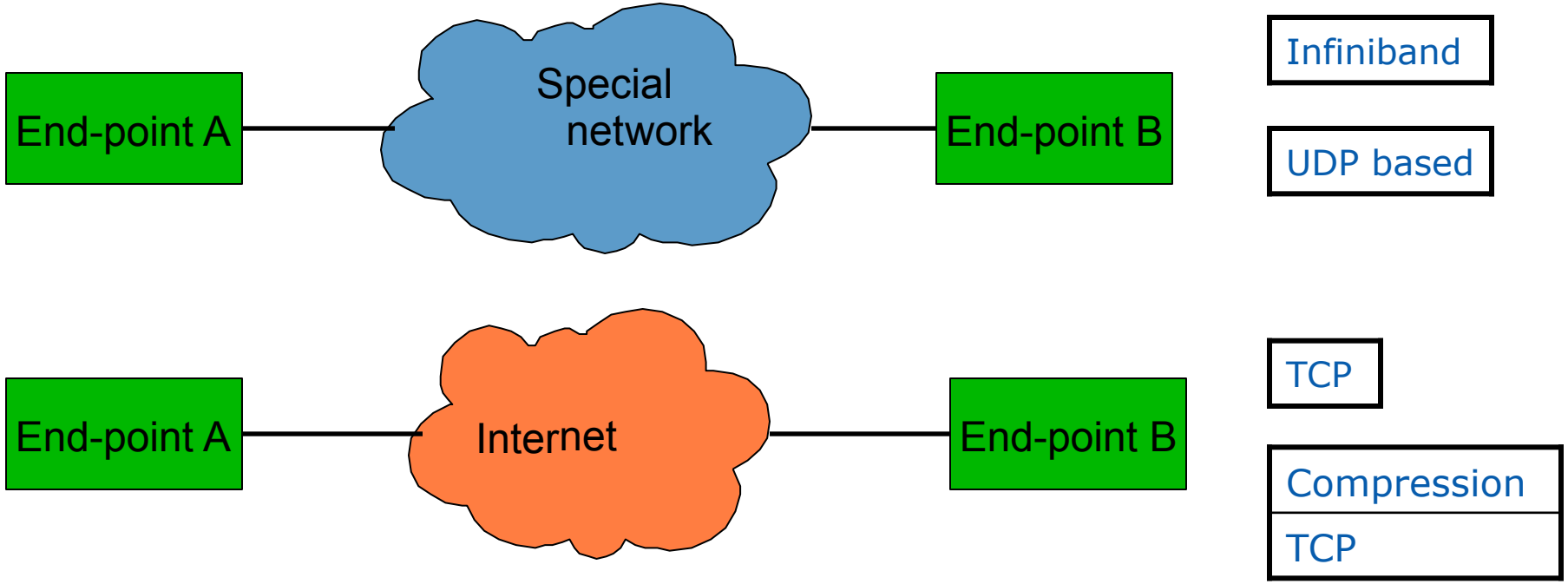
- Data movement in distributed environments is on best effort basis
- No Quality of Service (QoS) guarantees
- Network is shared
- Limited disk space
  - Destination might run out of space in the middle of a transfer
- End node, network, disk can fail any time

Managed Data Movement



Dynamic Selection of Protocols

- Compose protocol stack based on user needs and underlying network capabilities



Acknowledgments

- John Bresnahan
- Mike Link
- Ravi Madduri
- Martin Feller
- Gaurav Khanna
- Liu Wantao

Questions