

Data Management Technologies in Grids

Borja Bergua Guerra
Computer Architecture Group
Universidad Carlos III de Madrid



Topics

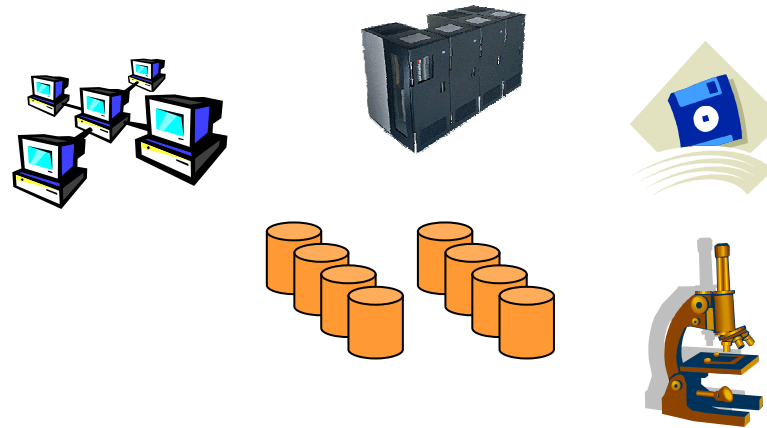


1. Motivation and some definitions
2. Data Management in Globus
3. Expand solution for cluster and grid computing

Motivation of grid computing

- Many resources scattered across multiple organizations

- CPU
- Networks
- Storage
- Data



- Goals:

- To take advantage of these resources.
- Sharing geographically distributed computation and storage resources.
- Users can share data and resources reliably and consistently.

Some definitions



- A **grid system** is a collection of distributed resources connected by a network, located at **different administrative domains**, accessible to users and applications in order to share resources and increase the performance.
- A **grid application** is an application that operates in a grid environment.
- A **grid middleware** is the software that facilitates writing grid applications and manages the grid infrastructure.

Other Definitions

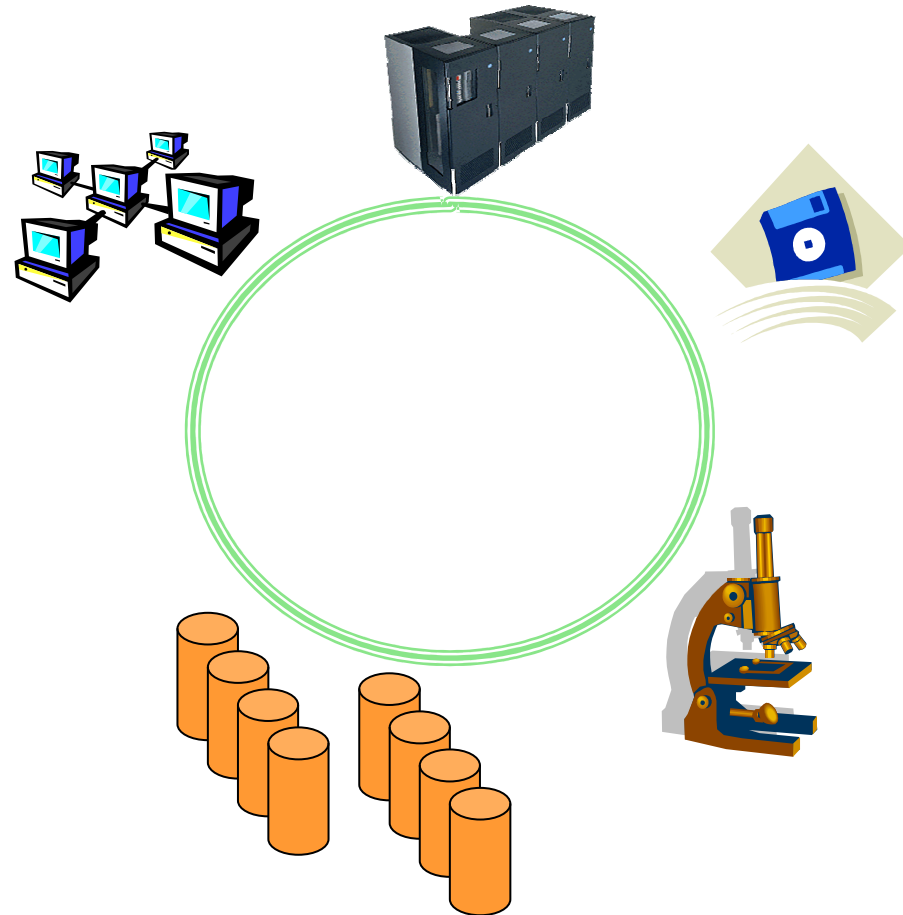


- The Globus Project defines Grid as:
 - Infrastructure that enables the integrated, collaborative use of high-end computers, networks, databases, and scientific instruments owned and managed by **multiple organizations**
- The Gridbus project defines Grid as:
 - Type of parallel and distributed system that enables the sharing, selection, and aggregation of **geographically distributed autonomous** resources dynamically at runtime depending on their availability, capability, performance, cost and users' quality-of-service requirements.

Resources



- Processors
- Storage
- Networks
- Software
- Data
- Special Devices



Grid Architectures



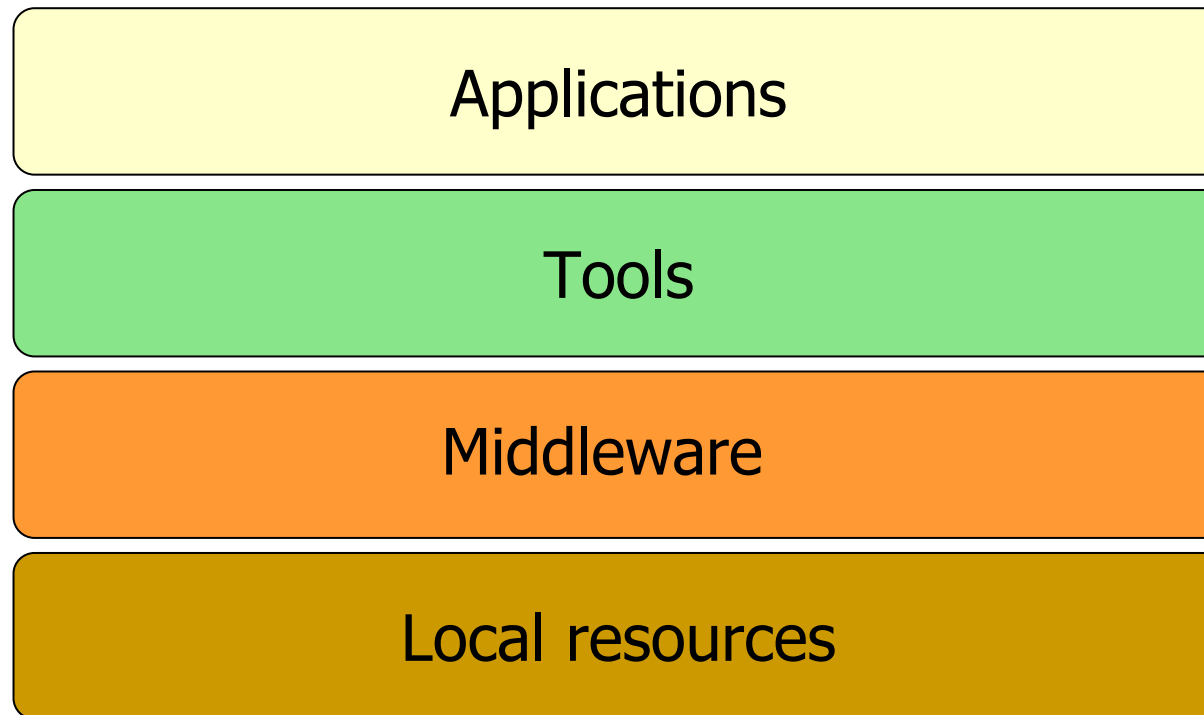
- Virtual organizations (VO)
 - Different physical organizations that share resources and collaborate in order to achieve a common goal
 - A VO defines the resources available and the rules for access
- Economic grid
 - Based on economic principles
 - Resources providers (owners) compete to provide the best service to resource consumers (users) who select appropriate resources based on their specific requirements



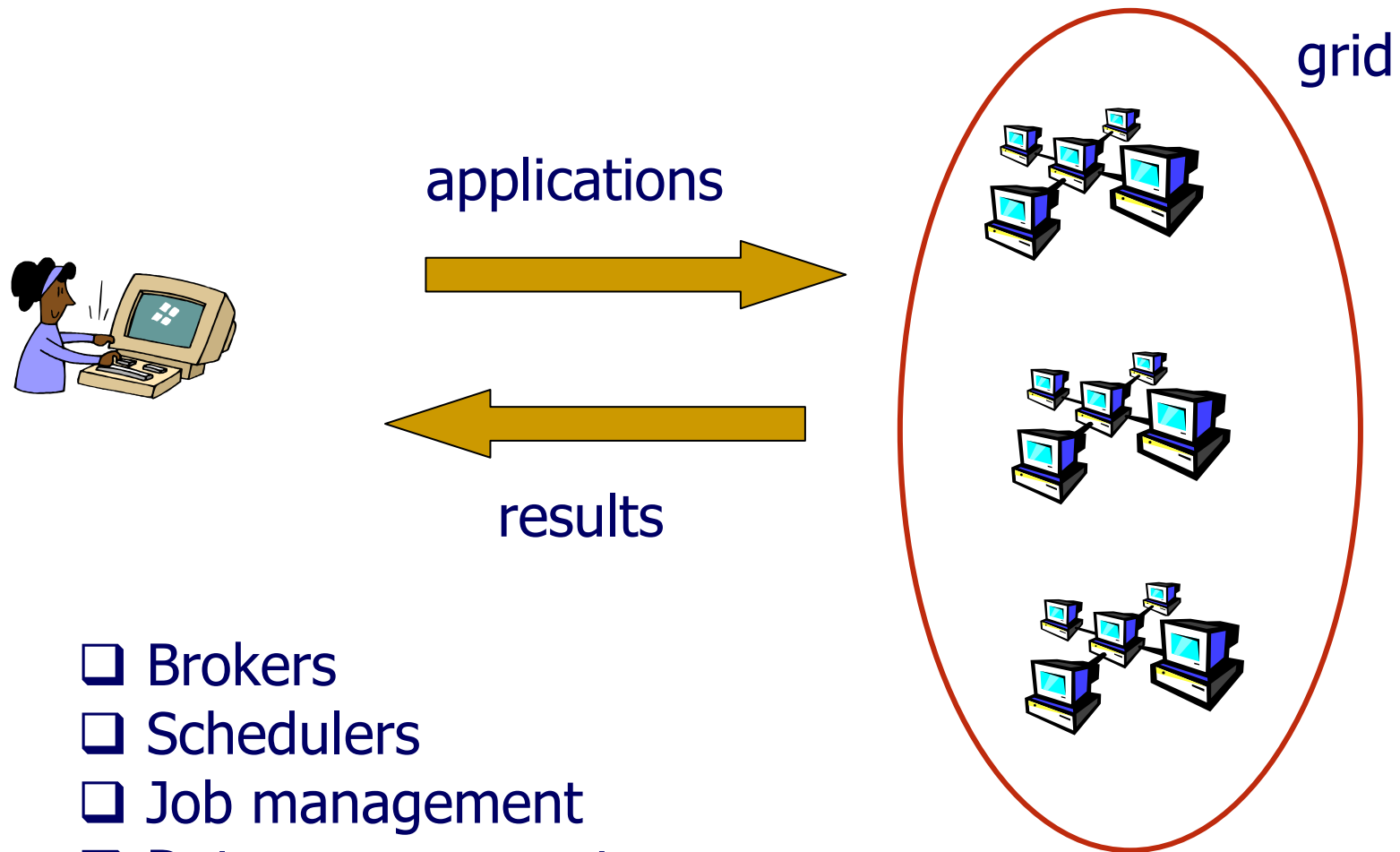
Main grid uses

- Distributed supercomputing
- High throughput computing
- On-demand computing
- Data-intensive computing
- Collaborative computing
- Multimedia computing

Grid components

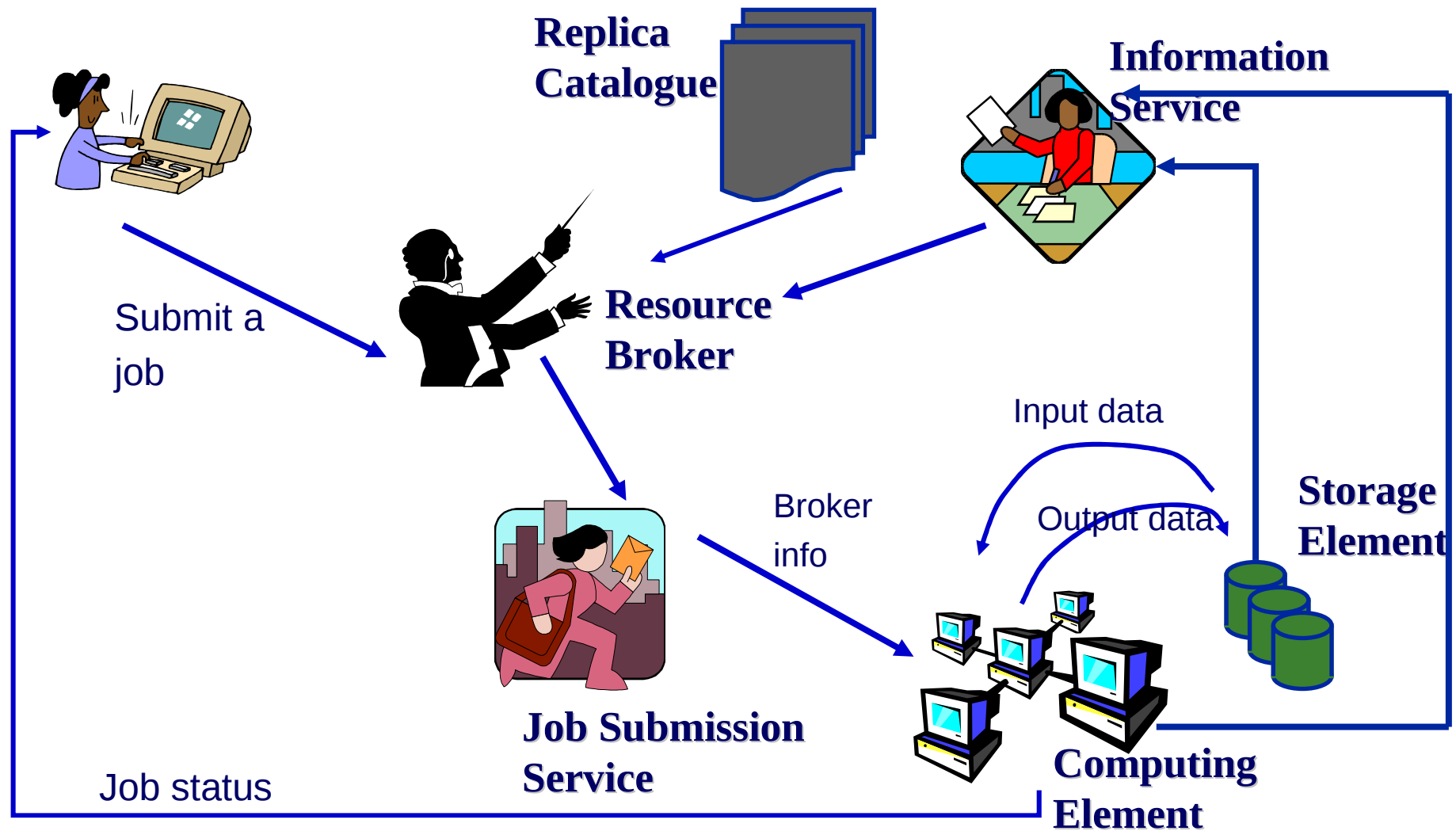


User point of view



- ☐ Brokers
- ☐ Schedulers
- ☐ Job management
- ☐ Data management
- ☐ Security model

Jobs execution model



Applications appropriate for grids



- *Perfect parallelism.* An application can be divided into sets of processes that require little or no communication.
- *Data Parallelism.* The same operation is performed on many data elements simultaneously.
- *Functional Parallelism.* Multiple operations are performed simultaneously, with each operation addressing a particular part of the problem.

Applications not appropriate for grids



- Parallel applications with many interprocess communication
- Transactions
- Applications with many interdependencies between jobs
- Applications with non standard network protocols



What need applications?

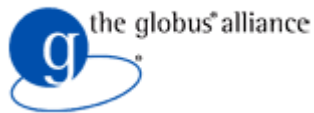
- Applications need **programs** that process **input data** and to produce **output data** in a reliable and efficient way
 - Executables transfer
 - Input data transfer
 - Output data gathering
 - *Checkpoints*
- Applications need **transparent access** to its input and output data is required
- A grid must to provide **transparent, secure, high-performance** access to data located in different administrative domains and organizations
 - File location services
 - File transfer services

Data management techniques



- Distributed File systems
 - NFS, AFS,...
- Storage technologies
 - NAS Devices
 - *Storage Area Networks (SAN)*
 - iSCSI Devices
- Parallel file systems
 - PVFS, GPFS, Expand,...
- Distributed Storage (wide-area)
 - File transfers
 - Replication
 - Grid file systems

Data Management in Globus



Globus Toolkit



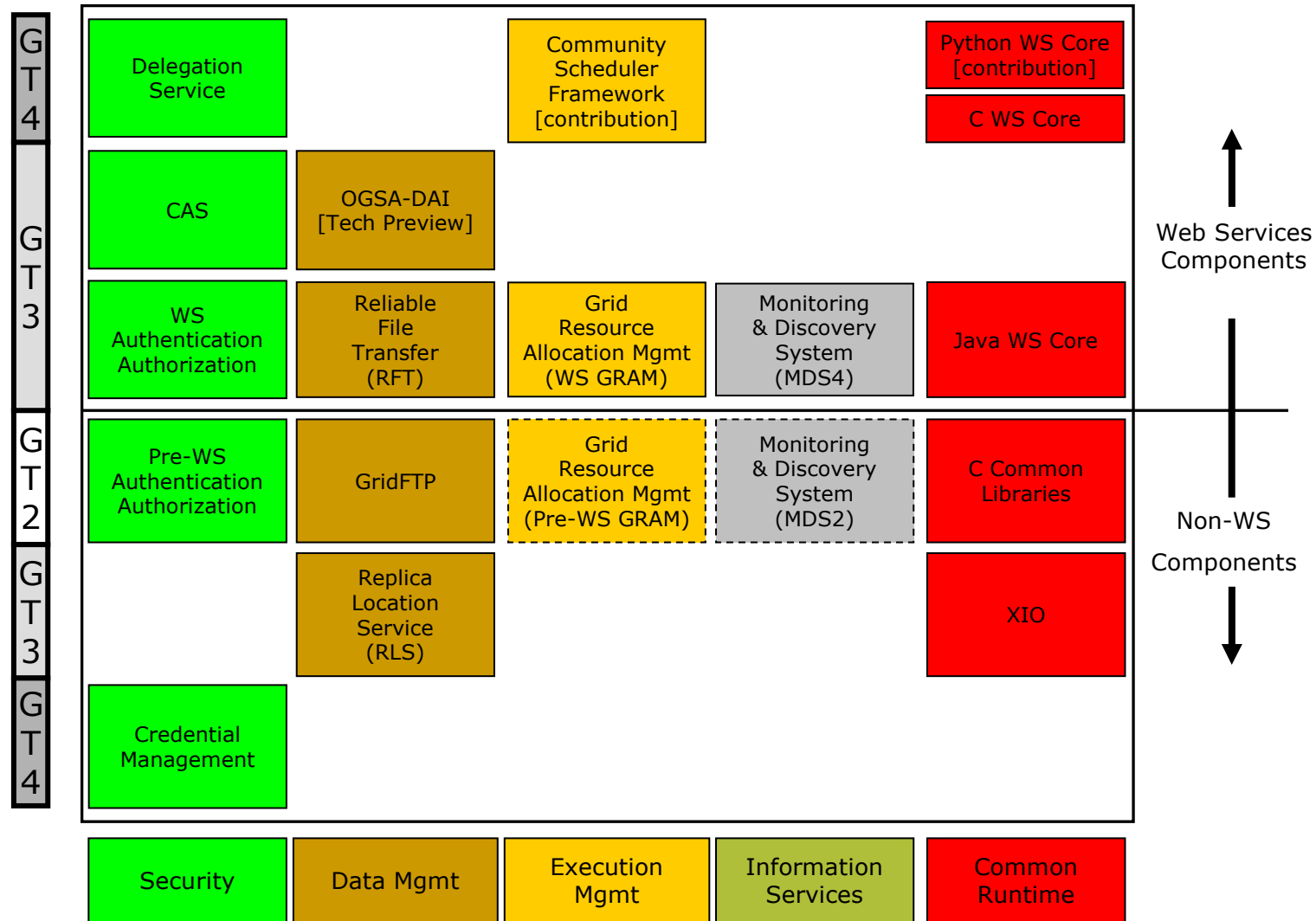
- Software toolkit, developed by The Globus Alliance
<http://www.globus.org>
- It can use to program grid-based applications.
- Includes *high-level services* that we can use to build Grid applications. It includes:
 - Resource monitoring and discovery service
 - A job submission infrastructure
 - A security infrastructure
 - Data management services

What is not the Globus toolkit?



- A user tool
- An application
- A scheduler

Globus Toolkit Components



Security

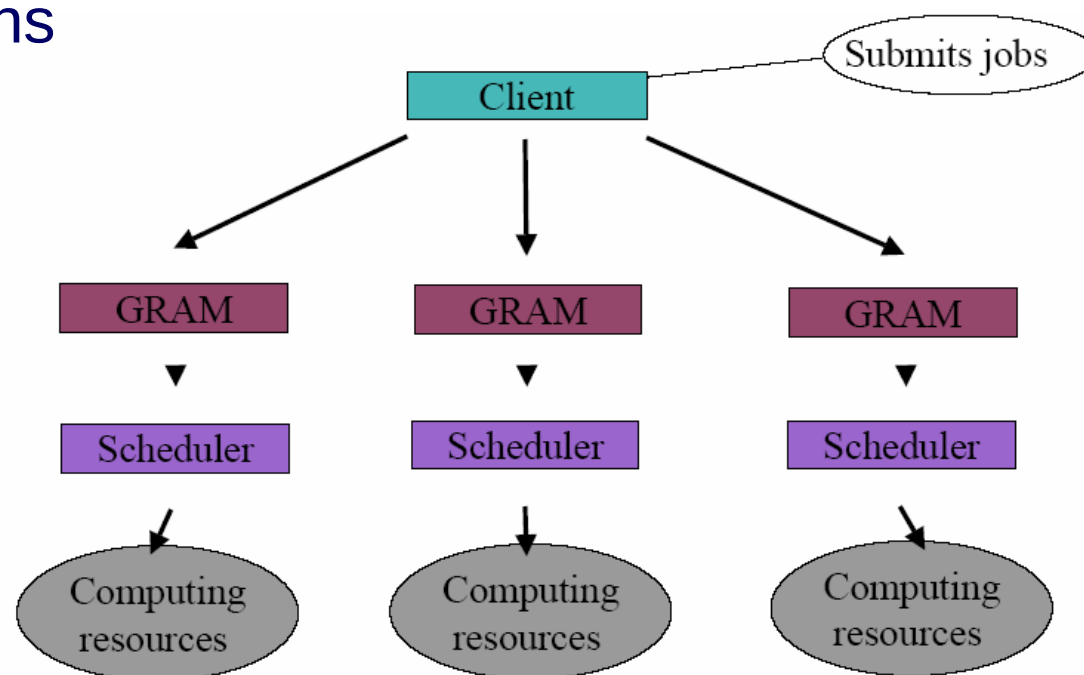


- Globus Toolkit proposed and implements the *Grid Security Infrastructure* (GSI)
 - Single sign-on for using grid services through user certificates
 - Resource authentication through host certificates
 - Data encryption
 - Authorization

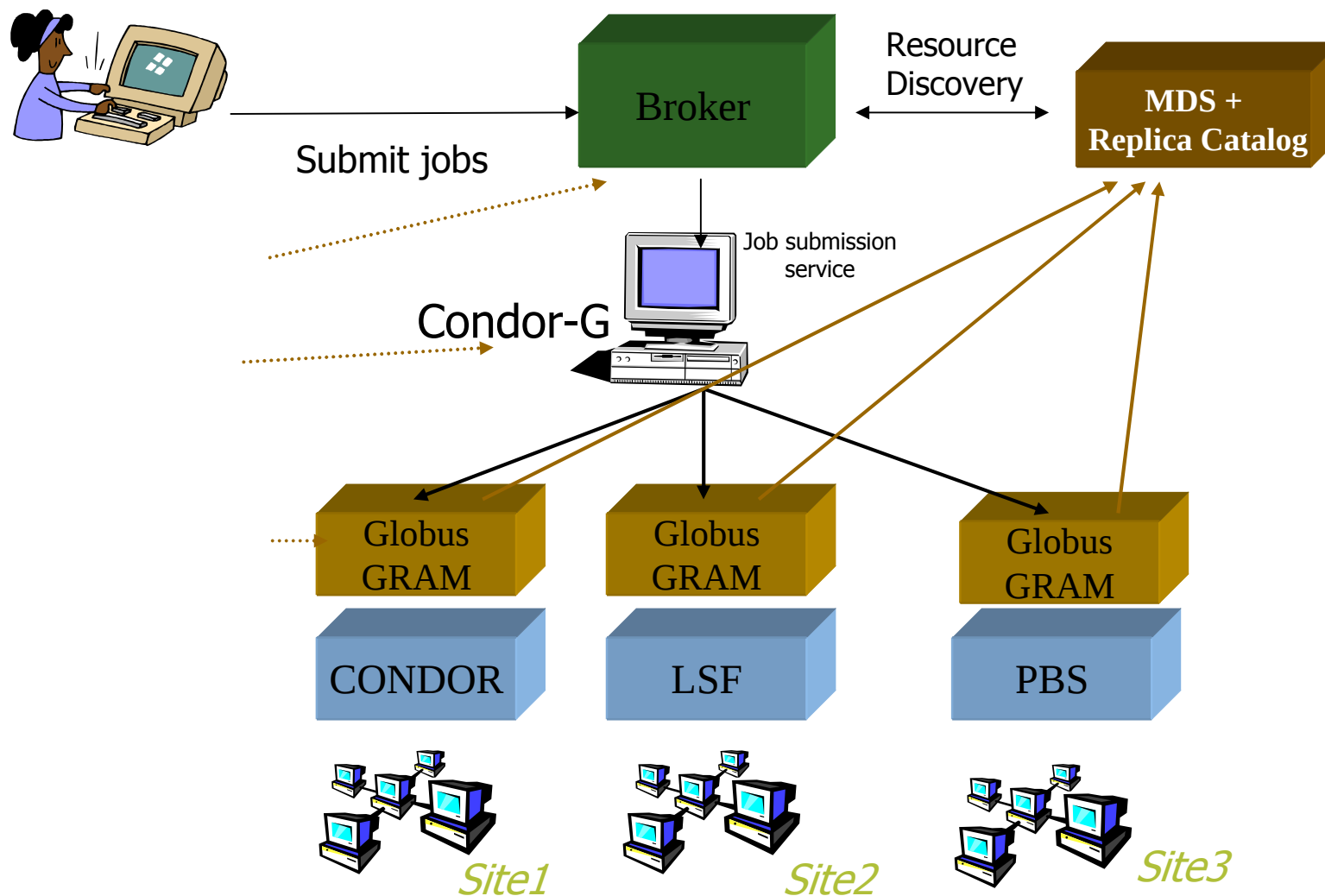


Execution management

- Grid Resource Allocation and Management (GRAM)
- GRAM allows to *submit, monitor* and *cancel* job.
- GRAM is not a scheduler.
- Two implementations
 - WS GRAM
 - Pre-WS GRAM



MetaSchedulers



Monitoring and Discovery System (MDS)



- Suite of web services for *monitoring* and *discovering* resources and services on Grids
- Examples of useful information
 - Characteristics of a compute resource
 - Software available, networks connected to, load, type of CPU, disk space
 - Characteristics of the Globus infrastructure
 - Hosts, resource managers, service availability

Data Management in Globus



- Data transfer
 - GridFTP
 - Reliable File Transfer (RFT) service

- Data Replication
 - Replica Location Service (RLS)

GridFTP



- GridFTP is a *high-performance, secure, reliable* data transfer protocol optimized for high-bandwidth wide-area networks
- Based on FTP protocol
 - GridFTP: Protocol Extensions to FTP for the Grid
 - Global Grid Forum Recommendation
 - <http://www.ggf.org/documents/GWD-R/GFD-R.020.pdf>
- Globus provides
 - Server implementation
 - Client tools (command line programs)
 - Development libraries

globus-url-copy



- Command line tool that can do multi-protocol data movement
- Mainly used for GridFTP, however, it supports many protocols
 - gsiftp:// (GridFTP)
 - ftp://
 - http://
 - https://
 - file://
- You must have a certificate to use it

globus-url-copy



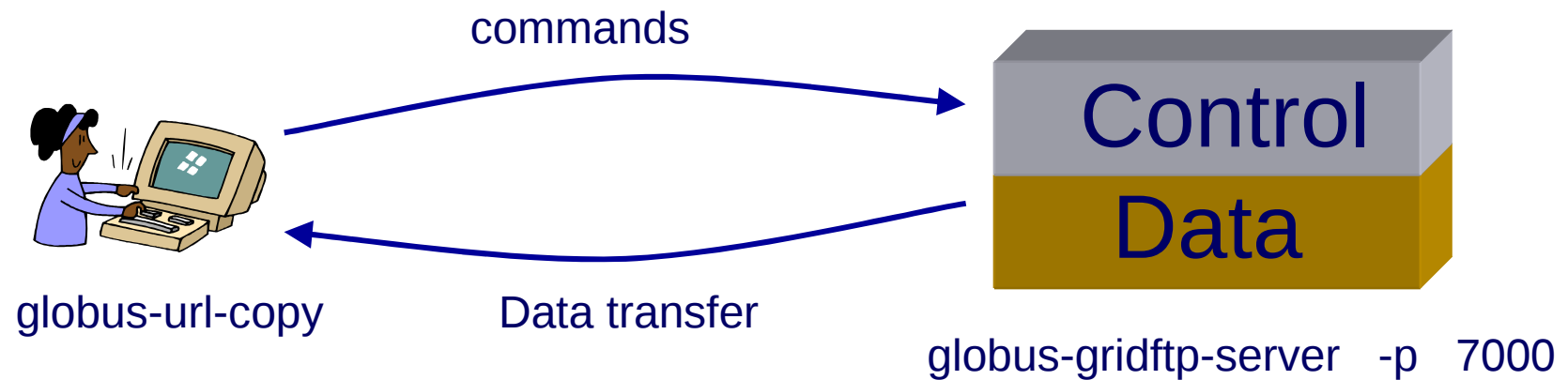
- Command line
 - `globus-url-copy [options] srcURL dstURL`
- Examples of URL
 - `gsiftp://host.domain.com:2890/dir/file.dat`
 - `http://host.domain.com/webpage/page.html`
 - `file:///localdirectory/file.dat`
- For GridFTP (`gsiftp://`) and FTP (`ftp://`), it is legal to specify a user name and password
 - `gsiftp://name:[password]@host.domain.com/foo.dat`

Server Architecture

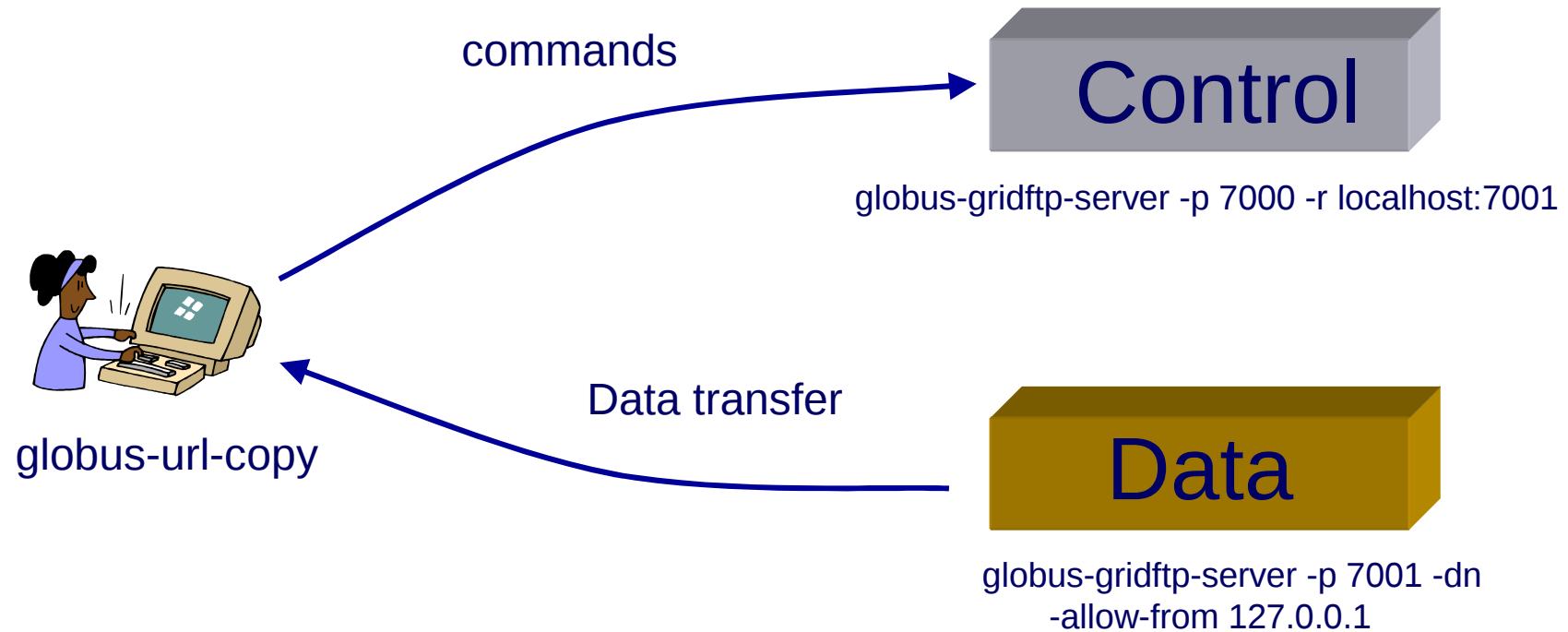


- GridFTP use two separate socket connections:
 - A control channel for commands and responses
 - A data channel for data transfer
- Control and data channel can be in separate process
- A single control channel can have multiple data channels
 - Used in striped operation

Same Server for control and data



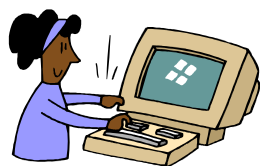
Different servers for control and data



- High level of security
 - The frontend can be run as any user
 - The backend is run as root, but configured to only allow connections from the frontend



Third party transfers

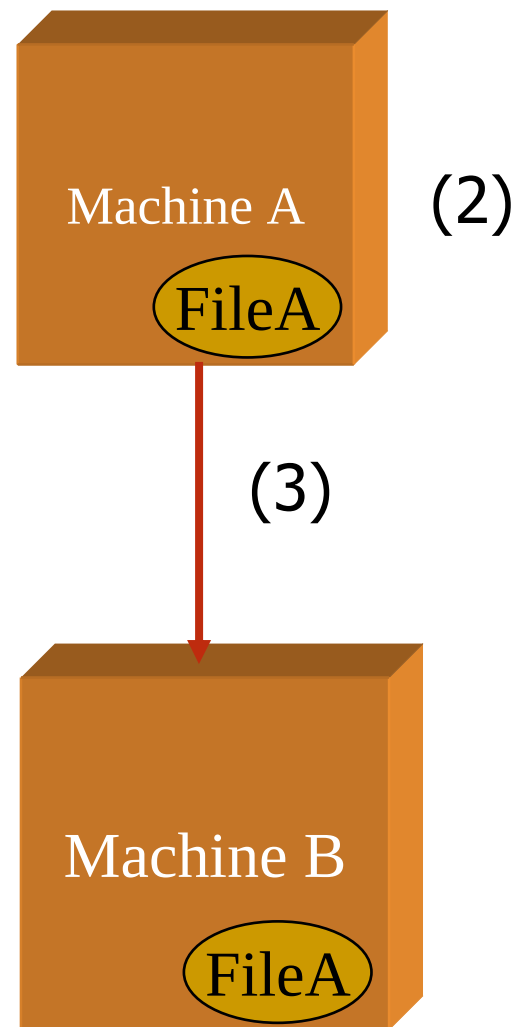


(1)

`globus-url-copy`

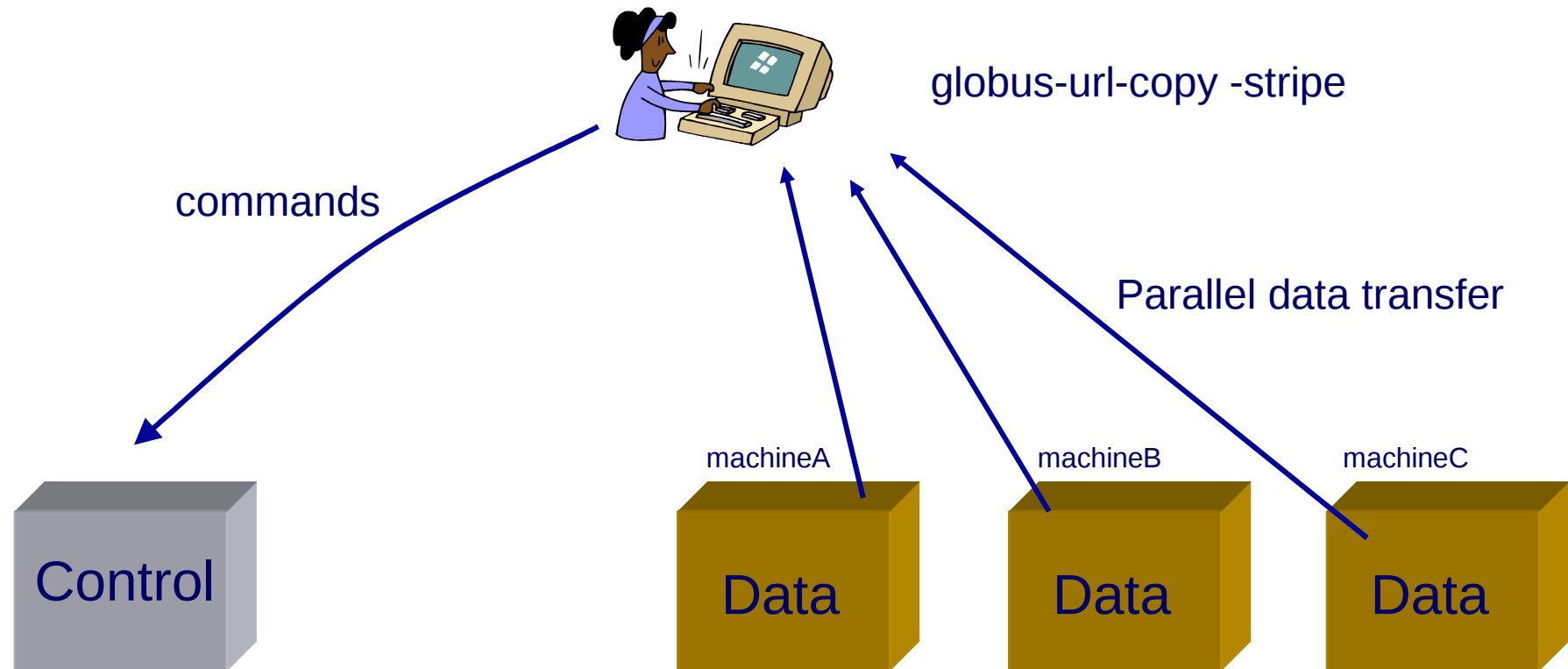
`gsiftp://machineA/tmp/FileA`

`gsiftp://machineB/tmp/FileB`



- (1) Client send data requests to machine A
- (2) Machine A init the transfer and client it disconnected
- (3) Machine B receives the file

Striped servers



```
globus-gridftp-server -p 7000  
-r machineA:5000,machineB:6000,machineC:4000
```

```
machineA> globus-gridftp-server -p 5000 -dn
```

```
machineB> globus-gridftp-server -p 6000 -dn
```

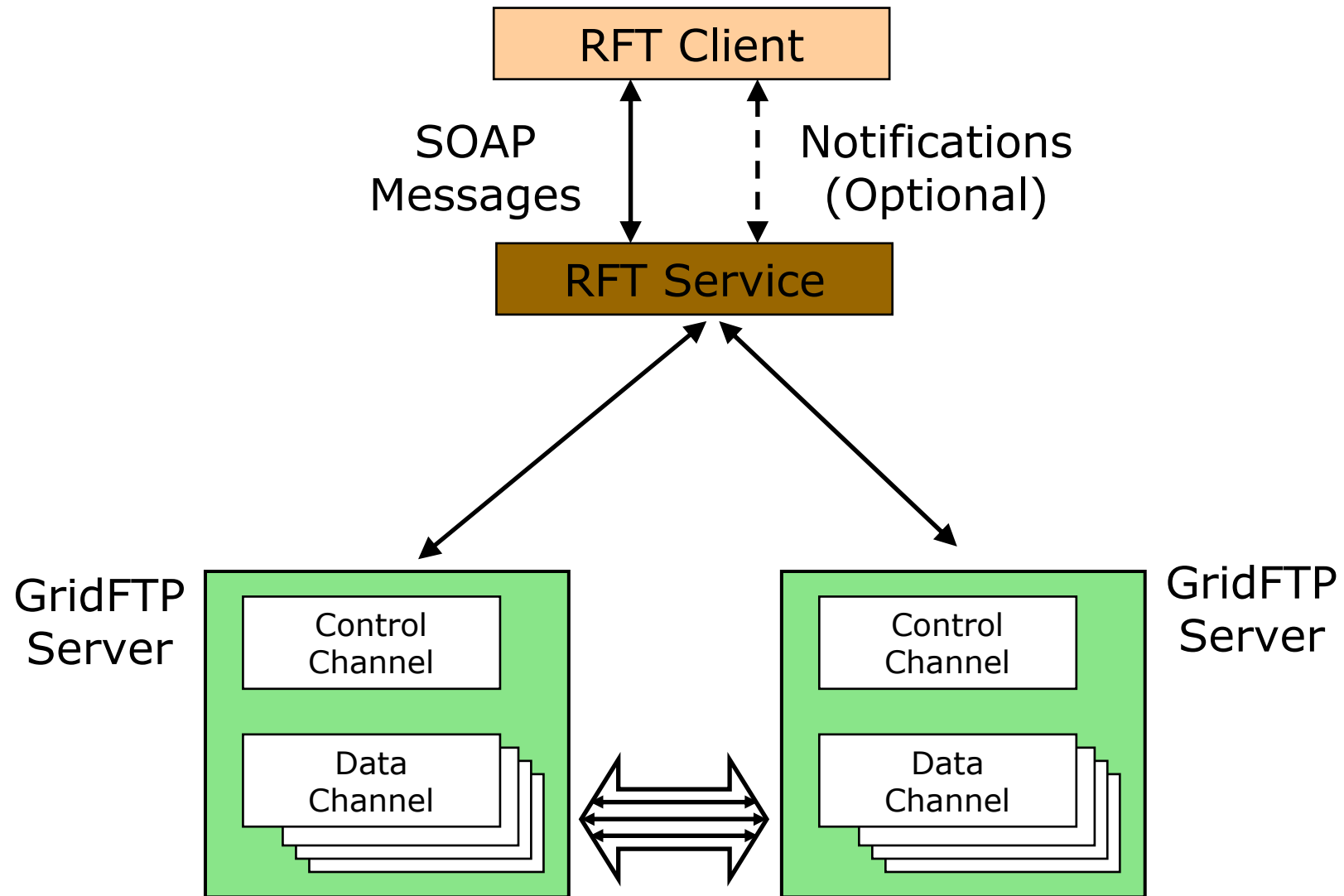
```
machineC> globus-gridftp-server -p 4000 -dn
```

Reliable File Transfer



- Uses standard SOAP messages over HTTP to submit and manage transfers
- Increase reliability transfer because state is stored in a database.
- The client can submit the transfer request and then disconnect and go away.

RFT: Third party transfer



Replication Location and management



- This component supports multiple locations for the same file across the grid
- Filenames
 - **Logical Filename (LFN)** is the name that refers to the full set of replicas for a file.
 - Logical identifier
 - Based on the LFN, all replicas can be looked up in a replica catalogue.
 - **Physical file name (PFN)** is the location of a copy of the file on a storage system .
 - Host + physical location
 - May be a *physical filename* of a file stored on disk

Filenames examples

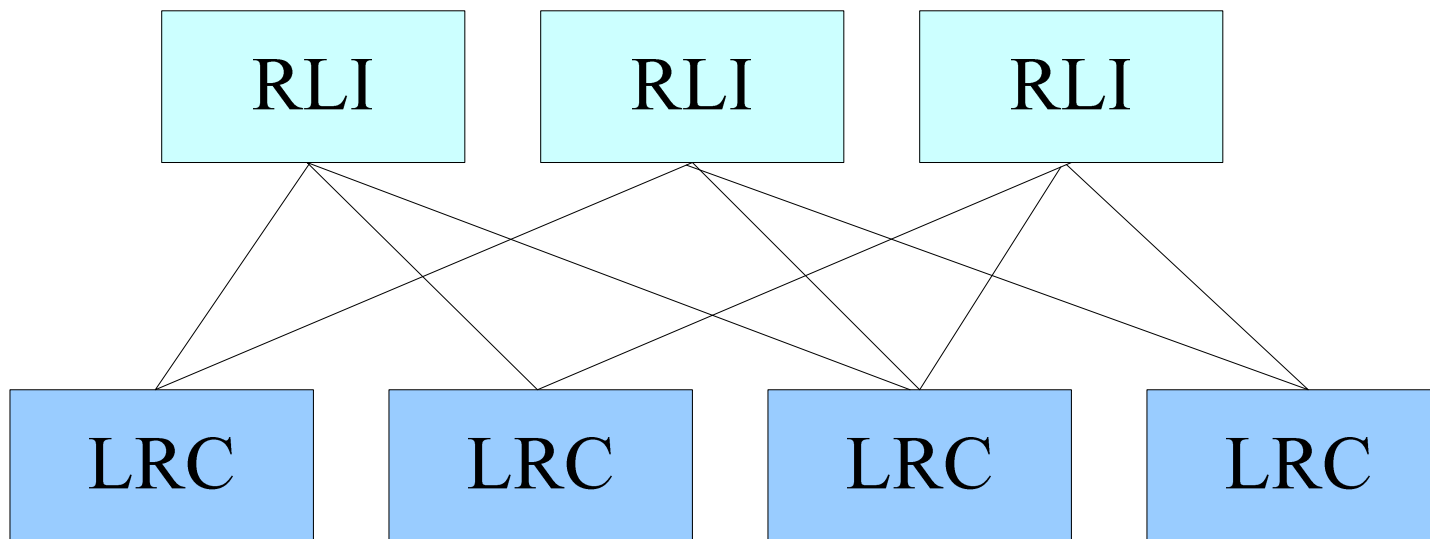


Logical File Name	My-data logical-name
Physical File name	gridftp://host1.domain/data/myfile gsiftp://host2.domain/data/myfile http://example.org/data/filename

RLS components



Replica Location Index Nodes



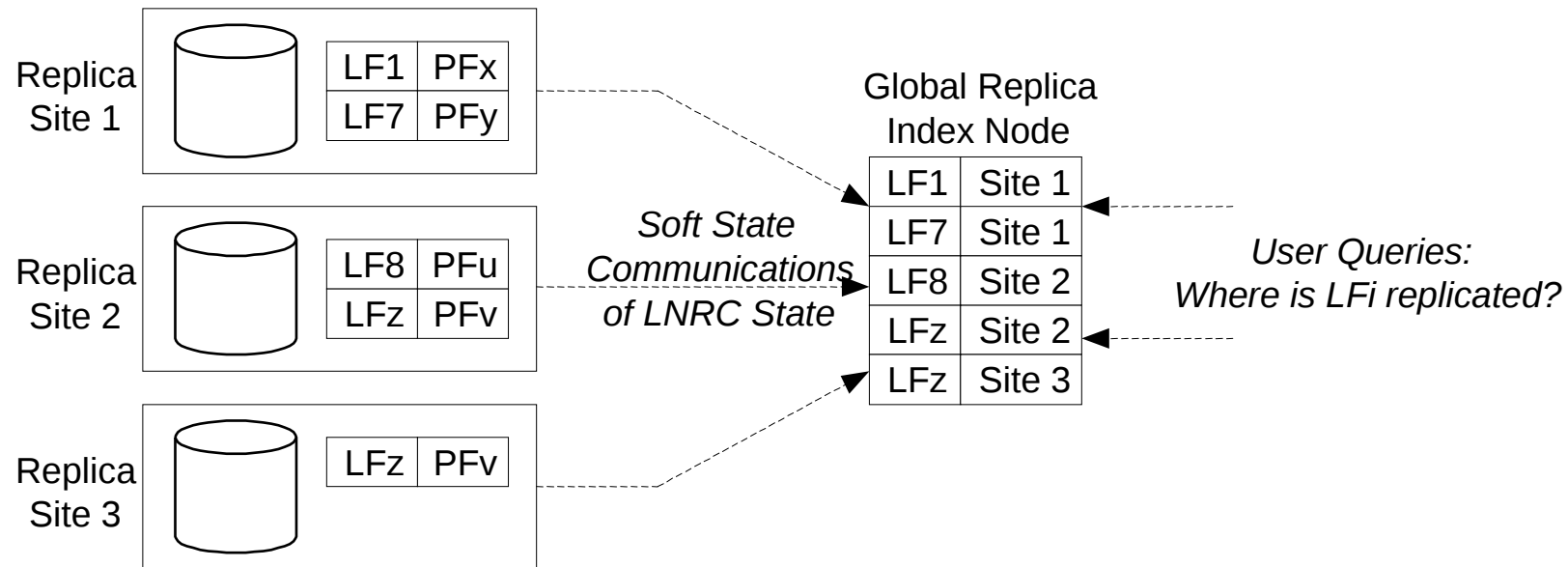
Local Replica Catalogs

RLS components



- **Local Replica Catalogs** (LRCs) stores mappings between logical names for data items and the physical locations of replicas of those items.
- **Replica Location Index** (RLI) nodes aggregate information about one or more LRCs.
- LRCs use soft state update mechanisms to inform RLIs about their state: relaxed consistency of index.
- Optional compression of state updates reduces communication, CPU and storage overheads
- Membership service registers participating LRCs and RLIs and deals with changes in membership

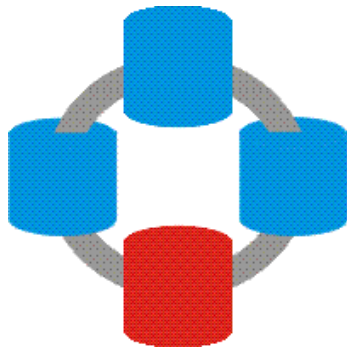
Example 1



Source: Globus Alliance

- Users submit a query (LF) to the RLI for obtaining the list of RLC that store the file
- User selects the appropriate replica site and submit a query (LF) for obtaining the physical name
- Finally the user uses the gridFTP with the PN for transfer the replica

Expand: Parallel File System for cluster and grid computing



Expand

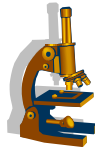


- A new approach to the construction of parallel file systems for cluster and grid computing
- Solution for clusters:
 - Use of standard servers
- Solution for grid:
 - Use of GridFTP and Globus Toolkit

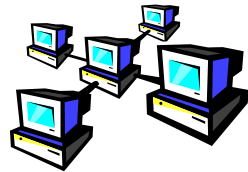
Parallelism in data access



Parallel applications



Parallel computers



Parallel file systems

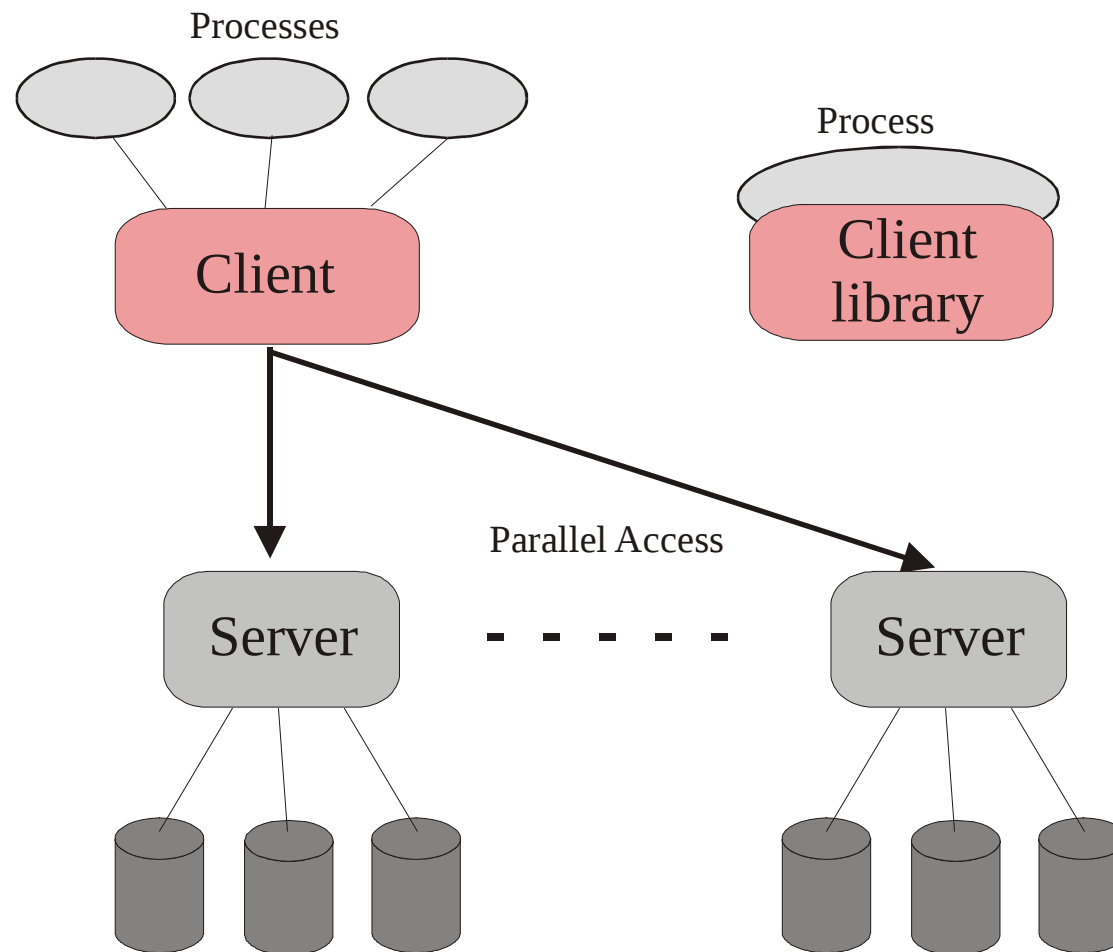


Parallel devices



- Parallelism can be apply in multiple levels
 - To exploit all levels

Typical architecture of a parallel file system





Motivation

- All parallel file systems do not use standard servers
 - Very difficult to use these systems in heterogeneous environments
 - There is no PFS for grid computing

- However:

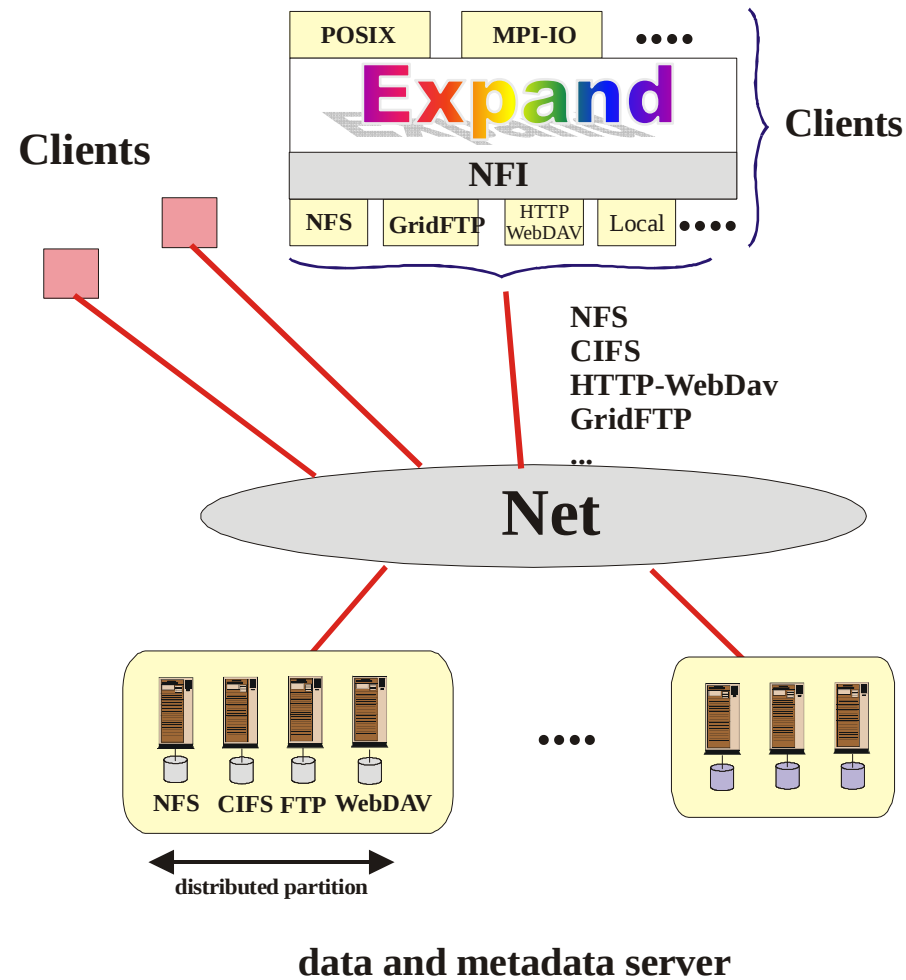
Why use proprietary, special-purpose servers for the parallel file system when you have most of the necessary functionality in many standard servers (NFS, GridFTP) already?

Expand Goals



- To propose a new approach to the construction of parallel file systems for cluster and grid computing
 - Based on existing and standard servers
 - NFS for cluster computing
 - GridFTP for grid computing
- Advantages
 - No changes to servers are required
 - All operations are implemented on the clients
 - Expand is independent of the operating system used in the clients
 - All operations based on sever protocol
 - The parallel file system construction is greatly simplified
 - Allows the using of servers with different architectures and operating systems

Architecture



File Structure

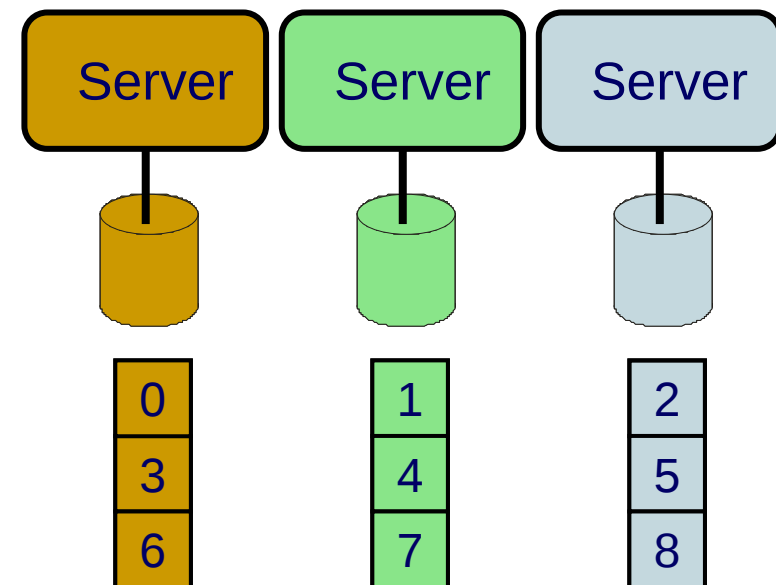


- A file consists of several subfiles, one for each server
- Data are distributed across all servers

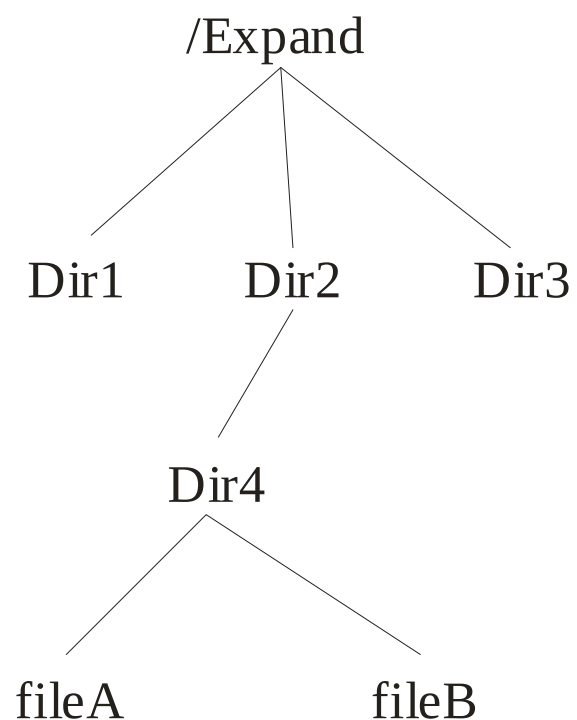
Expand File



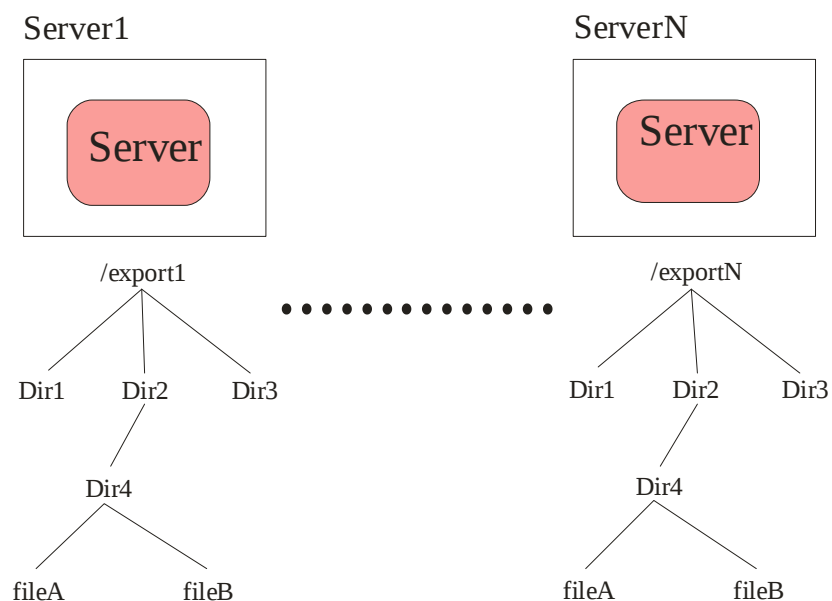
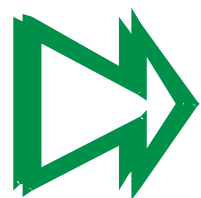
Server File
(subfiles)



Directory Structure

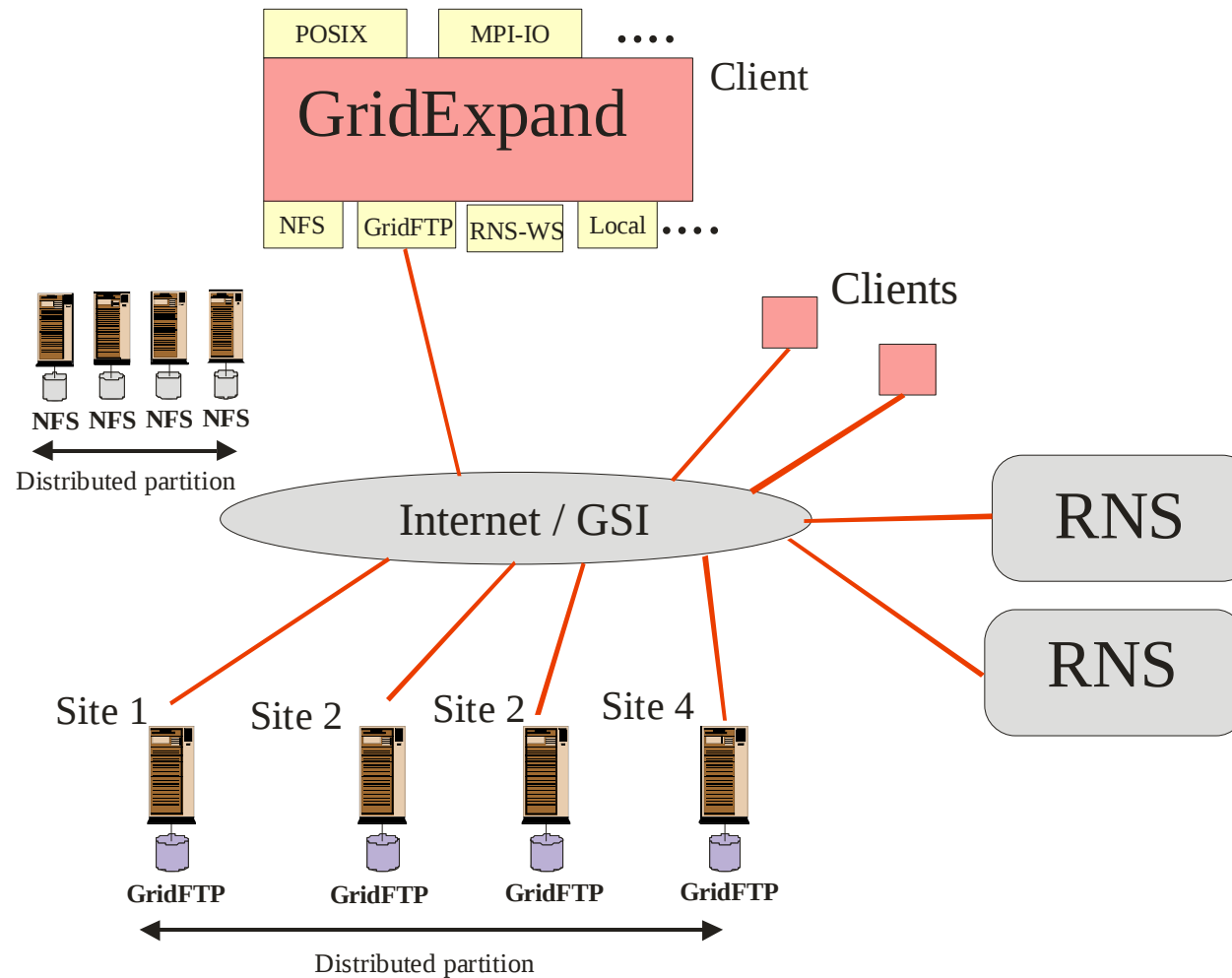


Logical user view



Directory mapping

Expand version for grid computing



Grid file system requirements



- Logical Hierarchical Name Space
 - Resource Namespace service
- Uniform Storage Interface and API
 - POSIX and MPI-IO
- Data Access/Transfer
 - GridFTP
- Security
 - GSI
- Optimization and Performance Improvements
 - Parallel I/O

Performance evaluation

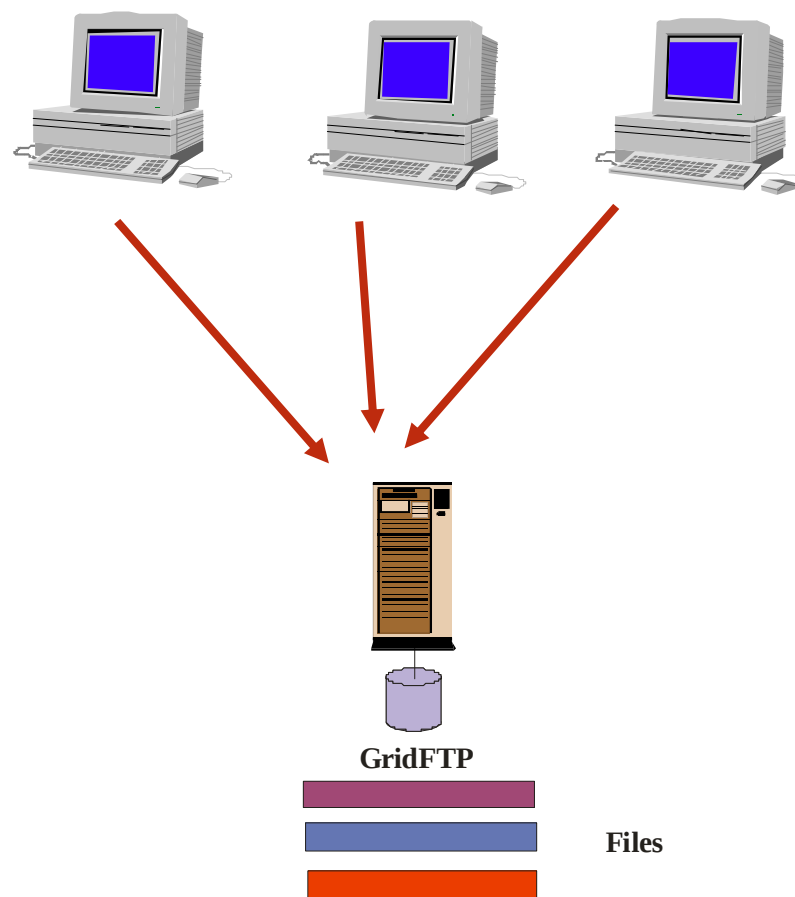


- 500 jobs
- Each job accesses to a random number of files (between 1 and 10 files) chosen among 1000 files.
- The size of each file is 500 MB.

Models evaluated



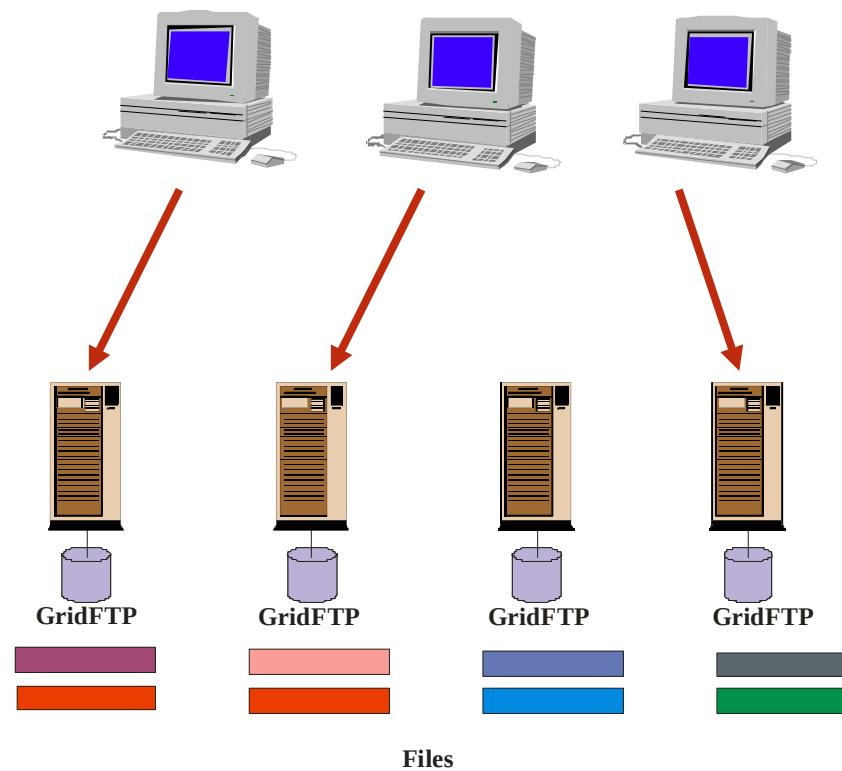
- 1 site
- globus-url-copy



Models evaluated



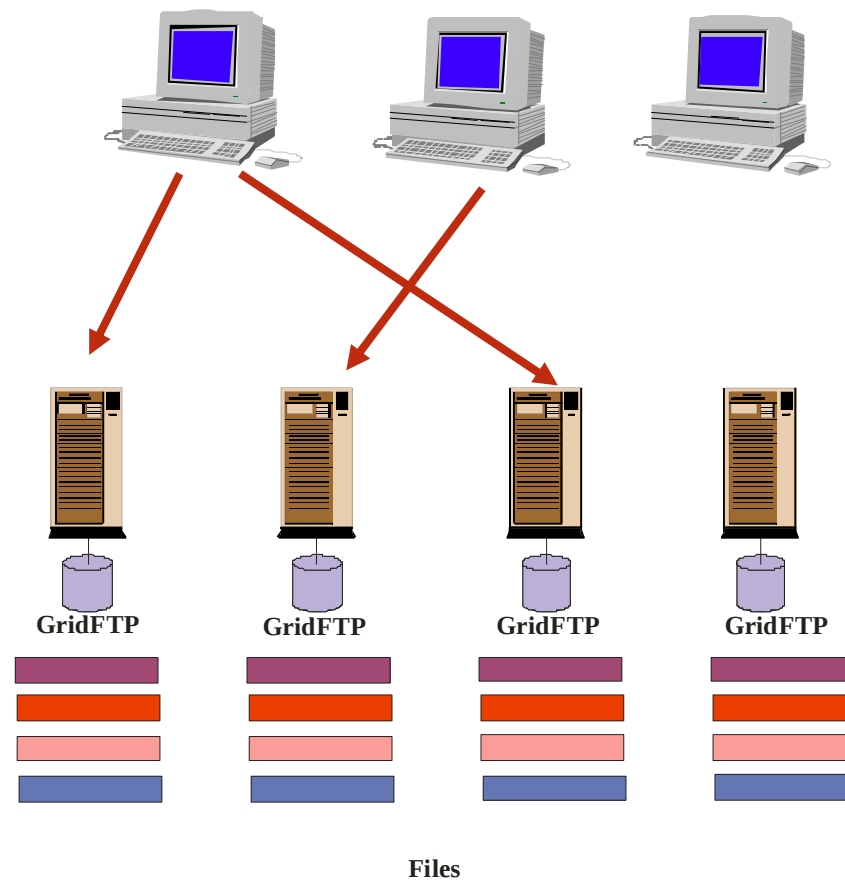
- 4 sites (distributed files, each server store 250 files)
- globus-url-copy



Models evaluated



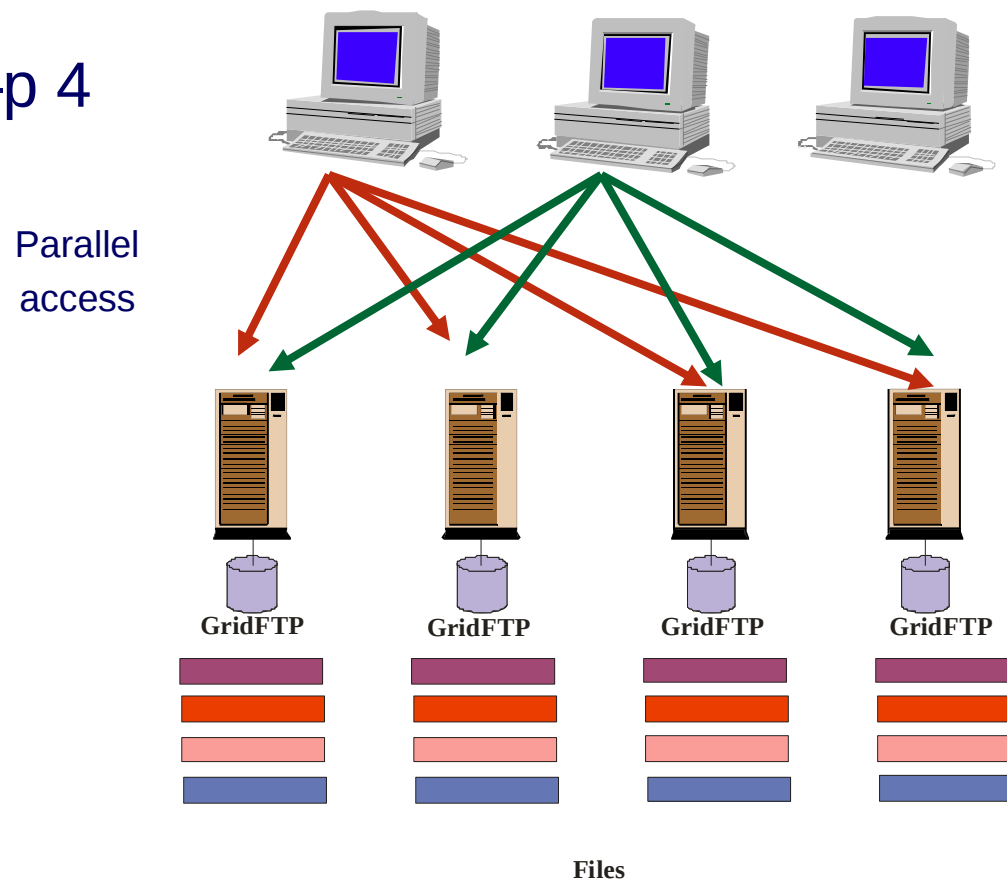
- 4 sites (full replication)
- globus-url-copy



Models evaluated



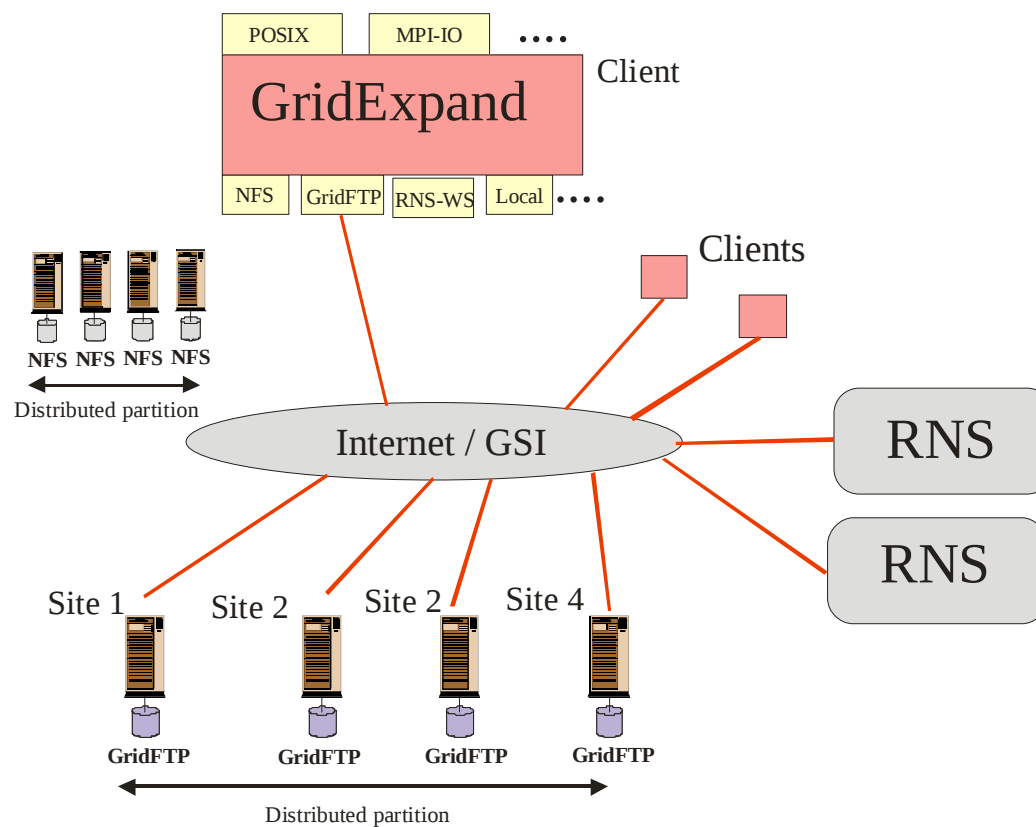
- 4 sites (full replication, parallel access)
- globus-url-copy -p 4



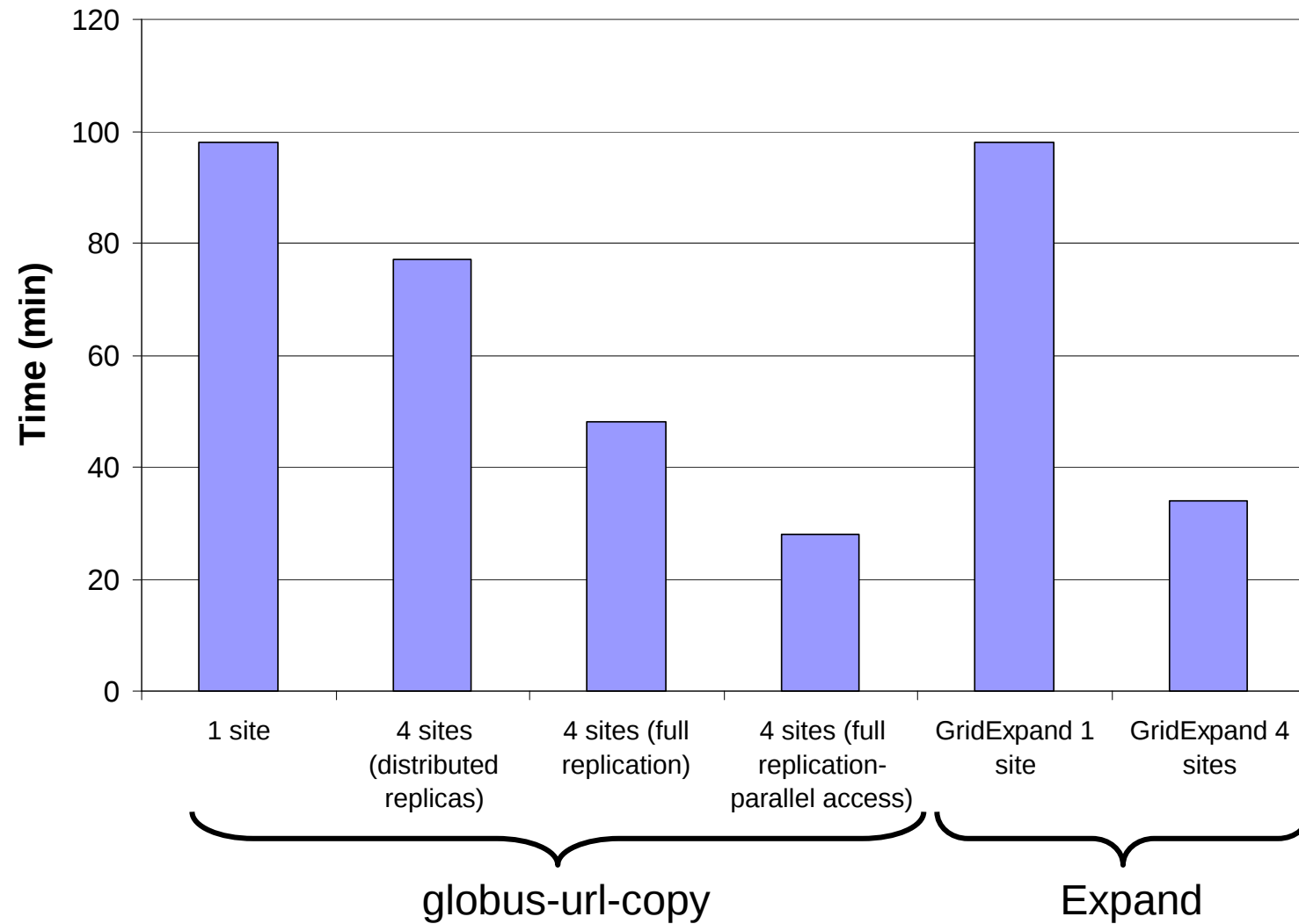


Modeles evaluated

- Grid Expand with 1 and 4 servers



Results



Some Grid File Systems



- Storage Resource Broker (SBR)
- Grid Datafarm
- Distributed StorageTank
- SlashGrid by Andrew