



UNIVERSITY OF DORTMUND

ROBOTICS RESEARCH INSTITUTE
INFORMATION TECHNOLOGY SECTION



HEP CG Scheduling Architecture

Stefan Freitag

Stefan.Freitag@udo.edu

University Dortmund, IRF-IT

29/11/2006



Current situation



- Allocation of jobs to sites may not incorporate knowledge about data availability
- Data may be on tertiary storage and first access is delayed
- Storage element does not provide advanced planning features
- Our work is based on gLite 3.0.x and dCache 1.7



Goals



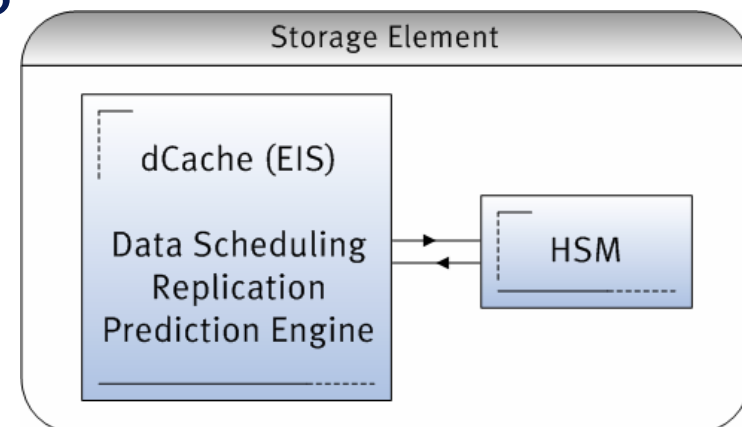
- Co-scheduling of jobs and data
- More precise planning and prediction of data availability prior to job allocation to compute elements
- Better integration of local job and data scheduling to improve response times and through-put



Extension of the SE dCache

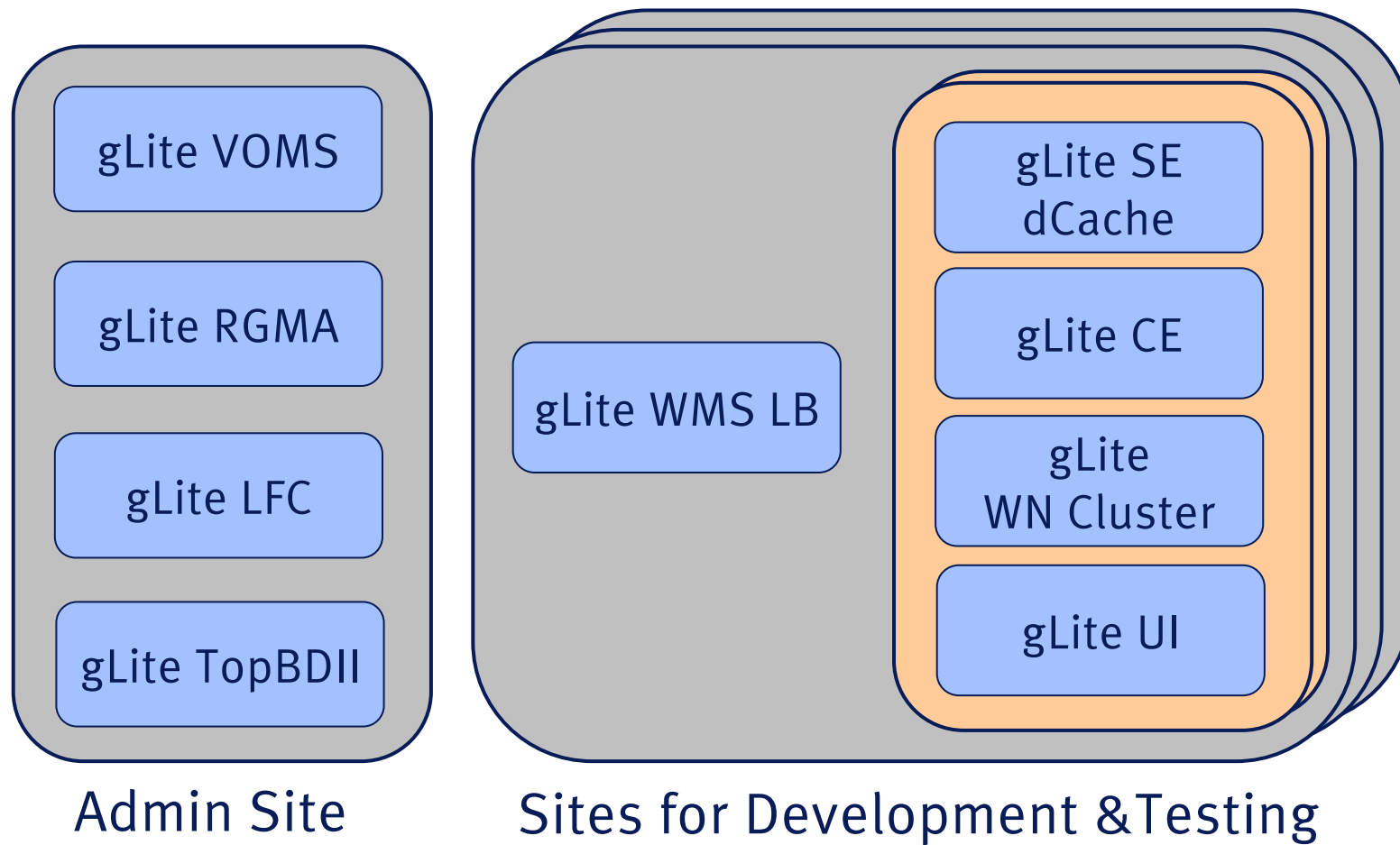


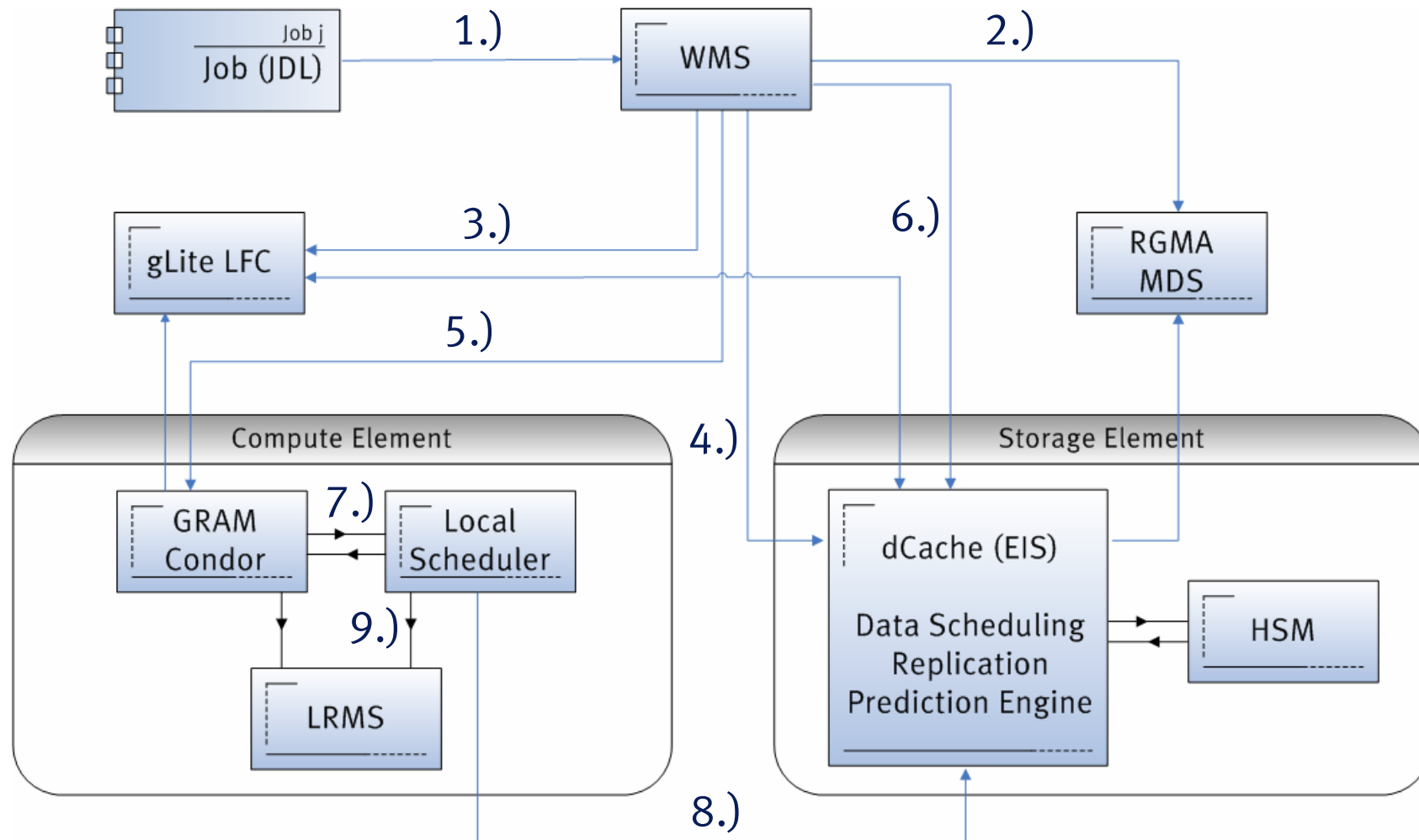
- File state information in HSM based systems can not be used properly
 - ➔ Data appears “available” on the local file system, but is located on tertiary storage
 - ➔ Staging on demand can take a long period of time
 - ➔ Ready to run jobs are delayed
- Extension of the dCache EIS
 - ➔ Prediction engine
 - ➔ Space reservation
 - ➔ SLA





Used Environment@IRF-IT







Extensions of gLite



- Several packages of gLite 3.0.0 Final CVS trunk modified

→ org.glite.wms. ...

- broker
- helper
- jdl
- jobsubmission
- manager



Extensions of gLite



■ New packages

→ org.glite.wms. ...

- dCache Extended Information Service (EIS) library
- dCache EIS Client
- WSDL definition (DESY)

→ org.glite.ce. ...

- Jobscheduler daemon
- Jobscheduler client



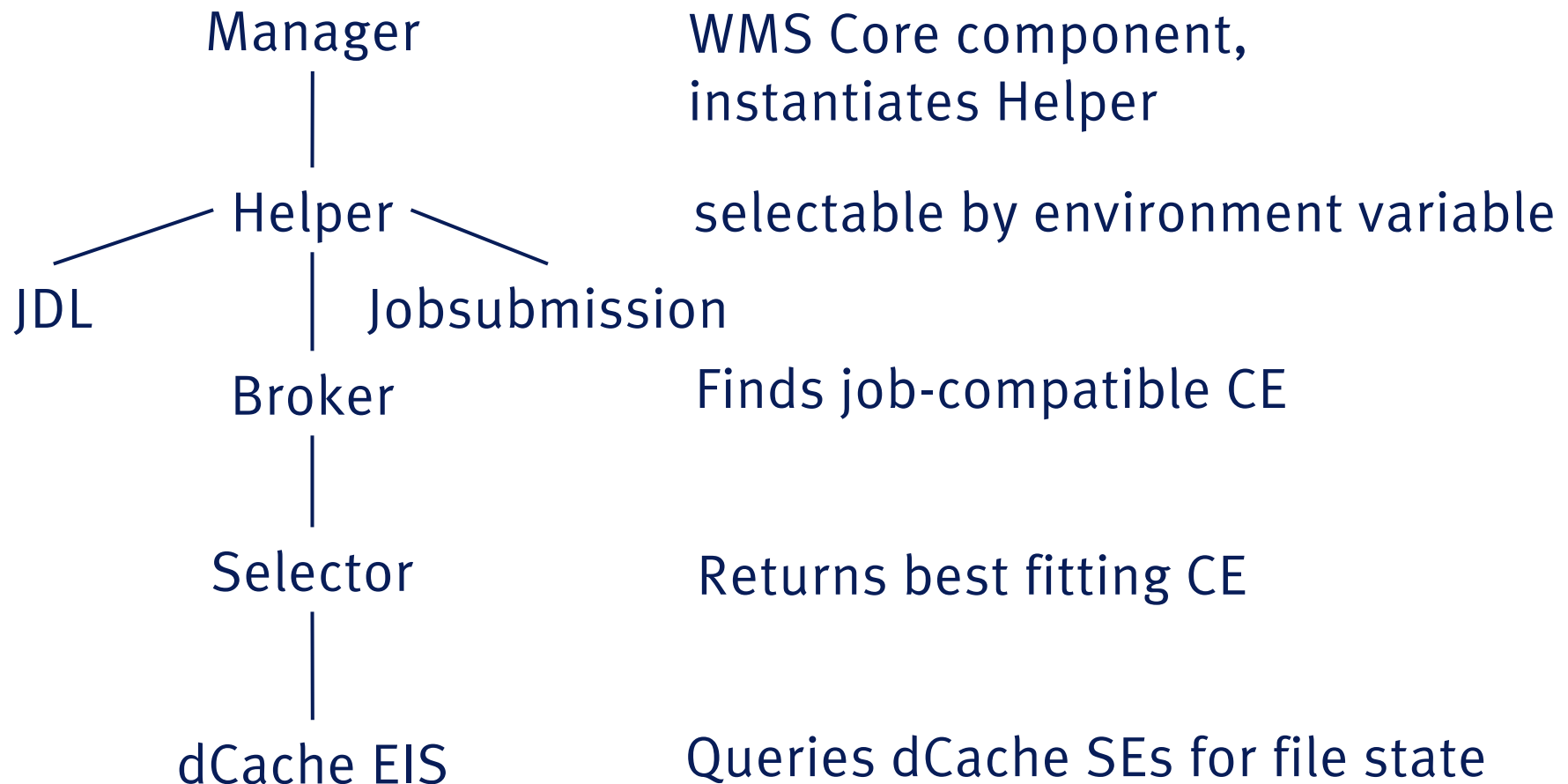
Workload Management System



- Receives jobs from the User Interface (JDL)
- Uses e.g. RGMA or MDS to find resources
- Resolves LFNs to PFNs and GUIDs through LFC
- Further actions and tests
 - ➔ Does the CE fulfill the requirements, e.g. from the jdl?
 - ➔ Data availability on the different SE
(Information of LFC query is used)
 - ➔ Send job to the CE whose close SE contain most of the required data
 - ➔ Possible inconsistencies between LFC entries and data on SE are not considered



Overview: Packages & Job Flow





Extension of the WMS (exemplified)



■ org.glite.wms.broker

➔ Implementation of a Resource Broker and Selector via defined interfaces

- RB fetches a multitude of VO specific CEs
- RB invokes the Selector which implements a Selection Schema
- Selection Schema defines how the “best fitting” CE is selected
- Selector interacts via dCache EIS with the SE to incorporate data state information into the selection process
- RB gets the “best fitting” CE from Selector
- Job data requirements are integrated into the submission description towards CE

■ If data replication is necessary a replication scheduler in dCache will be invoked



Extension of the CE (1)



- CE is extended by a Webservice-based scheduler and an adequate client
 - ➔ Both based on blahp source code
- Scheduler interacts between CondorG and the LRMS
- Scheduler client substitutes blahp



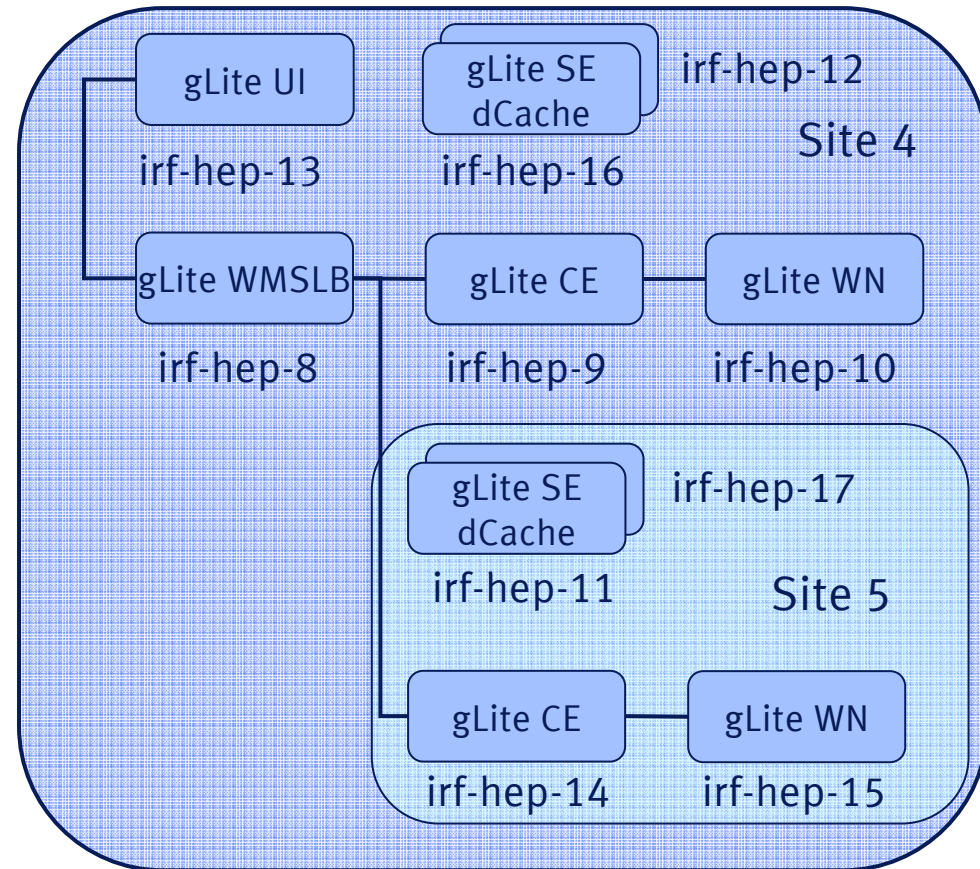
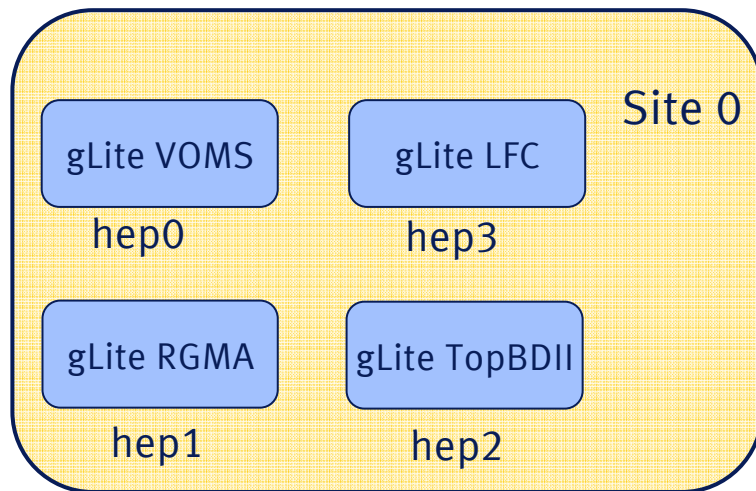
Extension of the CE (2)



- Initially jobs are set to hold status
 - ➔ Job Id and Job Description are propagated to the Scheduler
- Scheduler makes use of dCache EIS to check data availability/state
- Scheduler resumes jobs when data is available at the close SEs



Demo





Demo



■ JDL

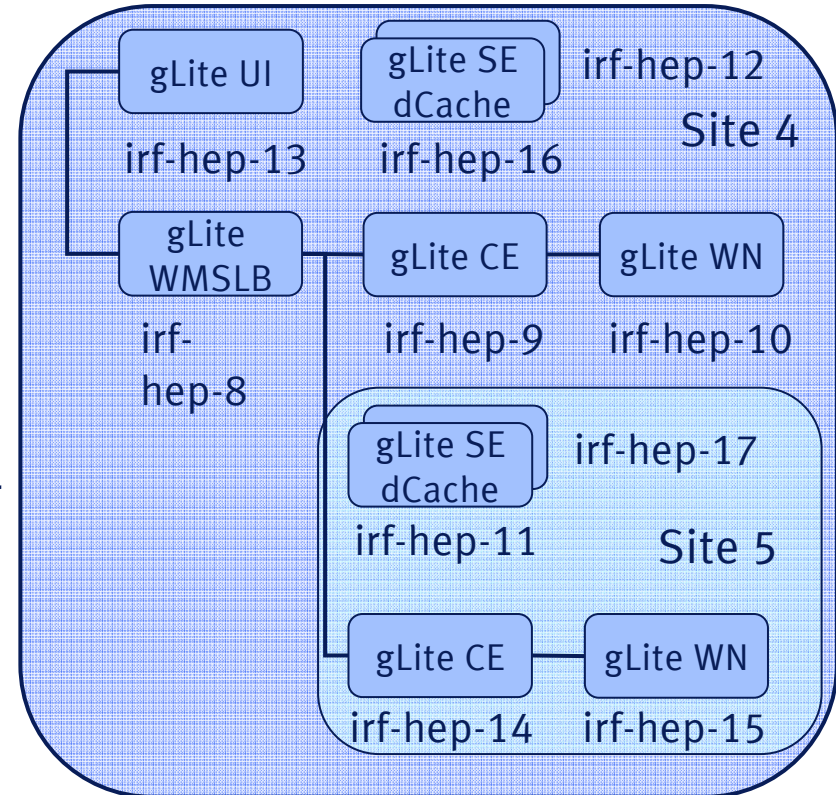
```
Type = "Job";  
JobType = "Normal";  
Executable = "/bin/lis";  
StdOutput = "input_files.out";  
StdError = "input_files.err";  
InputSandbox = {};  
OutputSandbox = {"input_files.err", "input_files.out"};  
Arguments = "";  
VirtualOrganisation="irfvo4";  
DataCatalog = "http://hep3.e-technik.uni-dortmund.de:8085";  
InputData = {"lfn:/grid/irfvo4/temp/test_data1"};  
DataAccessProtocol = {"gridftp", "srm", "gsiftp"};
```



Demo



- Send JDL from UI irf-hep-13 to WMS irf-hep-8
- WMS interacts via dCache EIS with SE irf-hep-11 (Site 5) and SE irf-hep-16 (Site 4)
- Both SE contain requested data
 - ➔ SE irf-hep-11: data is cached
 - ➔ SE irf-hep-16: data is not available (e.g. inconsistencies LFC - SE)
- CE close to SE irf-hep-11 is selected by Resource Broker
- Job forwarded to CE irf-hep-14





Demo



■ Logger in Workload Management

```
retrieveSFNsInfo: trying to resolve: lfn:/grid/irfvo4/temp/test_data1
put_results_in_bi_data: srm://irf-hep-11.e-technik.uni-dortmund.de/pnfs/...
    /file79e22236-705b-4edb-9a6a-93cc5e912a8b
put_results_in_bi_data: srm://irf-hep-16.e-technik.uni-dortmund.de/pnfs/ ...
    /file7f7119da-e8f8-4b7f-b604-fef40138ca48
retrieveSFNsInfo: finishing retrieveSFNsInfo
```

```
retrieveCloseSAsInfoFromISM: irf-hep-14.e-technik.uni-dortmund.de:2119/blah-
    pbs-long is close to irf-hep-11.e-technik.uni-dortmund.de mountable on /tmp
```

```
Storage Element:irf-hep-11.e-technik.uni-dortmund.de
```

```
LFN: lfn:/grid/irfvo4/temp/test_data1
```

```
SFN: srm://irf-hep-11.e-technik.uni-dortmund.de/pnfs/ ... /file79e22236-705b-
    4edb-9a6a-93cc5e912a8b
```

```
State: CACHED
```



Demo



■ Logger in Workload Management

retrieveCloseSAsInfoFromISM: irf-hep-9.e-technik.uni-dortmund.de:2119/blah-pbs-short is close to irf-hep-16.e-technik.uni-dortmund.de mountable on /tmp

Storage Element:irf-hep-16.e-technik.uni-dortmund.de

LFN: lfn:/grid/irfvo4/temp/test_data1

SFN: srm://irf-hep-16.e-technik.uni-dortmund.de/pnfs/ ... /file7f7119da-e8f8-4b7f-b604-fef40138ca48

State: NOT AVAILABLE

Best fitted CE is irf-hep-14.e-technik.uni-dortmund.de:2119/blah-pbs-long



Demo



List of cached files:

LFN: lfn:/grid/irfvo4/temp/test_data1

SFN: srm://irf-hep-11.e-technik.uni-dortmund.de/pnfs/ ... /file79e22236-705b-4edb-9a6a-93cc5e912a8b

List of available files:

List of files on other SE:

LFN: lfn:/grid/irfvo4/temp/test_data1

SFN: srm://irf-hep-16.e-technik.uni-dortmund.de/pnfs/ ... /file7f7119da-e8f8-4b7f-b604-fef40138ca48

Files NOT to replicate:

LFN: lfn:/grid/irfvo4/temp/test_data1

SFN: srm://irf-hep-11.e-technik.uni-dortmund.de/pnfs/... /file79e22236-705b-4edb-9a6a-93cc5e912a8b

Files to replicate:



Questions?

