eGee

Enabling Grids for E-science in Europe

Grid Data Management

Vered Kunik - Israeli Grid NA3 Team
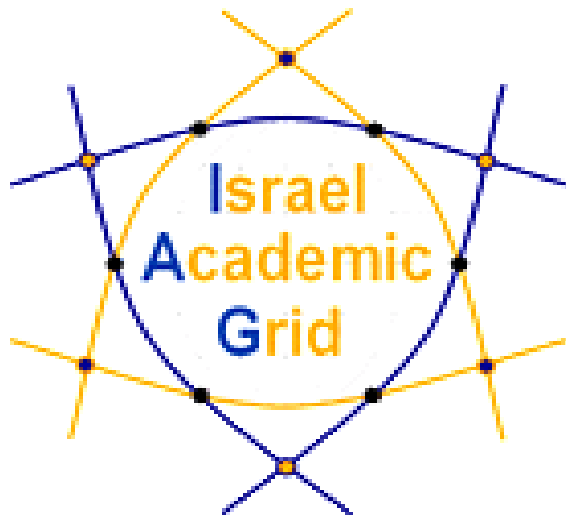
Israel Academic Grid

Outline
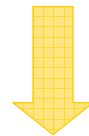
- **Introduction**

- **Data Management Services**

- **File catalogues**

- **Replica management commands**

- **Hands on**

Introduction

What does "Data Management" mean ?

- Users and applications produce and require data

  → Users and applications need to handle files <u>on the Grid</u>

- Files are stored in appropriate permanent resources called "Storage Elements" (SE)

- Appropriate data management utilities/services hide the internal structure of the SE's and transfer protocols

  → The users / applications do not need to know where the data is located

Data Management Services

The Data Management Services enable Grid users to:

- move files in and out of the Grid
- Replicate files on different SE's
- Locate files on various SE's

   provided by the **R**eplica **M**anagement **S**ystem (**RMS**)

This is achieved by:

- Transferring data using a number of protocols (mainly gsiftp)
- Interacting with a central file catalogue = the **R**eplica **L**ocation **S**ervice (**RLS**)
- Data Management focuses on "large" files (> 20MB – a few hundred MB)

Files & replicas

- **Globally Unique Identifier (GUID)**
  - A file can always be identified by its GUID (based on UUID)
  - A GUID is of the form: guid:<unique_string>
  - All replicas of a file will share the same GUID
  - **Example**: "guid:f81d4fae-7dec-11d0-a765-00a0c91e6bf6"

- **Logical File Name (LFN)**
  - An alias created by the user to refer to some file
  - A LFN is of the form: lfn:<any_alias>
  - **Example**: lfn:/grid/gilda/importantResults/Test1240.dat

➡ refer to **files** (**not** to **replicas)**

Files & replicas – cont'd

- **Storage URL (SURL)**
- **(AKA: Physical/Storage File Name (PFN/SFN))**
  - Used by the RMS to find where the replica is physically stored
  - A SURL is of the form:
    sfn://<SE_hostname><SE_Accesspoint><VO_path><file_name>
  - **Example**: sfn://tbed1.cern.ch/flatfiles/SE00/gilda/project1/testSUTL.dat

- **Transport URL (TURL)**
  - Temporary locator of a physical replica including the access protocol understood by a SE
  - A TURL is of the form:
    <protocol>://<SE_hostname><SE_Accesspoint><VO_path><filename>
  - **Example**: gsiftp://tbed1.cern.ch/flatfiles/gilda/project1/testTURL.dat

  provide info about the physical location of the **replica**

File Catalogs

So…

- How do I keep track of all my files on the Grid?
- Even if I remember all the LFNs of my files, what about someone else's files ?
- How does the Grid keep track of associations lfn/GUID/SUTL ?

- **Well… for that we have a FILE CATALOG**

The Replica Management System

The Replica Location Server (RLS) is composed of:

- **Local Replica Catalog = LRC:**
  - Stores GUID→SURL mappings
  - Accessible by edg-lrc

- **Replica Metadata Catalog = RMC:**
  - Stores LFN→GUID mappings
  - Accessible by edg-rmc
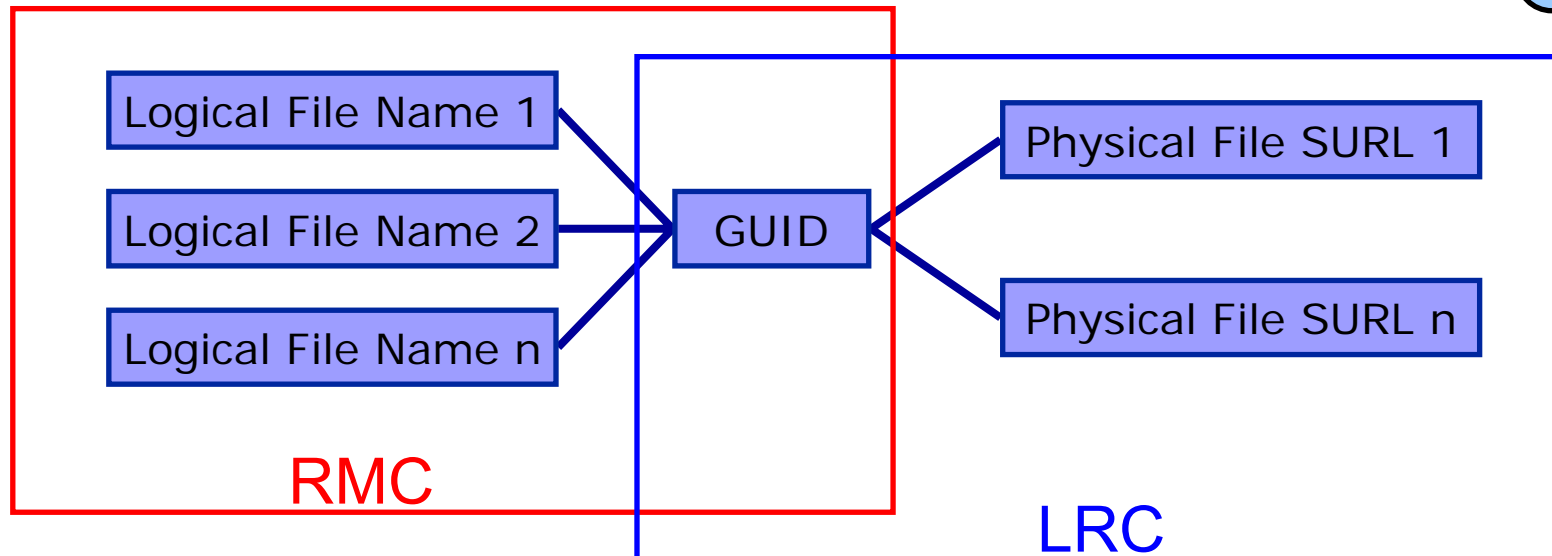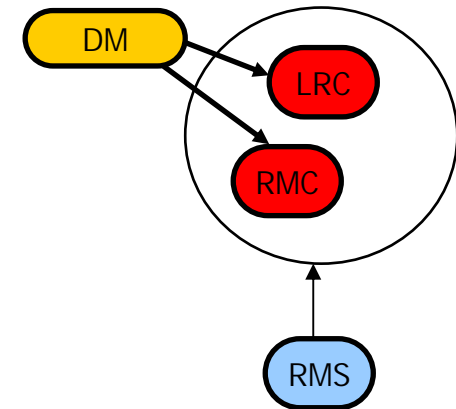
Replica management commands

- **lcg-cp**   Copies a Grid file to a local destination

- **lcg-cr**   Copies a file to a SE and registers the file in the LRC

- **lcg-del**   Deletes one file (either one replica or all replicas)

- **lcg-rep**   Copies a file from SE to SE and registers it in the LRC

- **lcg-se**   set file status to "Done" in a specified request

Replica management commands – cont'd

- **lcg-aa**   Adds an alias in RMC for a given GUID

- **lcg-gt**   Gets the TURL for a given SURL and transfer protocol

- **lcg-la**   Lists the aliases for a given LFN, GUID or SURL

- **lcg-lg**   Gets the GUID for a given LFN or SURL

Replica management commands – cont'd

- **lcg-lr**   Lists the replicas for a given LFN, GUID or SURL

- **lcg-ra**   Removes an alias in RMC for a given GUID

- **lcg-rf**   Registers a SE file in the LRC (optionally in the RMC)

- **lcg-uf**   Un-registers a file residing on an SE from the LRC

To Data Management tutorial