Data Management

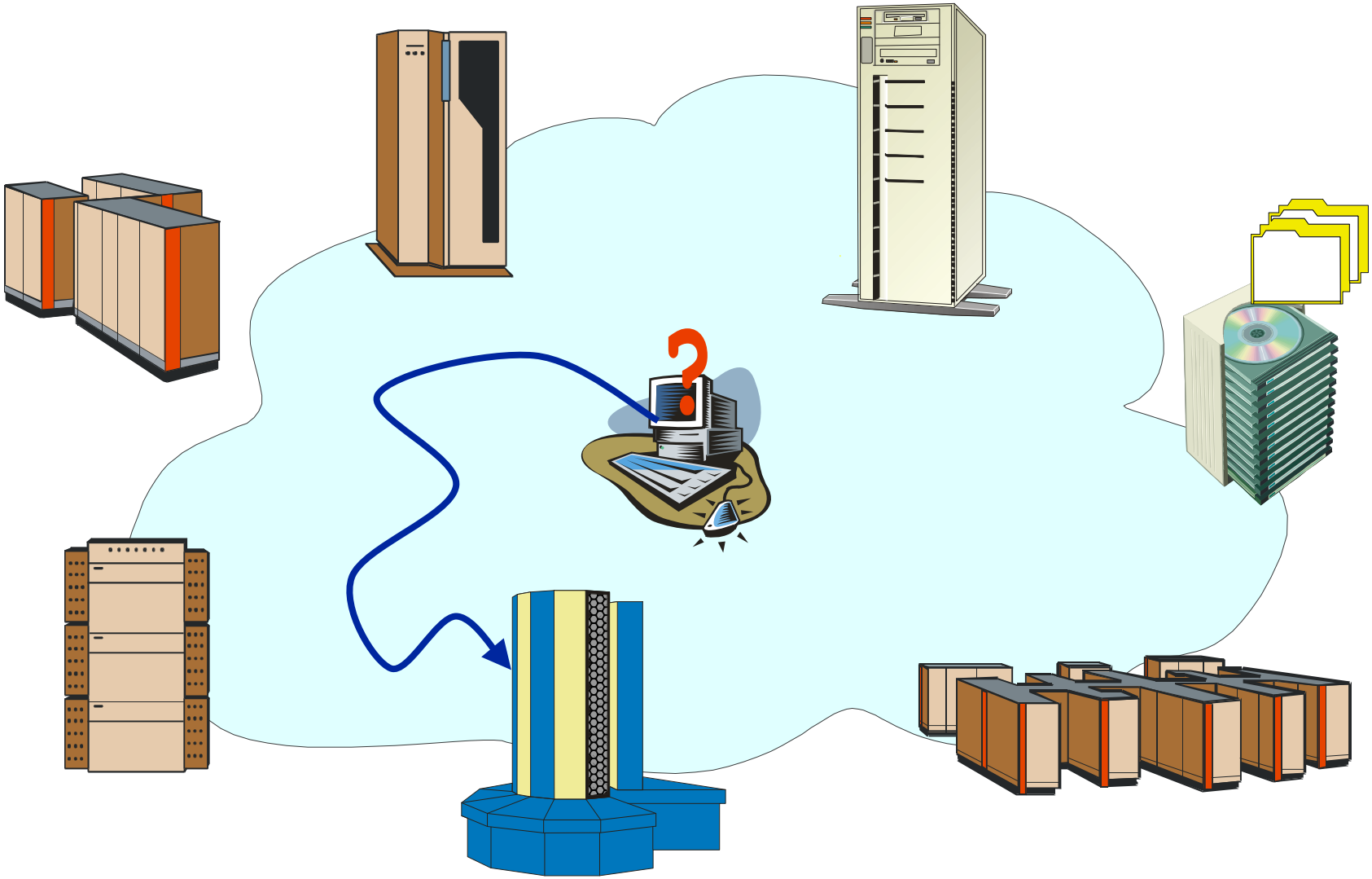
The European DataGrid Project Team

http://www.eu-datagrid.org

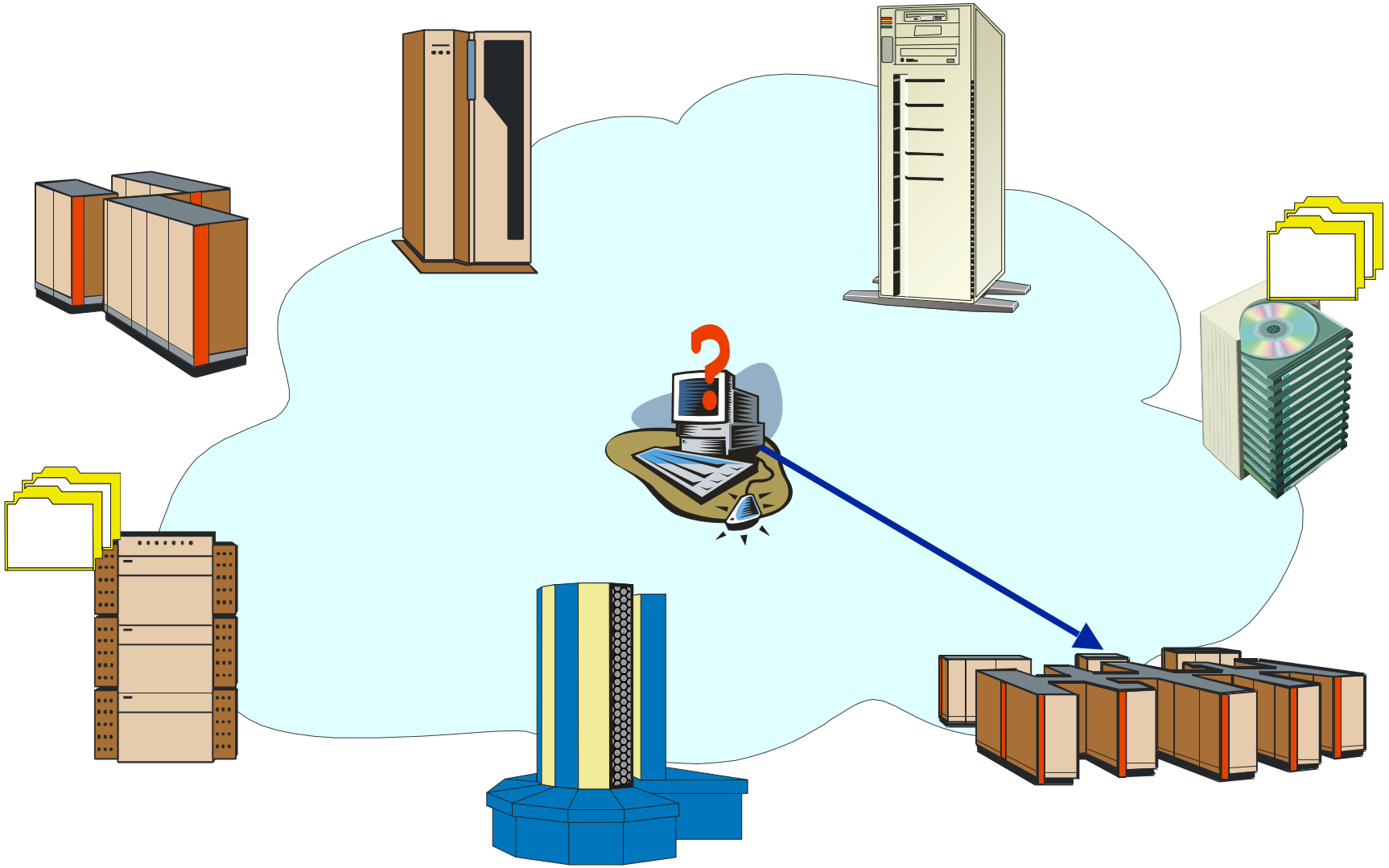
Overview

- Data Management Components
- Replica Location Service
- Replica Manager
- Metadata Management
- GDMP

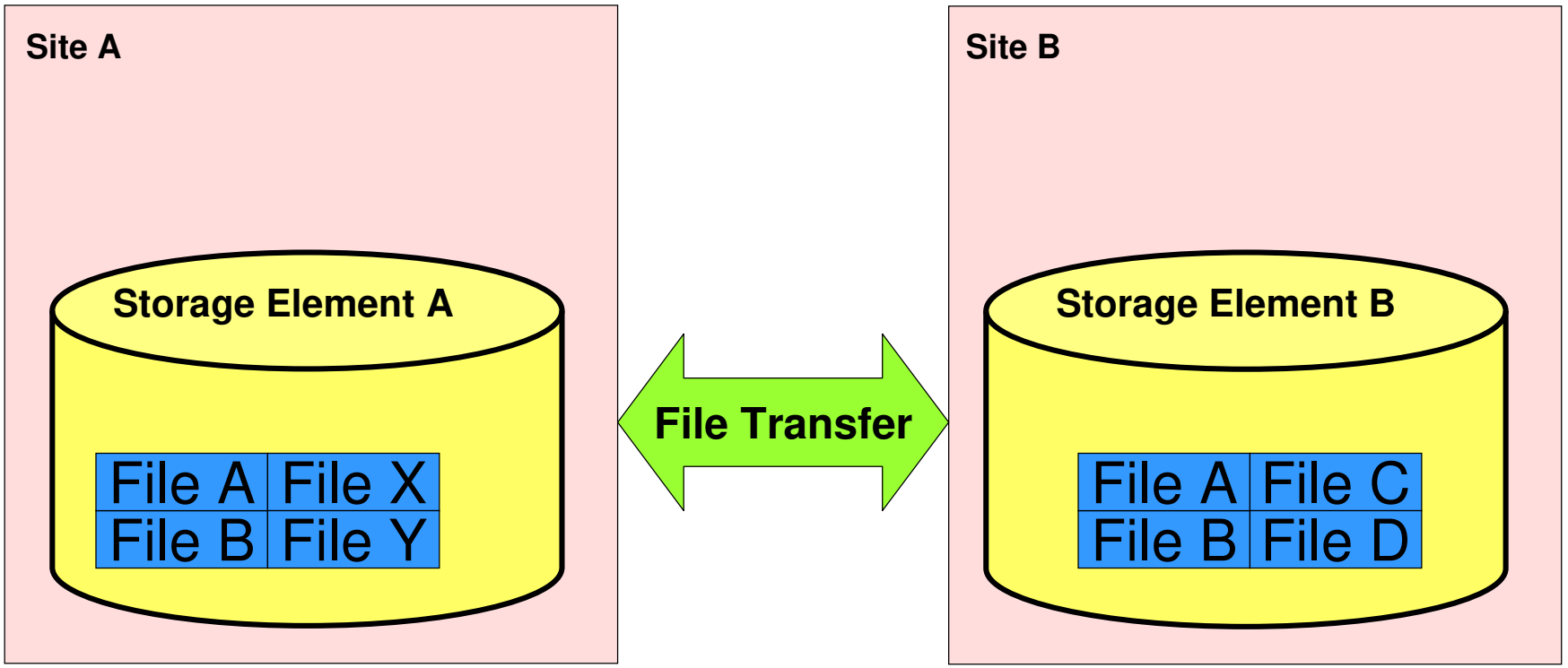
Data Management Issues



Data Management Issues

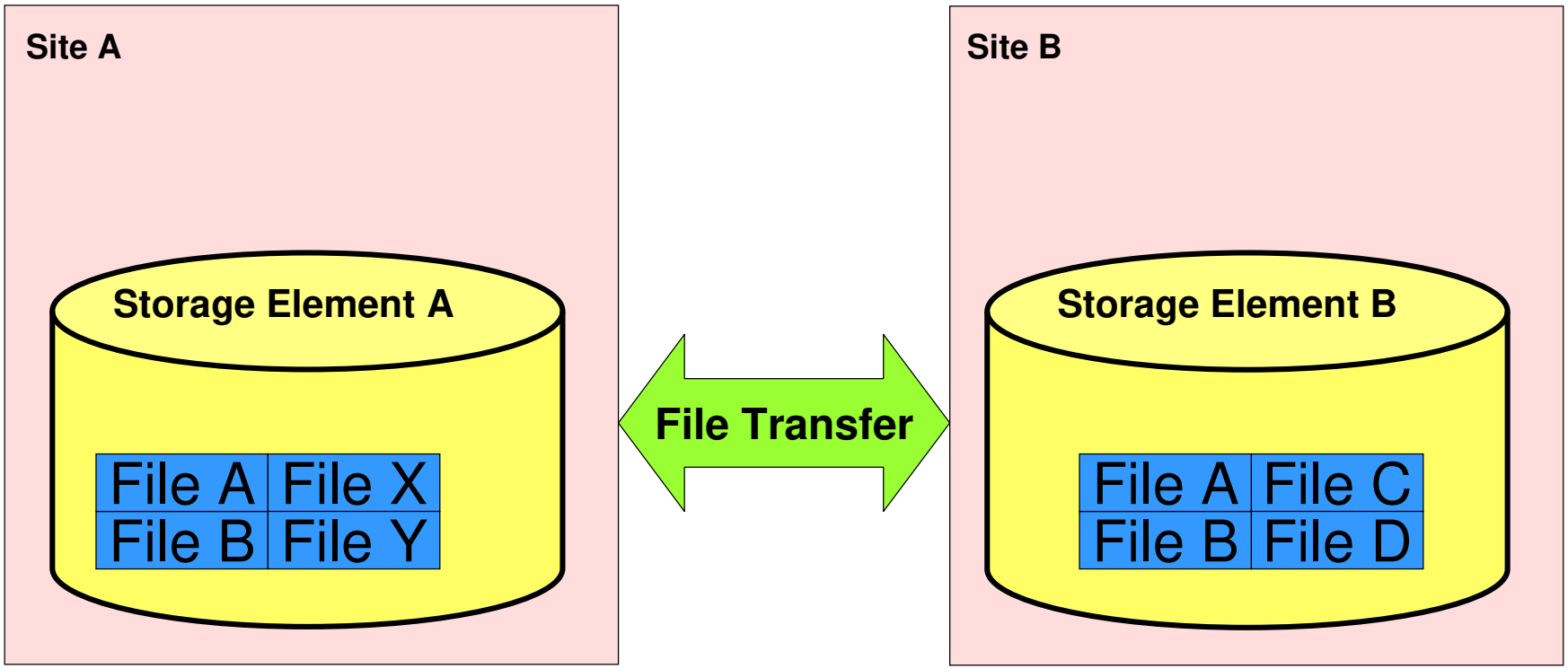


File Management



File Management

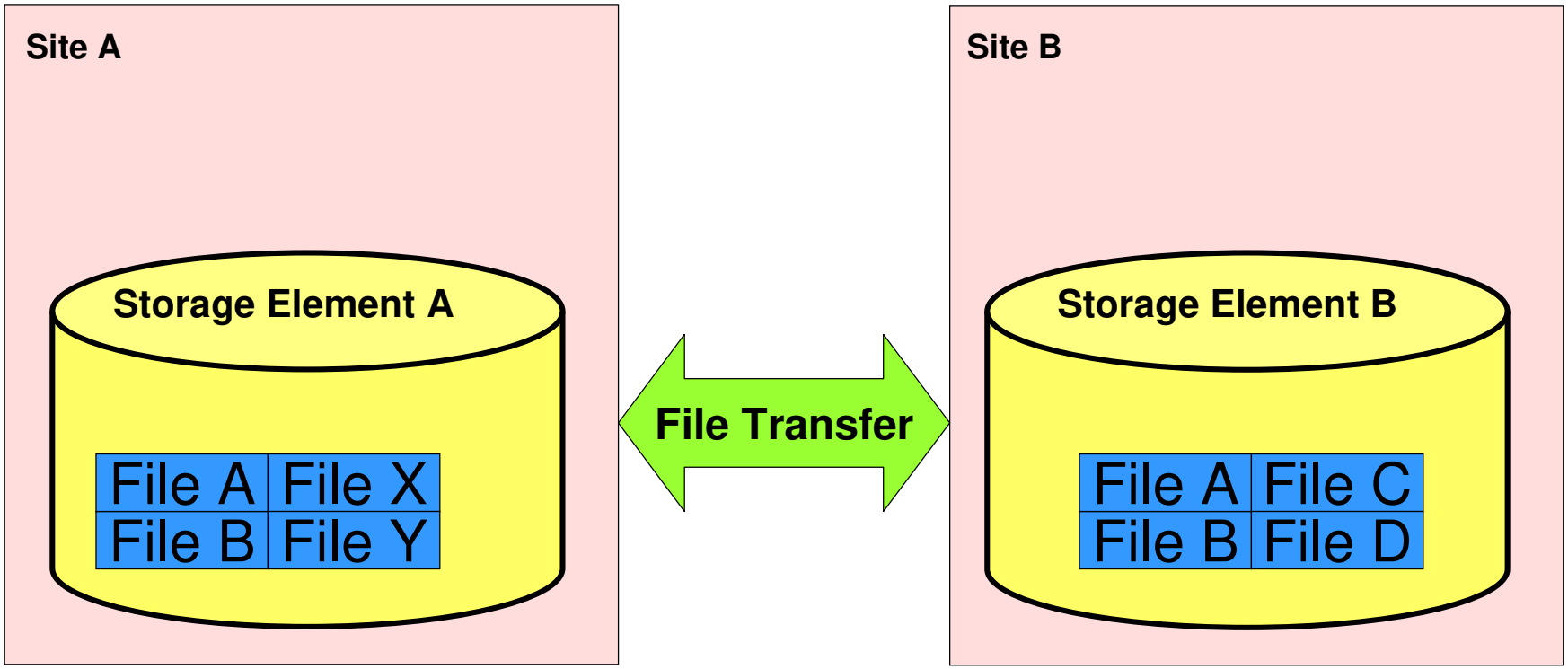
Replica Catalog:
Map Logical to Site files



File Management

Replica Catalog:
Map Logical to Site files

Replica Selection:
Get 'best' file

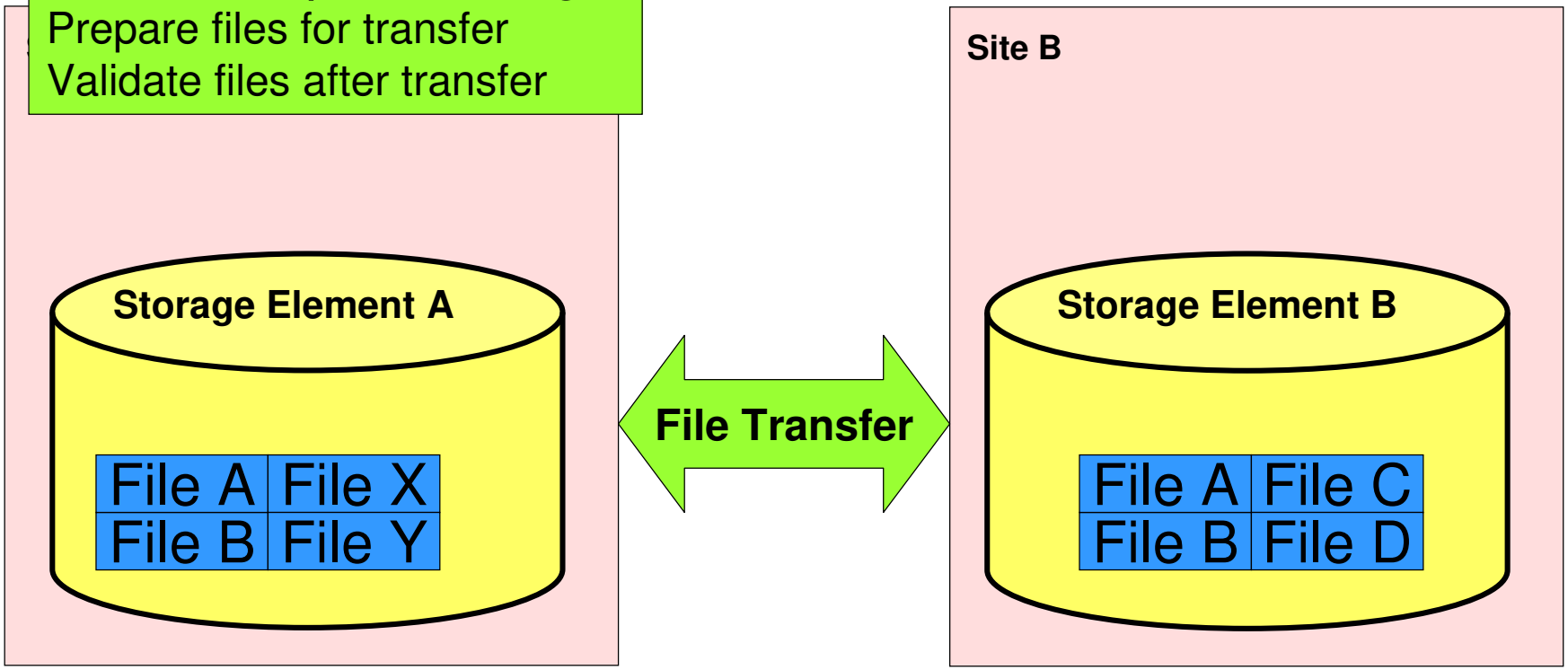


File Management

Replica Catalog:
Map Logical to Site files

Replica Selection:
Get 'best' file

Pre- Post-processing:
Prepare files for transfer
Validate files after transfer



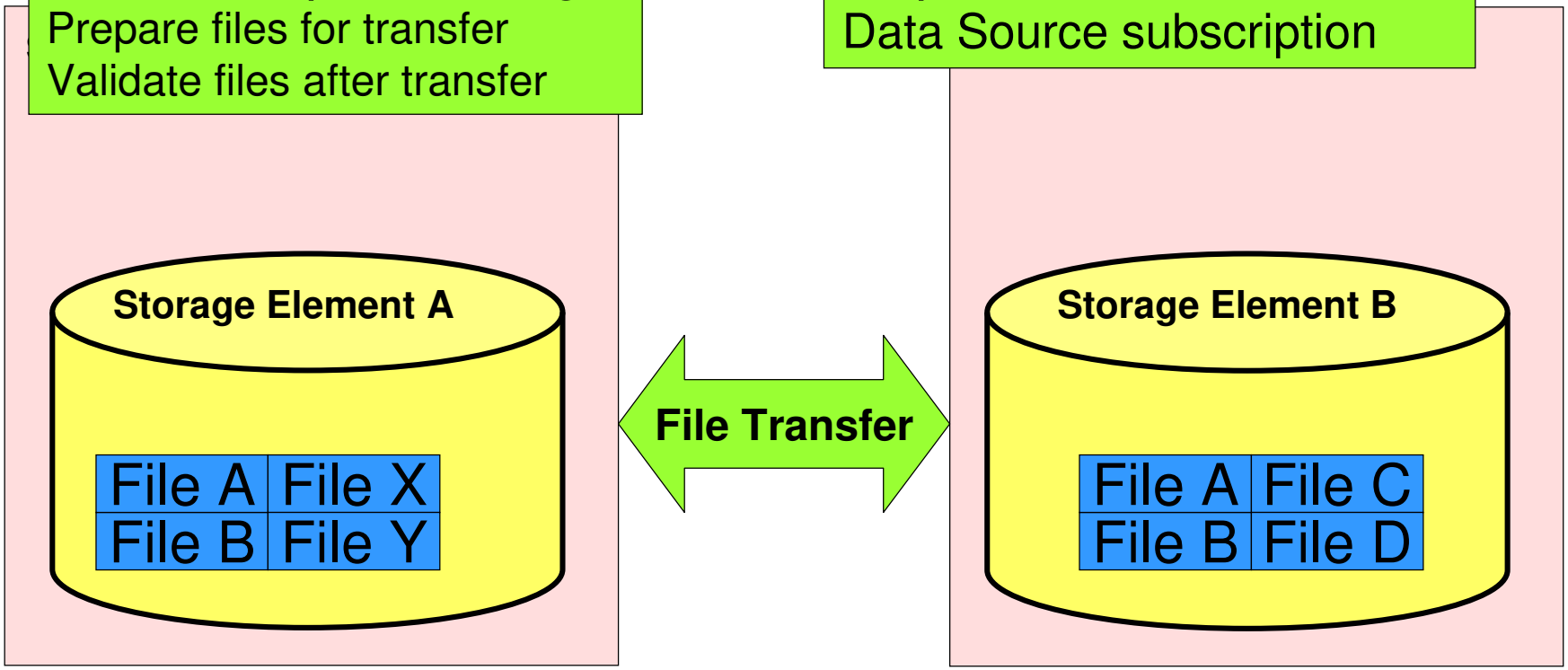
File Management

Replica Catalog:
Map Logical to Site files

Replica Selection:
Get 'best' file

Pre- Post-processing:
Prepare files for transfer
Validate files after transfer

Replication Automation:
Data Source subscription



File Management

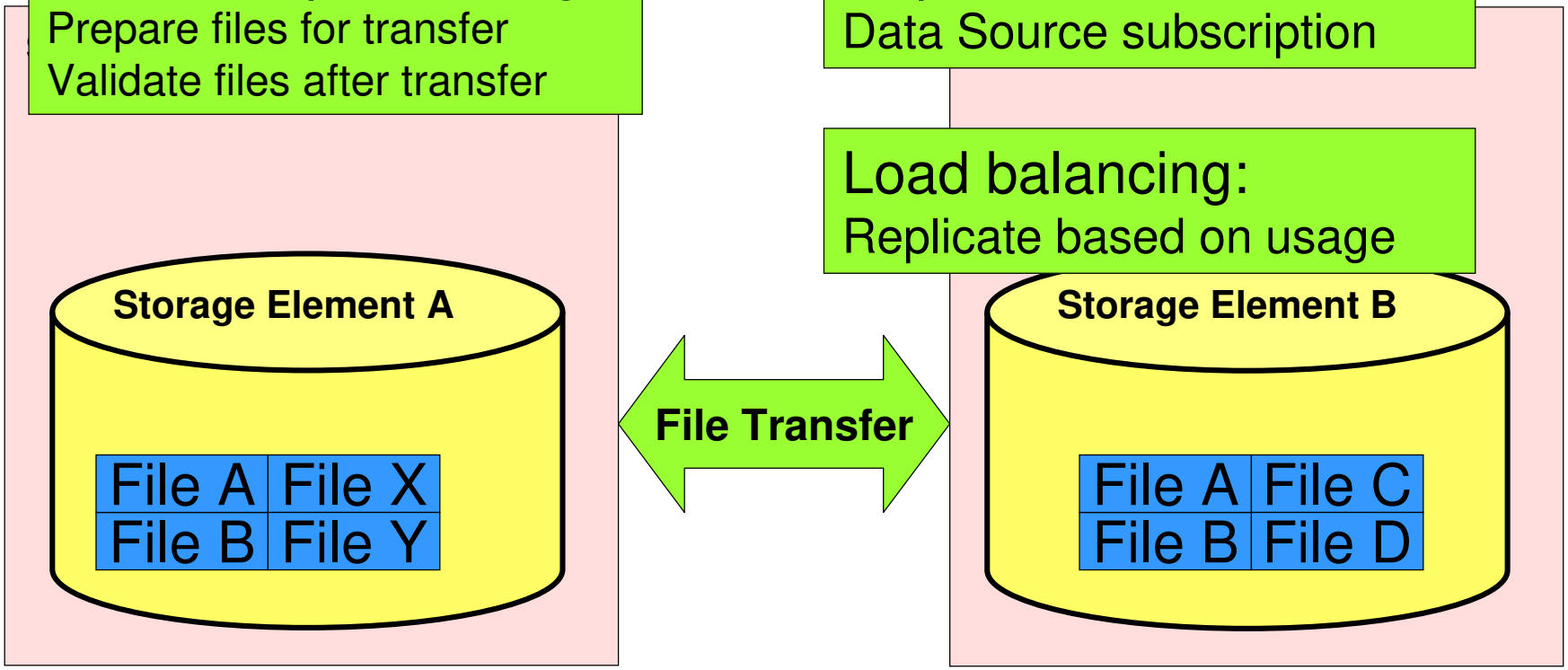
Replica Catalog:
Map Logical to Site files

Replica Selection:
Get 'best' file

Pre- Post-processing:
Prepare files for transfer
Validate files after transfer

Replication Automation:
Data Source subscription

Load balancing:
Replicate based on usage



File Ma

Replica Manager:
'atomic' replication operation
single client interface
orchestrator

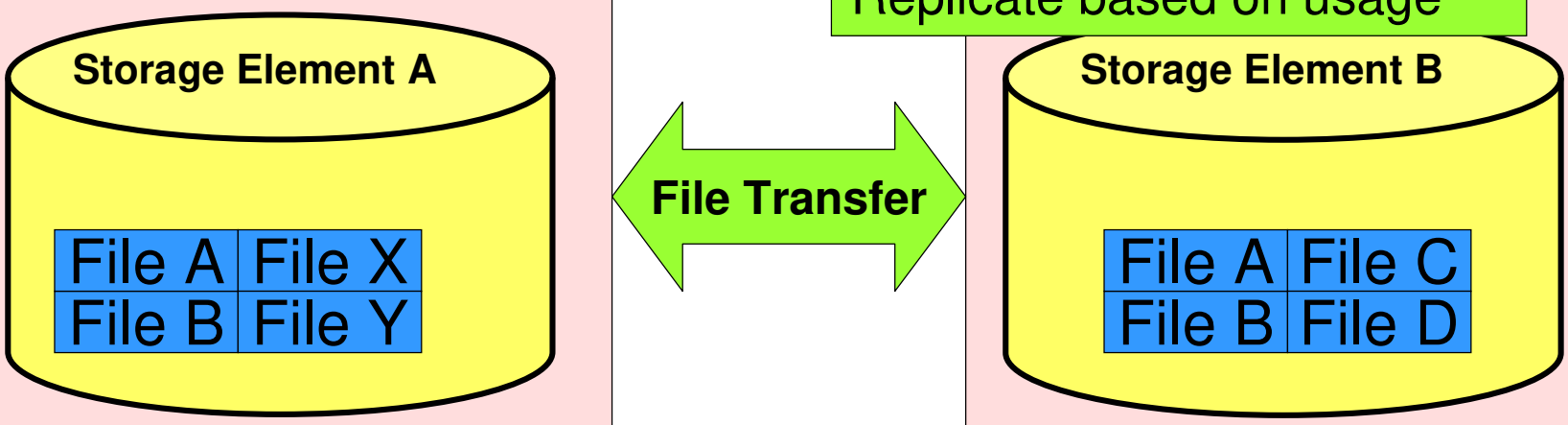
Replica Catalog:
Map Logical to Site files

Replica Selection:
Get 'best' file

Pre- Post-processing:
Prepare files for transfer
Validate files after transfer

Replication Automation:
Data Source subscription

Load balancing:
Replicate based on usage



File M

Replica Manager:
‘atomic’ replication operation
single client interface
orchestrator

Replica Catalog:
Map Logical to Site files

Replica Selection:
Get ‘best’ file

Pre- Post-processing:
Prepare files for transfer
Validate files after transfer

Replication Automation:
Data Source subscription

Load balancing:
Replicate based on usage

Metadata:
LFN metadata
Transaction information
Access patterns

File A | File X
File B | File Y

ile Transfer

Storage Element B

File A | File C
File B | File D

File M

Replica Manager:
‘atomic’ replication operation
single client interface
orchestrator

Replica Catalog:
Map Logical to Site files

Replica Selection:
Get be

Pre- Post-processing:
Prepare files for transfer
Validate files after transfe

Replication Automation:
Data Source subscription

Security

Metadata:
LFN metadata
Transaction information
Access patterns

Load balancing:
Replicate based on usage

File Transfer

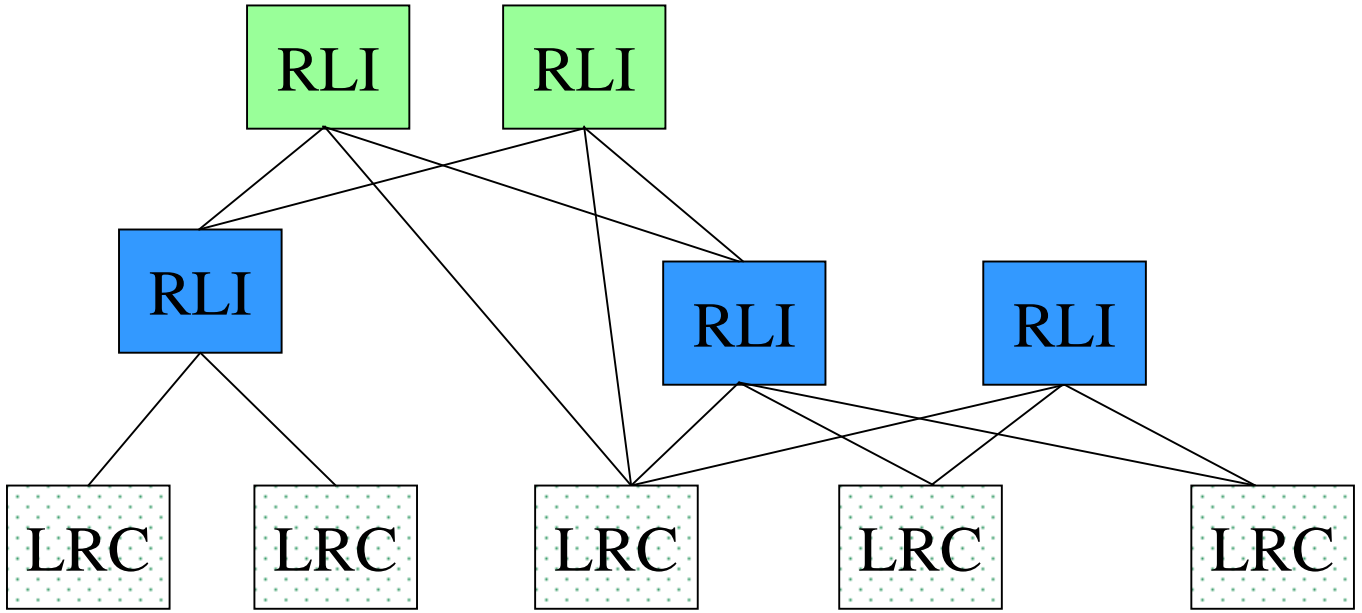
Storage Element B

File A	File X
File B	File Y

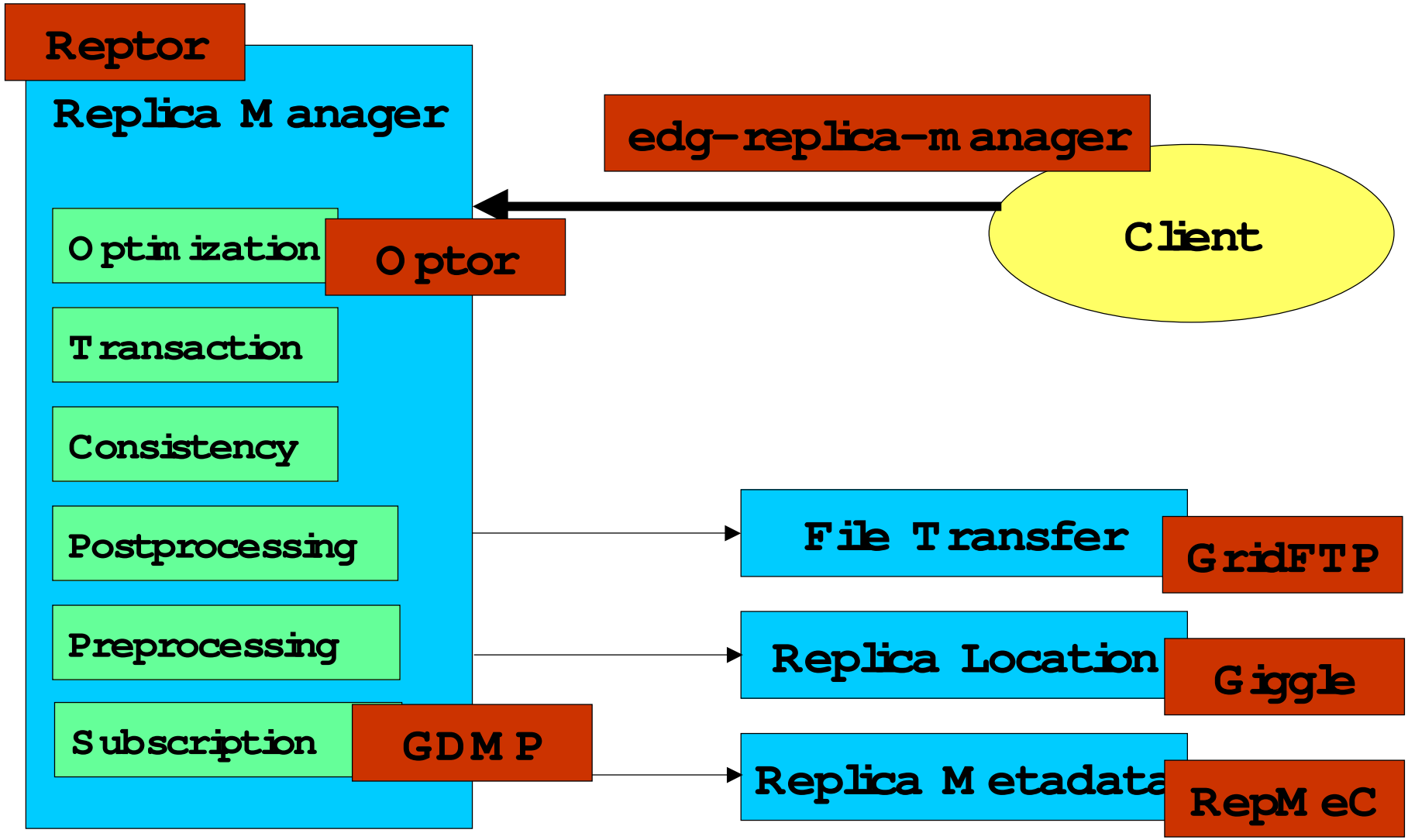
File A	File C
File B	File D

Replica Location Service

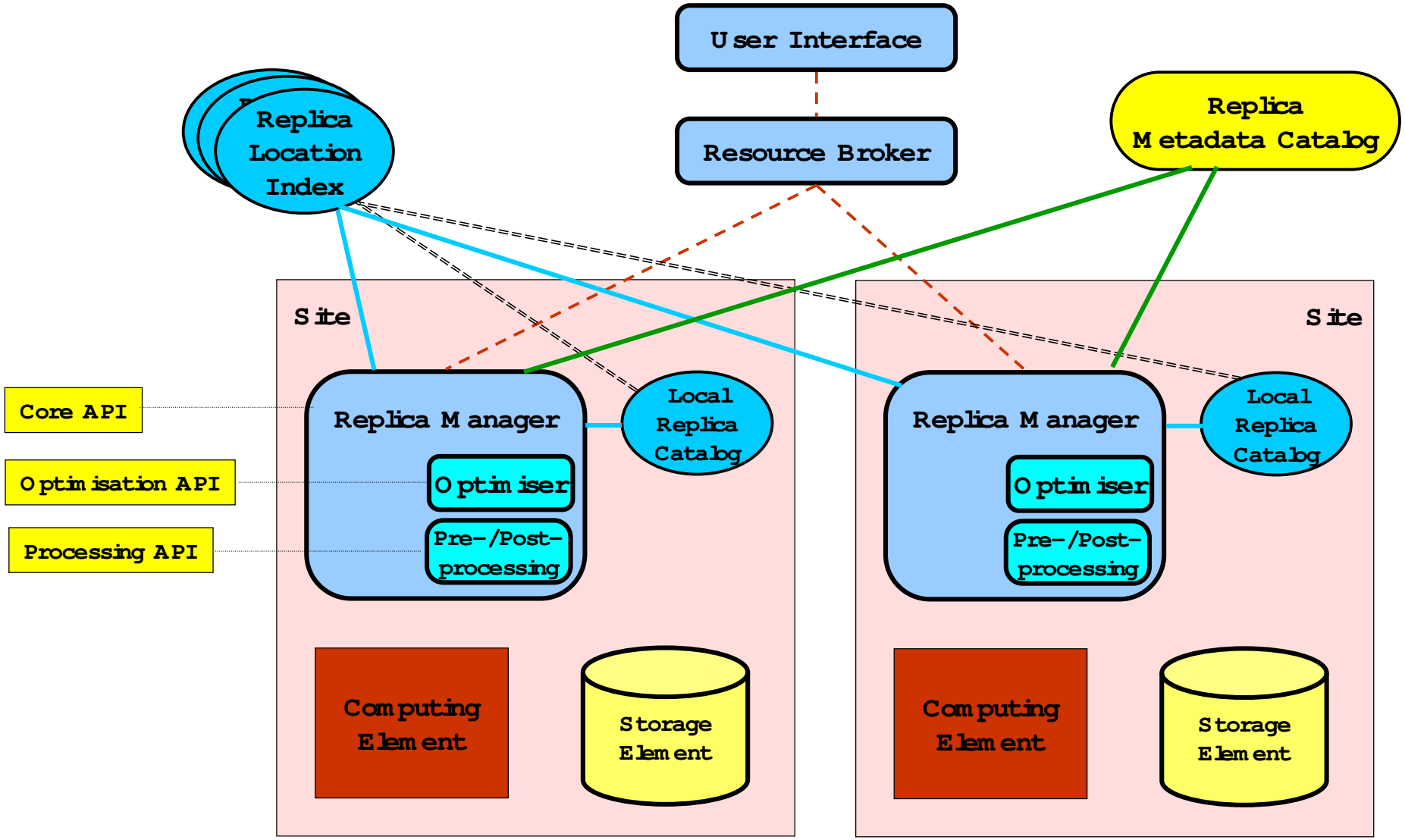
- Scalable, distributed replica location service (RLS)
- Joint collaboration between WP2 and Globus
- Local replica catalogs (LRCs) & collective replica location indices (RLI)
- LRC holds reliable local state, RLI unreliable collective state with soft-state update



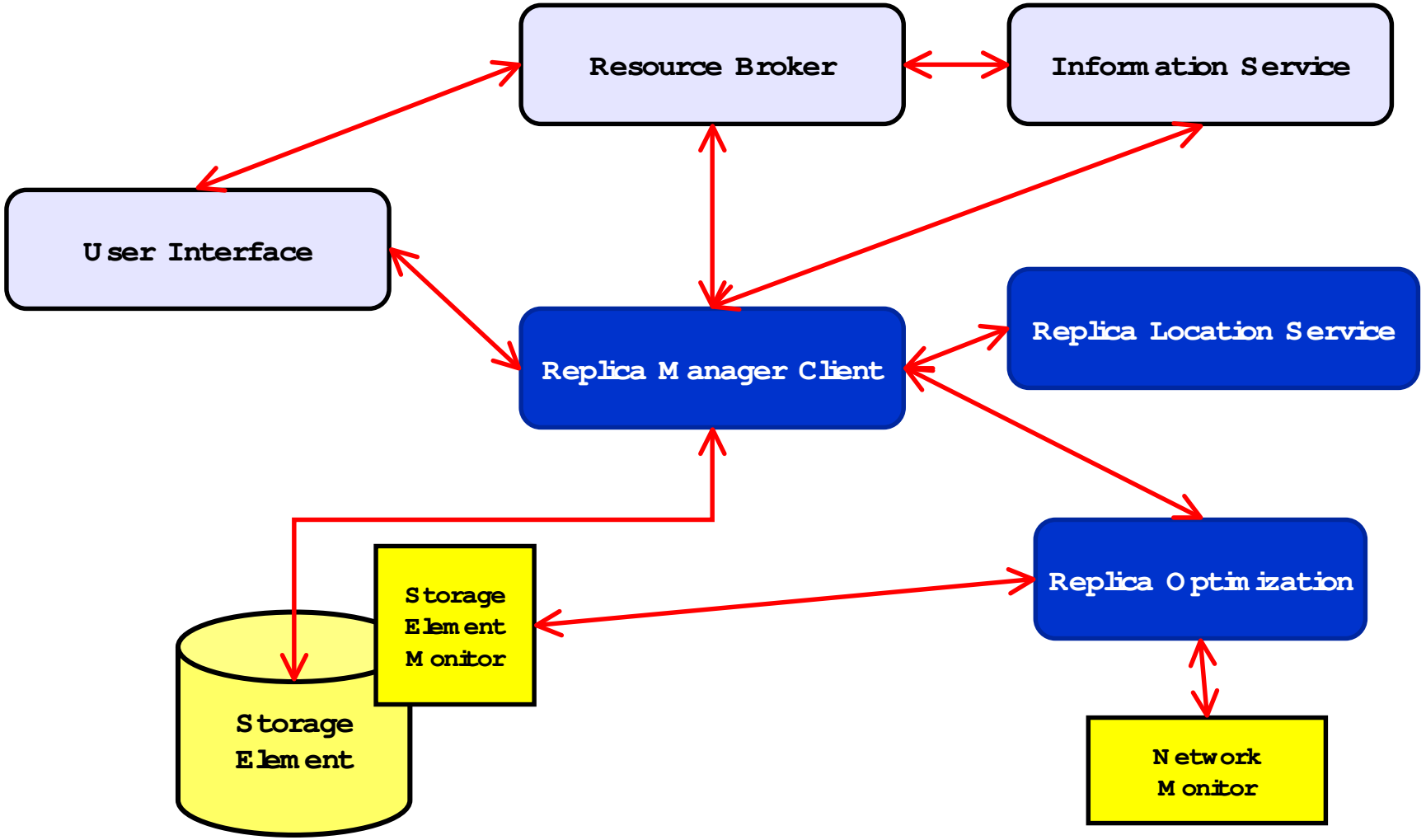
Replication Services: EDG Replica Manager



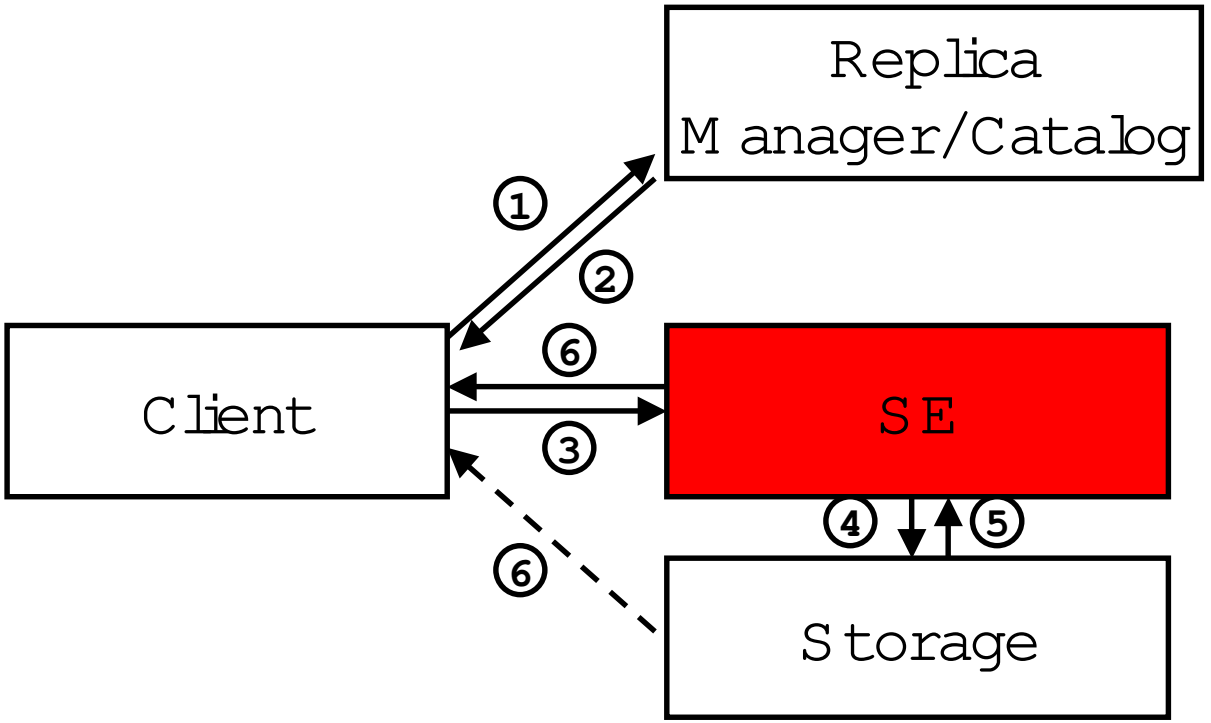
Replication Services Architecture



Replication Optimization Architecture



SE Interaction



1. The Client asks a catalog to provide the location of a file
2. The catalog responds with the name of an SE
3. The client asks the SE for the file
4. The SE asks the storage system to provide the file
5. The storage system sends the file to the client through the SE or
6. directly

Metadata Management and Security

Project Spitfire

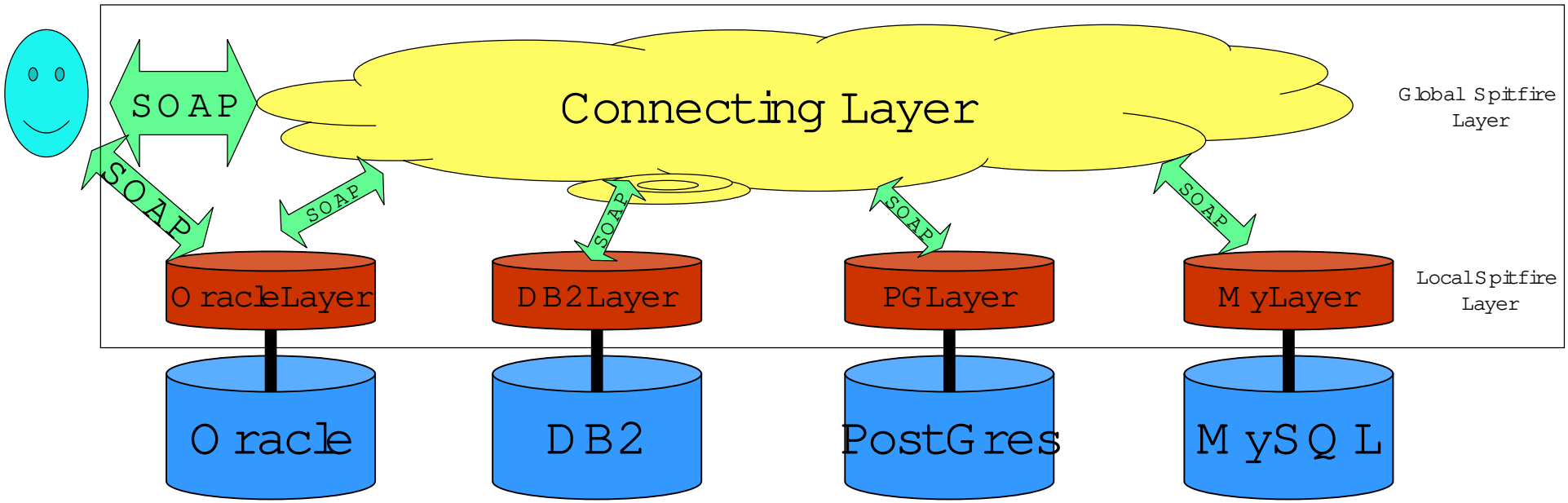
- 'Simple' Grid Persistency
  - Grid Metadata
  - Application Metadata
  - Unified Grid enabled front end to relational databases.
- *Metadata Replication and Consistency*
- *Publish information on the metadata service*

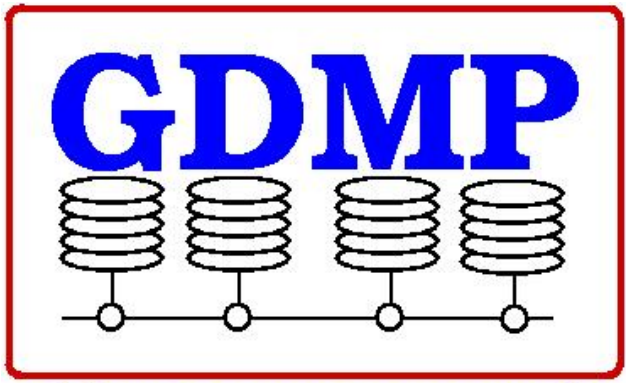
Secure Grid Services

- Grid authentication, authorization and access control mechanisms enabled in Spitfire
- Modular design, reusable by other Grid Services

Spitfire Architecture

- XSQL Servlet as one access mode for 'simple' web access
- Web/Grid Services Paradigm
  - SOAP interfaces
  - JDBC interface to RDBMS
- Plugability and extensibility
- Atomic RDBMS is always consistent
- No local replication of data
- Role-based authorization

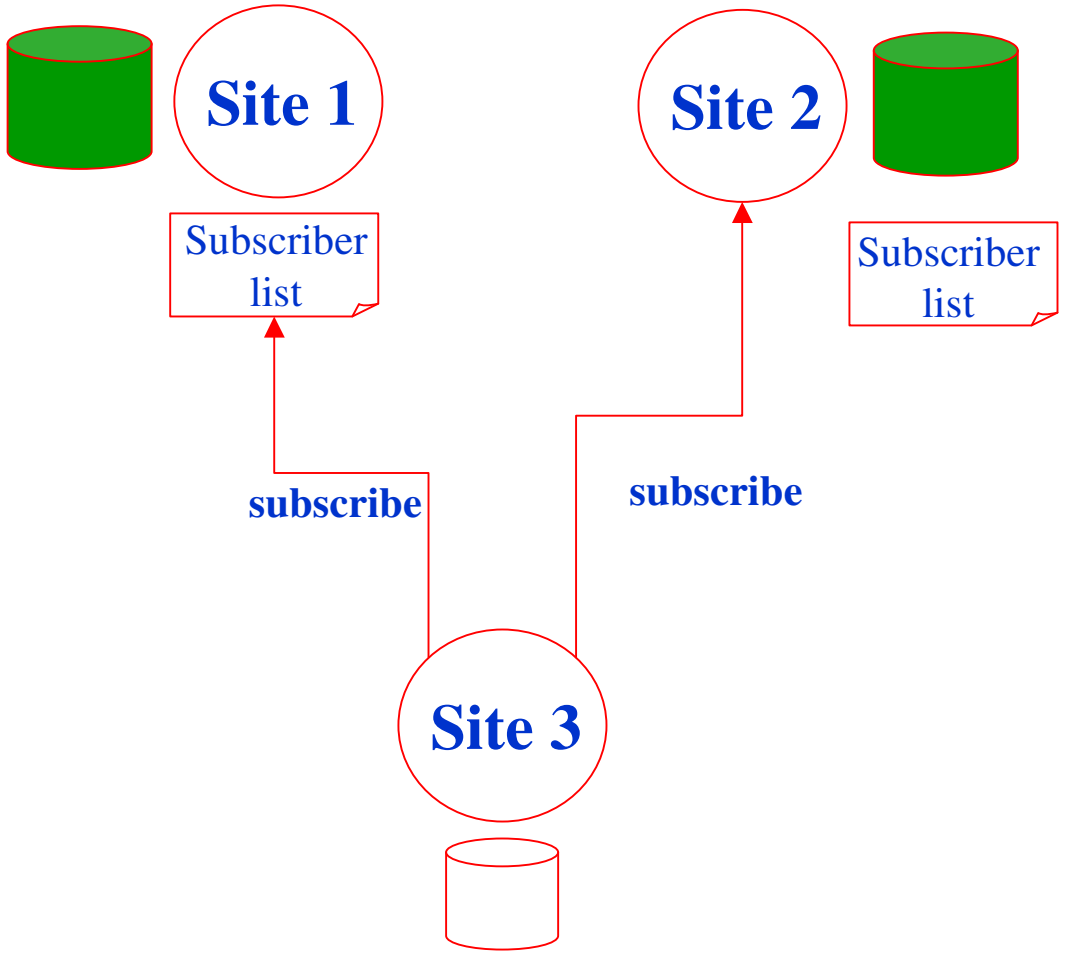


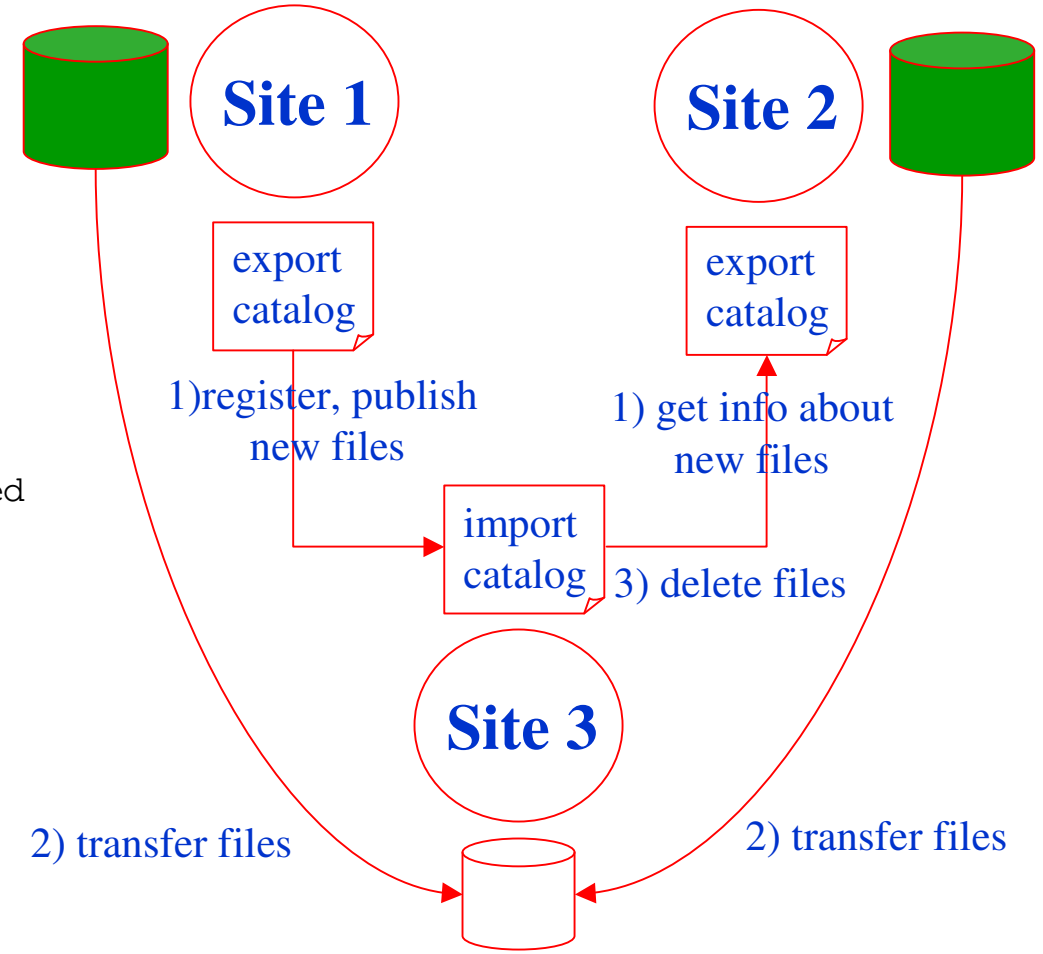


- based on CMS requirements for replicating Objectivity files for High Level Trigger studies
- **production** prototype project for evaluating Grid technologies (especially Globus)
- experience for will directly be used in DataGrid
  - input also for PPDG and GriPhyN
- http://cern.ch/GDMP
- implementation: Asad Samar, Heinz Stockinger, Shazhad Muzaffar, Flavia Donno, Aleksandr Konstantinov

Subscription Model

- All the sites that subscribe to a particular site get notified whenever there is an update in its catalog.



Export / Import Catalogue

- Export Catalog
  - information about the new files produced.
  - is published
- Import Catalog
  - information about the files which have been published by other sites but not yet transferred locally
  - As soon as the file is transferred locally, it is removed from the import catalogue.
- Possible to pull the information about new files in your import catalogue.