The Analysis & Mining of Globally Distributed Data

Chapter 5. Data Grids

Robert Grossman
Laboratory for Advanced Computing
University of Illinois at Chicago
&
Open Data Partners

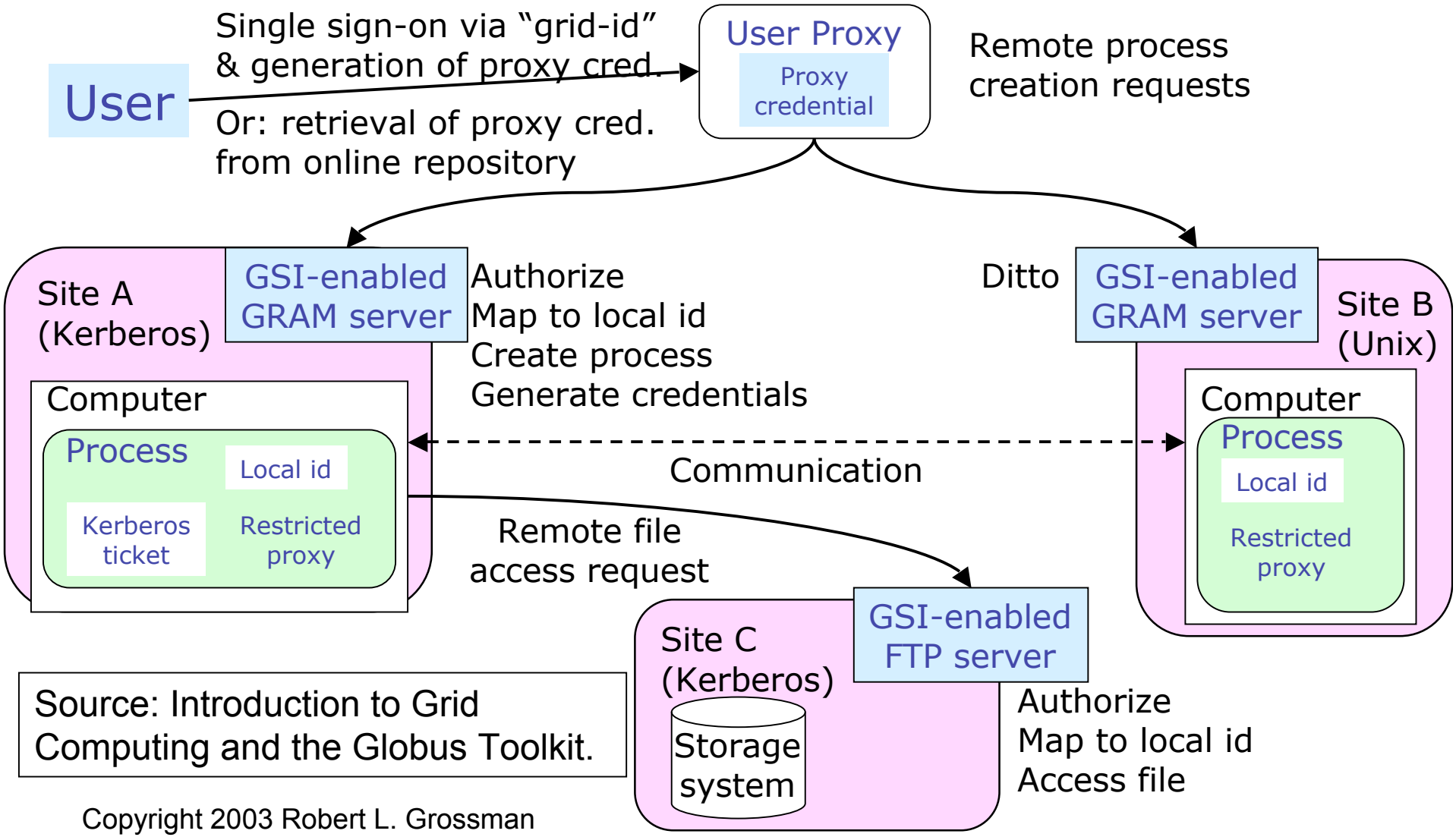
5.1 Data Grids

Data meets Globus Security Infrastructure (GSI).

Four Key Globus Protocols for Computational Grids

- Problem: It is very difficult to use other people's computational resources today.
- *Security*: Grid Security Infrastructure (GSI)
- *Resource Management*: Grid Resource Allocation Management (GRAM)
- *Information Services*: Grid Resource Information Protocol (GRIP)
- *Data Transfer*: Grid FTP (GridFTP)

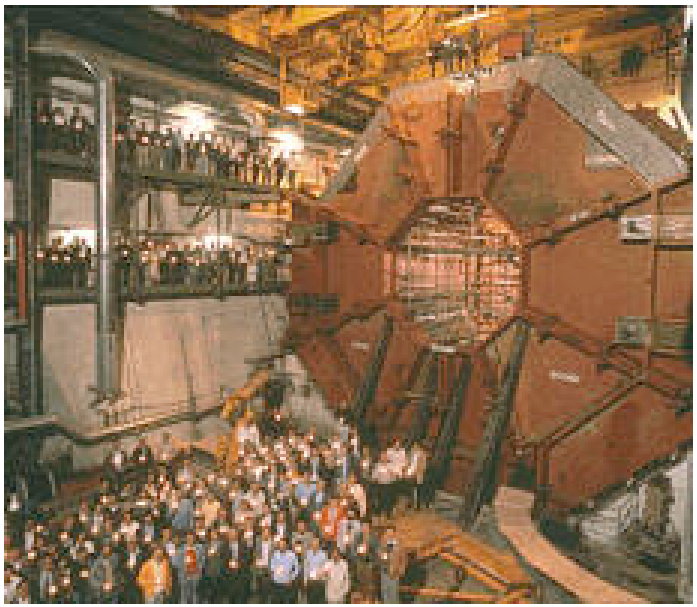
Example GSI: "Create Processes at A and B that Communicate & Access Files at C"

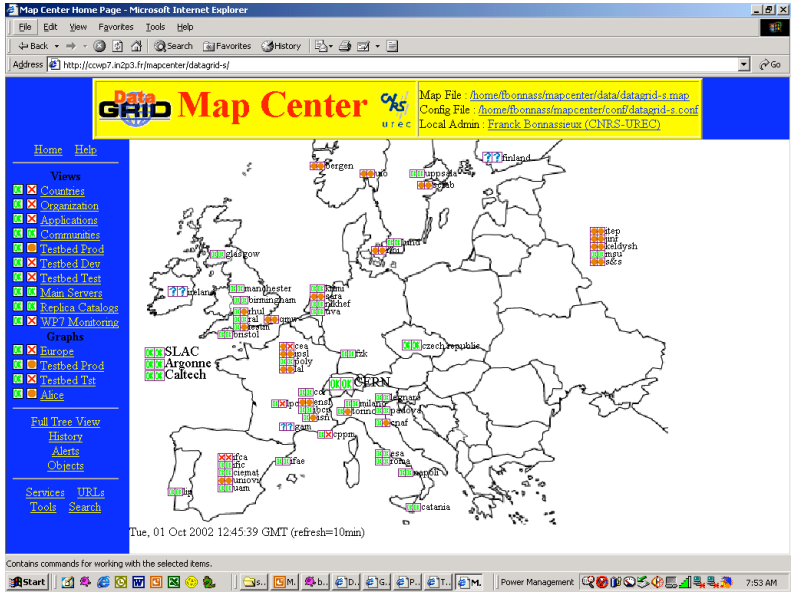


Source: Introduction to Grid Computing and the Globus Toolkit.

Data Grids – In a Word

- Data and users must be authenticated before accessing resources.





Philosophy

- Robust and flexible authentication, integrity and confidentiality features are critical when transferring or accessing files. GridFTP must support GSI and Kerberos authentication, with user controlled settings of various levels of data integrity and/or confidentiality.

Allcock et. al., Protocols and Services for Data Intensive Sciences,

Three Essential Data Grid Services

1. Grid Security Infrastructure (GSI)
2. GridFTP
   - supports parallel TCP
   - GridFTP must support GSI and Kerberos
3. Globus Replica Managment

Security Terminology

- Authentication: Establishing identity
- Authorization: Establishing rights
- Message protection (integrity & confidentiality)
- Non-repudiation
- Digital signature
- Accounting
- Certificate Authority (CA)

Grid Security Infrastructure (GSI)

- Based upon extensions to standard protocols & APIs
  - Standards: SSL/TLS, X.509 & CA, GSS-API
  - Extensions for single sign-on and delegation
- Globus Toolkit includes a reference implementation of GSI
- Deployed by NSF Teragrid, NASA Information Power Grid, DOE Science Grid, …

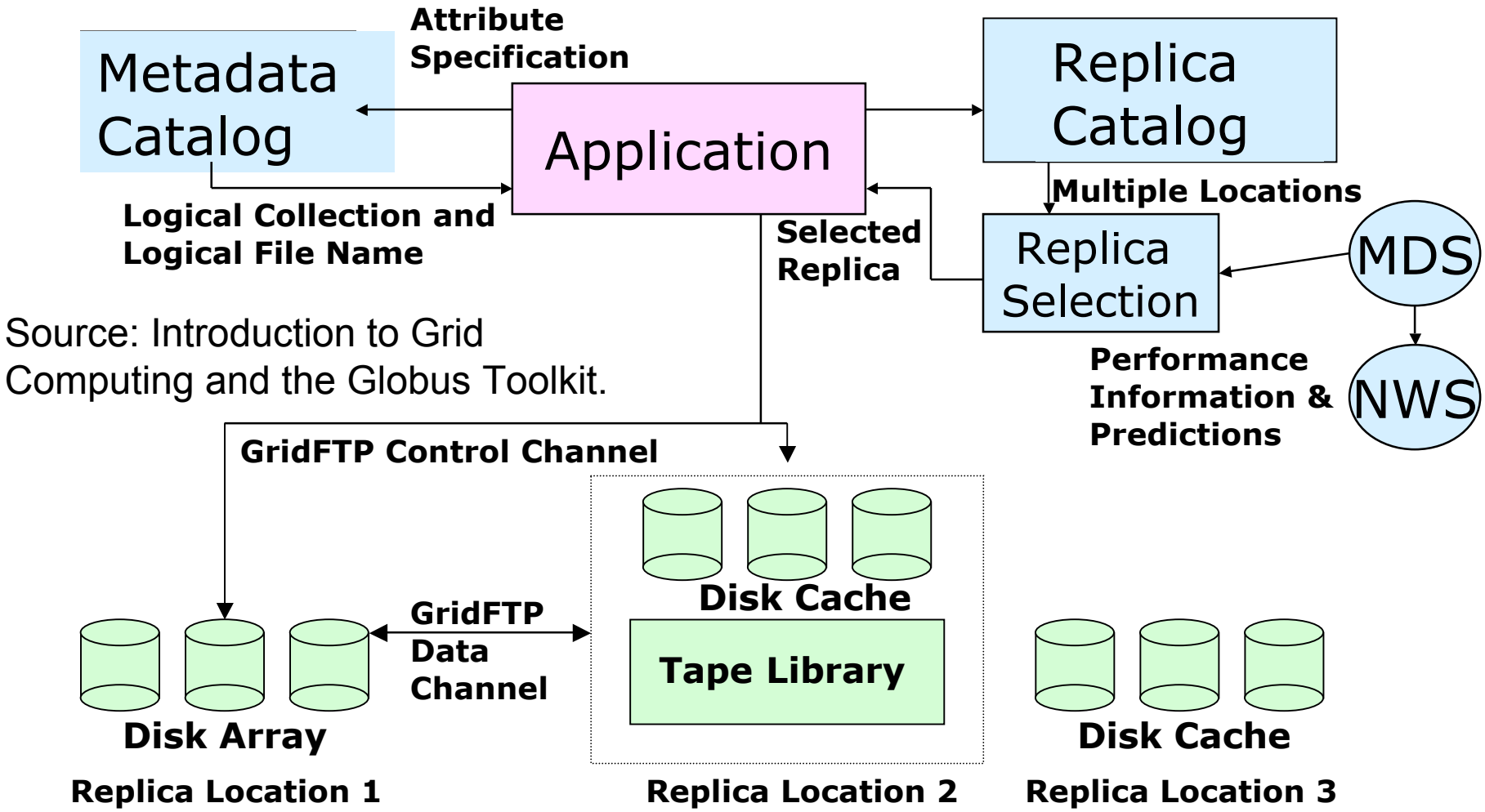
Globus GridFTP

- FTP is defined by several IETF RFCs
- GridFTP includes get/put, 3rd-party transfer as usual
- GridFTP extend FTP to include:
  - striped/parallel data channels
  - automatic & manual TCP buffer setting
  - progress monitoring,
  - restart

Globus Replica Managment

- Creating new copies of complete or partial data set
- Registering new copies in a replica catalog
- Supporting querying the catalog so users can identify all copies of a file or collection of files
- Selecting the “best” replica for access based on storage and network performance predictions

A Model Architecture for Data Grids



Source: Introduction to Grid Computing and the Globus Toolkit.

Layered Grid Architecture

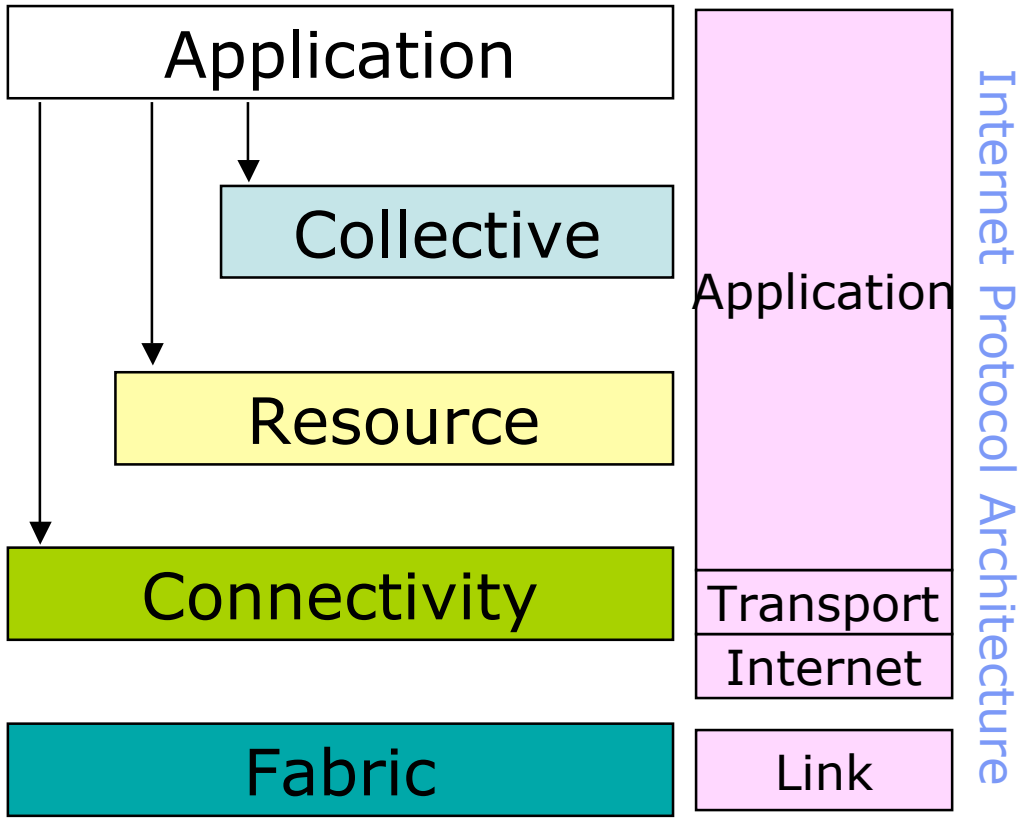
Source: Introduction to Grid Computing and the Globus Toolkit.

"Coordinating multiple resources": ubiquitous infrastructure services, app-specific distributed services

"Sharing single resources": negotiating access, controlling use

"Talking to things": communication (Internet protocols) & security

"Controlling things locally": Access to, & control of, resources



Resource Layer Protocols & Services

- Grid Resource Allocation Mgmt (GRAM)
  - Remote allocation, reservation, monitoring, control of compute resources
- GridFTP protocol (FTP extensions)
  - High-performance data access & transport
- Grid Resource Information Service (GRIS)
  - Access to structure & state information
- Network reservation, monitoring, control

Collective Layer Protocols & Services

- Index servers aka metadirectory services
  - Custom views on dynamic resource collections assembled by a community
- Resource brokers (e.g., Condor Matchmaker)
  - Resource discovery and allocation
- Replica catalogs
- Replication services
- Co-reservation and co-allocation services
- Workflow management services, etc.