



Enabling Grids for
E-science in Europe

www.eu-egee.org

Padova, JRA1 all-hands meeting, 15-17 Nov 2004

Data Requirements in the JDL

F. Pacini,
Datamat S.p.A.



EGEE is a project funded by the European Union under contract IST-2003-508833

Content



- WMS/DM interaction in LCG
- New Requirements
- WMS/DM interaction in EGEE
- JDL for gLite

WMS and DM in LCG-2 (1/2)

- The LCG-2 WMS Resource Broker uses file locations as the base for determining where input data is located:
 - the Replica Location Service (RLS) is queried via the replica manager method *listReplicas*
 - LFNs/GUIDs are mapped to PFNs
 - PFNs are used to get the list of SEs where files are located
- Supported Input Data types are:
 - File - Logical File Name (LFN)
 - Global Unique Identifier (GUID)

WMS and DM in LCG-2 (2/2)

- Data Requirements for a Job are expressed as follows:

```
[  
...  
Executable = "mytest.exe";  
...  
InputData = {  
    "lfn:EOTestfile" ,  
    "lfn:ATLATESTfile",  
    "guid:135b7b23-4a6a-11d7-87e7-9d101f8c8b70"  
};  
DataAccessProtocol = {"file", "gridftp"};  
...  
]
```

New Requirements

- Datasets rather than logical files:
 - Many application users want to deal with sets of files (collections or datasets) rather than single files addressed by LFNs
 - The concept of a Logical Data Set (LDS) is required.
- Allow for a generic query string:
 - Once a dataset catalogue is in place, application users typically express their input data requirements in an experiment specific way to query data sets
 - The query is resolved by the dataset catalogue which returns locations of the requested datasets
- Unique query interface to Data Catalogues:
 - There are currently several ongoing projects that are building replica catalogue systems
 - Make these catalogues expose a common query interface is a goal to be pursued

WMS and DM in EGEE (1/6)

- The LCG Data Location Interface (DLI) provides a unique query interface to:
 - File Replica Catalogues (e.g. EDG RLS)
 - Dataset Catalogues (e.g. CMS Dataset catalogue)
- DLI allows the specification of generic queries
- This leads to the following list of supported Input Data types:
 - File - Logical File Name (LFN)
 - Global Unique Identifier (GUID)
 - Dataset - Logical DataSet (LDS)
 - Generic query (query)

WMS and DM in EGEE (2/6)

- The **gLite Storage Index** provides a replica catalogue interface for locating files stored in SEs (LFN-PFN mapping)
- Supported InputData types are hence:
 - File - Logical File Name (LFN)
 - Global Unique Identifier (GUID)
- Query responses (coming from both **DLI** and **SI**) are always used by WMS to determine the list of SEs where input data are located
- *How can these different InputData types supported by different catalogues be specified within the JDL so that the WMS can appropriately use them for the matchmaking?*

WMS and DM in EGEE (3/6)

- Each of them can be assigned with a corresponding prefix to be used in the JDL specification. We propose the following:
 - **Data Location Interface**
 - lfn: <valid LFN String>
 - guid: <valid GUID String>
 - lds: <valid Dataset String>
 - query: <valid Query String>
 - **Storage Index**
 - si-lfn: <valid LFN String>
 - si-guid: <valid GUID String>

WMS and DM in EGEE (4/6)

```
[  
...  
Executable = "mytest.exe";  
...  
InputData = {  
    "lfn:E0testfile" ,  
    "guid:135b7b23-4a6a-11d7-87e7-9d101f8c8b70" ,  
    "lds:mu03_tt_4mu/mu_Hit245_2_g133" ,  
    "query:my last week CMS datasets" ,  
    "si-lfn:ALICEtestfile" ,  
    "si-guid: 9876abdc-my67-223ed-84e5-9f141f6c0f22"  
};  
DataAccessProtocol = {"file", "gridftp"};  
...  
]
```

Storage Index {

} **DLI**

WMS and DM in EGEE (5/6)

- InputData types from different catalogues can be mixed in the same JDL specification
- The WMS relies on the prefix for choosing the appropriate catalogue interface to query for a given InputData item
- The default catalogue for a particular VO can be either obtained through service discovery in the Information Service or set as a parameter in the WM configuration
- The user can specify a non-default (experiment specific) catalogue to be used for determining data location by specifying the service endpoint URL in the JDL.
- The corresponding JDL attributes for the two query interfaces supported by the WMS are respectively :
 - **DataCatalog**
 - **StorageIndex**

WMS and DM in EGEE (6/6)

```
[  
  ...  
  InputData = {  
    "ids:mu03_tt_4mu/mu_Hit245_2_g133",  
    "query:my CMS last week datasets",  
    "si-lfn:ALICETestfile",  
    "si-guid: 9876abdc-my67-223ed-84e5-9f141f6c0f22"  
  };  
  DataCatalog = "https://example.org:8443/CMSDataSetCatalog";  
  StorageIndex = "https://glite.org/data/StorageIndex";  
  ...  
]
```

The gLite's JDL

- JDL specification (the one “understood” by the WMS) should be ready at the end of next week
- How many flavors of JDL do we have in gLite?
- At least two:
 - Based on Classads
 - Semantically equivalent
 - Syntactically incompatible
- Can we at least converge on a common sub-set of attributes?