NORDUGRID

Grid Solution for Wide Area
Computing and Data Handling

The ARC Information System

Quick overview
Available information

Maiken Pedersen, University of Oslo
Florido Paganelli, Lund University
On behalf of Nordugrid ARC

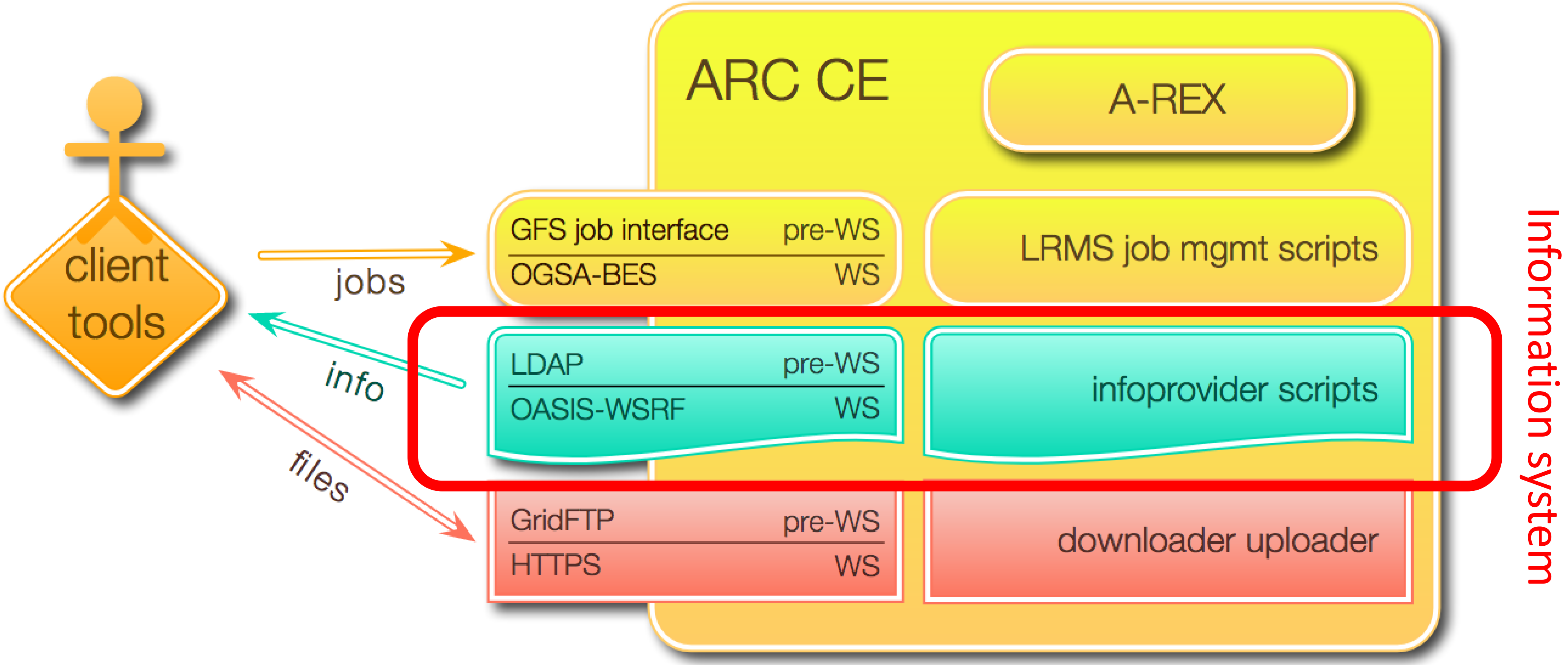
Outline

- ARC key concepts and technologies

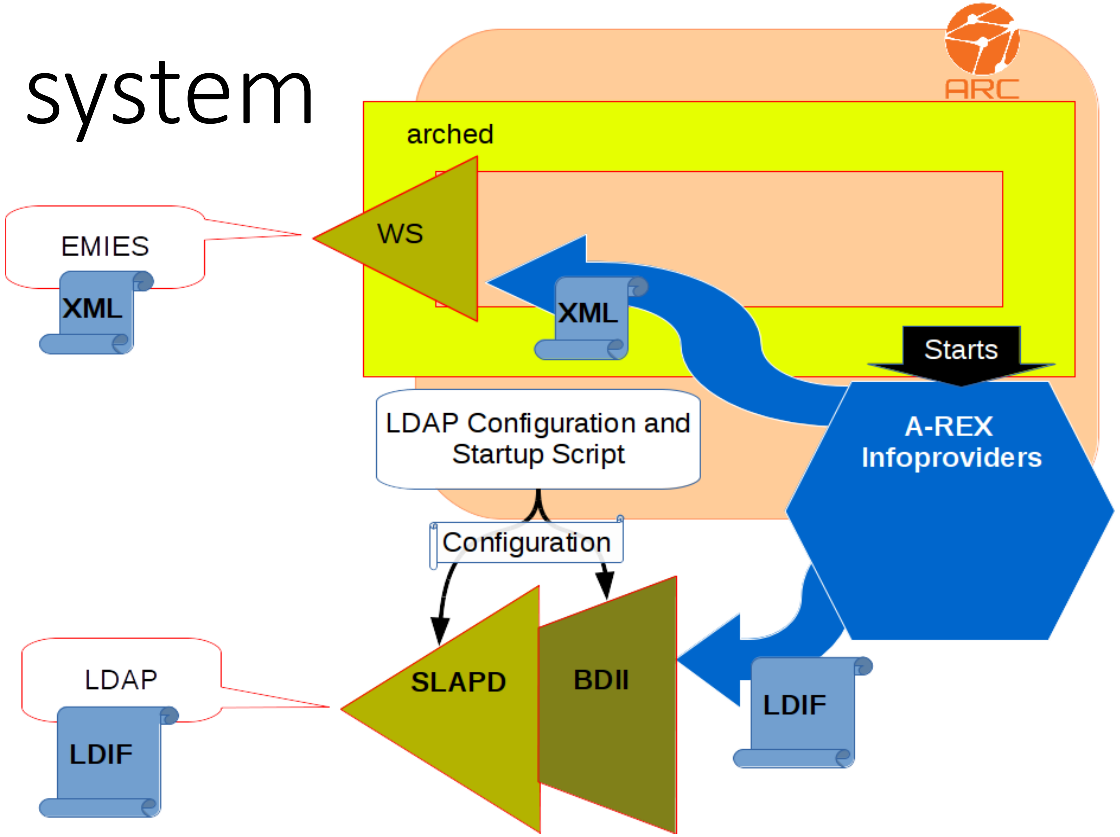
- Integration challenges past, present and future

- What information is available from ARC information system

ARC CE with and without (pre-) WS



Overview of ARC Information system



- ARIS: ARC Resource Information Service
  - Installed on the CE or storage resource
  - Publishes via
    - LDAP
    - OASIS-WSRF (OASIS)
  - Info on: OS, architecture, memory, running and finished jobs, what users are allowed to run, trusted certificate authorities, …
  - GLUE2.0 schema (GLUE 1.2, Nordugrid schema)
    - **GLUE2:** An attempt to unify information and make it independent from the specific LRMS
- Infoproviders collect dynamic state info
  - Grid layer (A-REX or GridFTP server)
  - Local operating system (/proc area)

- Interfaces to
  - UNIX fork
  - PBS-family
  - Condor
  - Sun Grid Engine
  - IBM LoadLeveler
  - SLURM

Output:

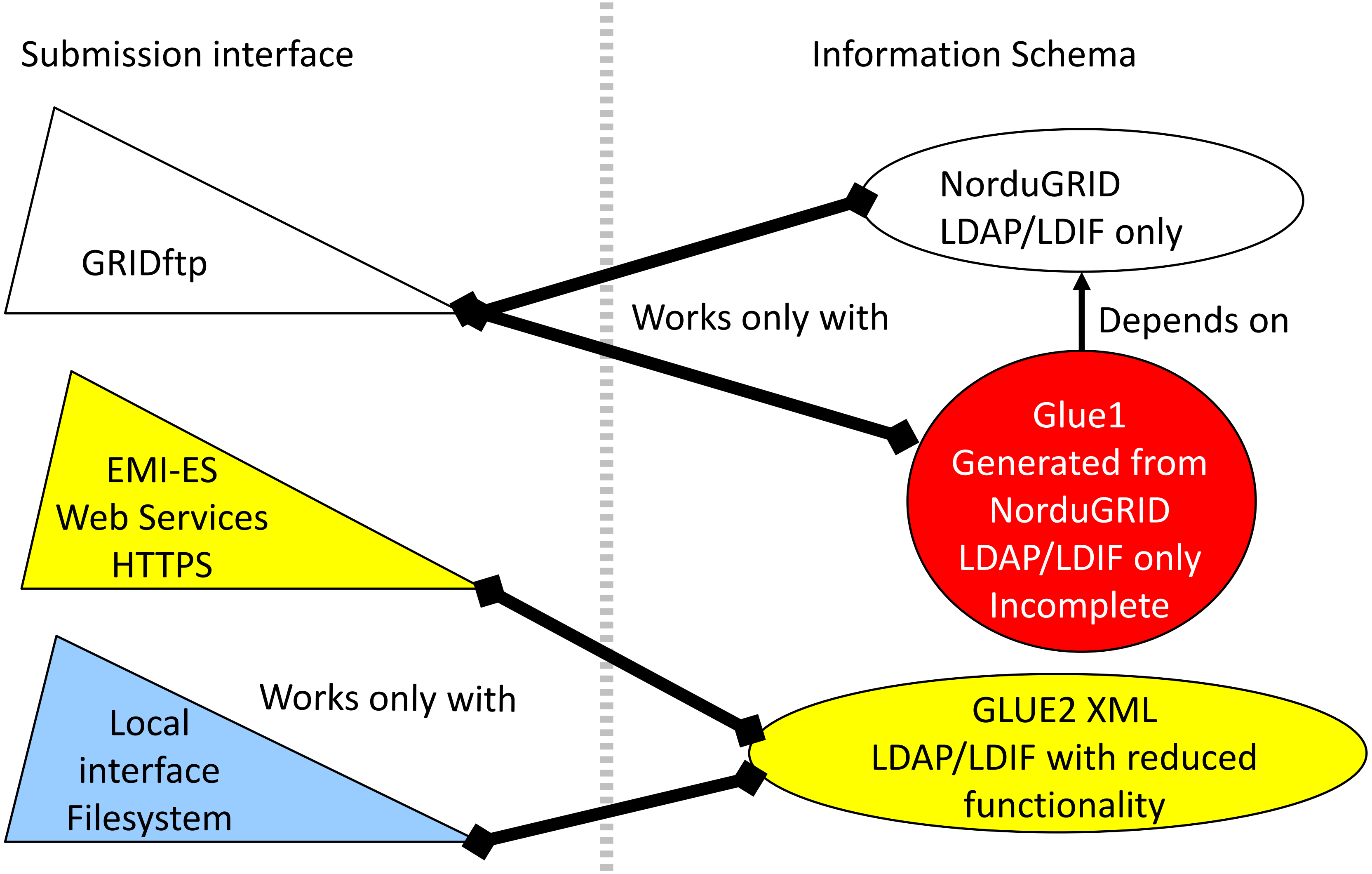
LDIF format populates local LDAP tree

XML format for OASIS-WSRF

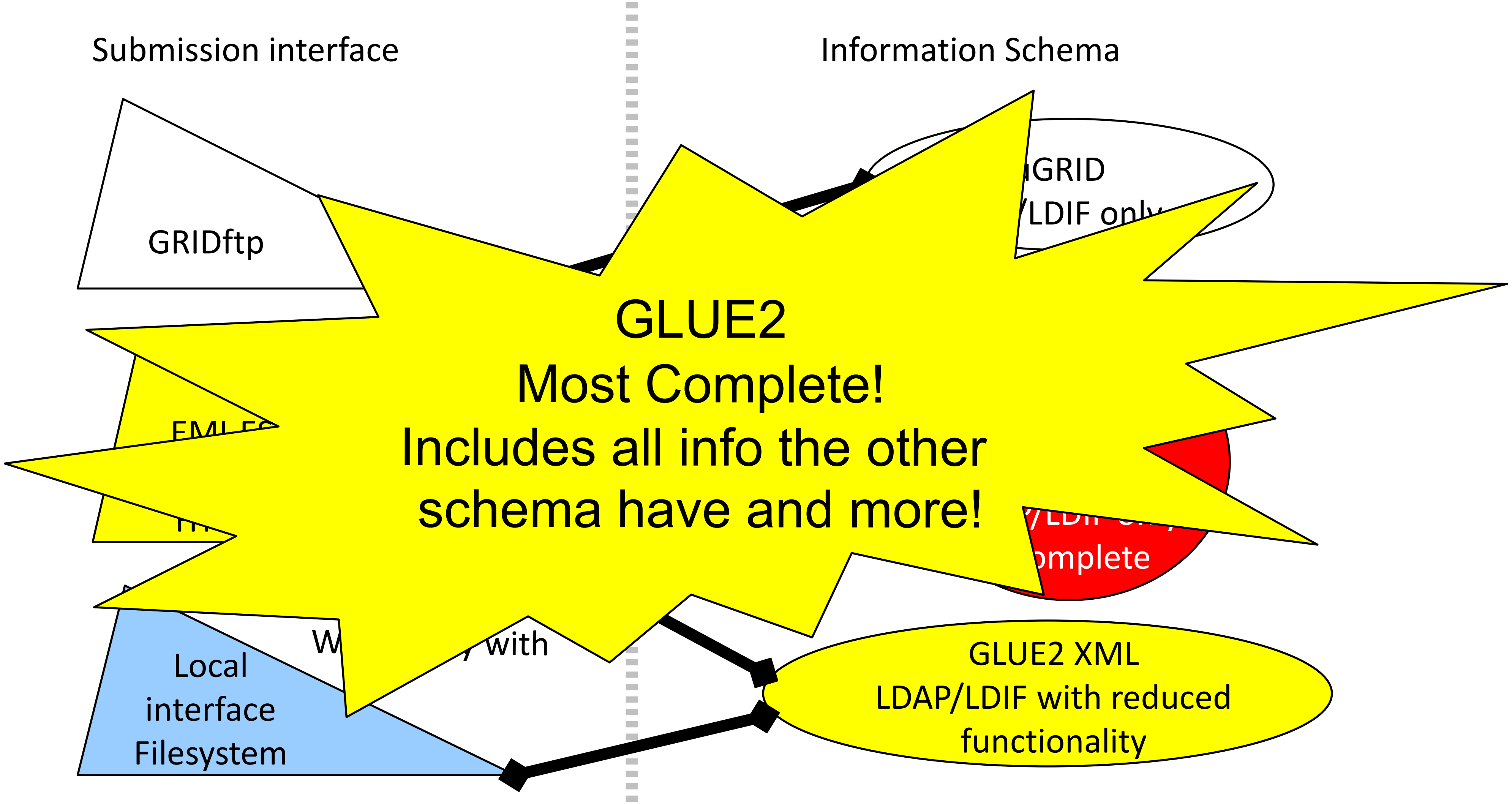
Briefly how info is collected and served

- Infoproviders are Perl/python LRMS modules that continuously run and collect information from the batch system
- Information is temporarily stored in an internal datastructure
- The content of this datastructure is reshaped according to the two official ARC information schema, NorduGRID and GLUE2
- The information is rendered as LDIF for LDAP and as XML for webservices

Supported interfaces and schemas



Supported interfaces and schemas



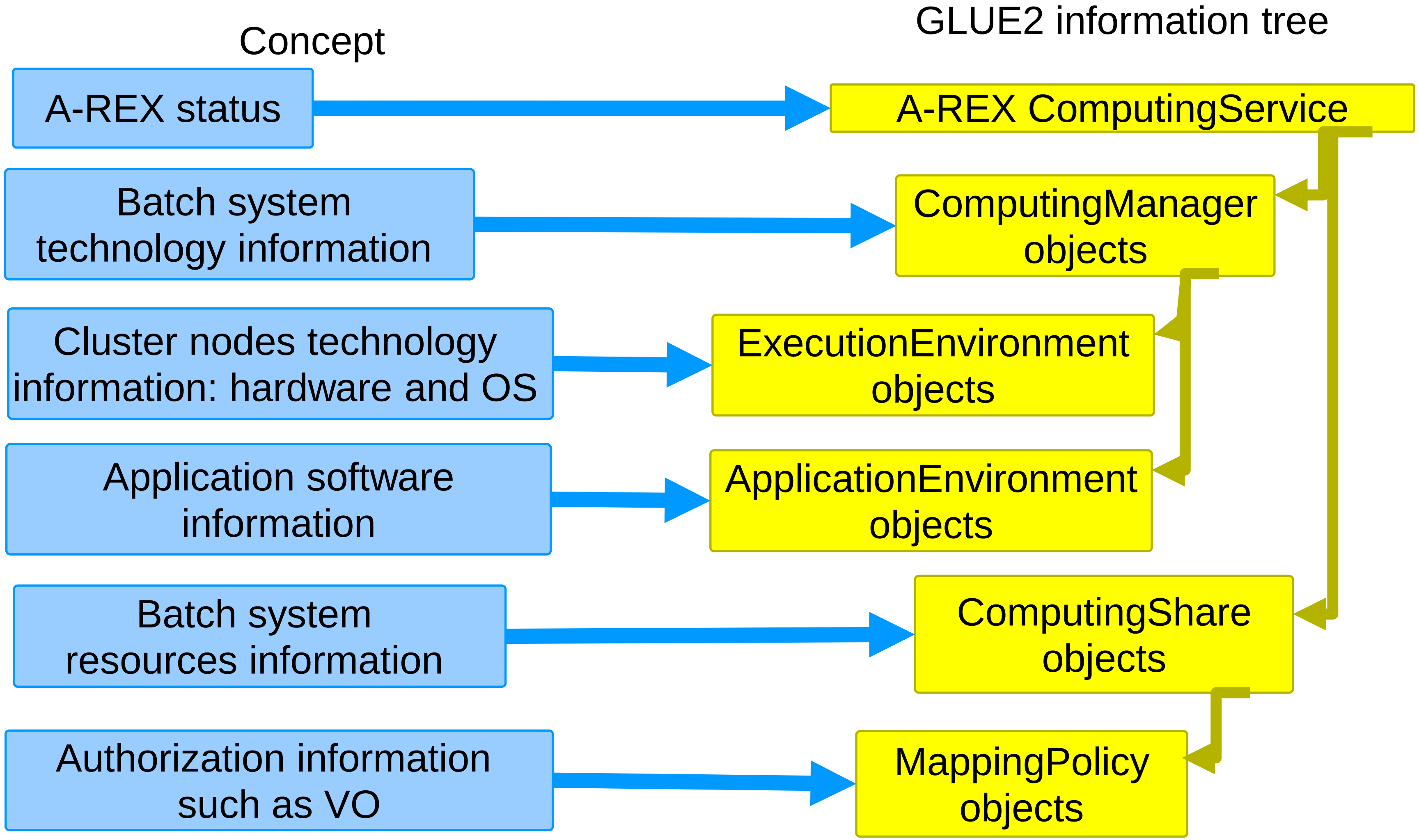
GLUE2
Most Complete!
Includes all info the other schema have and more!

Resource
Information Consumers

- LDAP:
  - Can simply be queried using ldapsearches
  - ldapsearch -x -h piff.hep.lu.se -p 2135 -b 'GLUE2GroupID=resource,o=glue’

- Web Services:
  - Communication encrypted (certificates)
  - EMI-ES: Any client capable of speaking this protocol
    - E.g. arcinfo –c arctest1.hpc.uio.no

GLUE2 mapping of relevant cluster information


Concept
GLUE2 information tree
A-REX status
A-REX ComputingService
Batch system technology information
ComputingManager objects
Cluster nodes technology information: hardware and OS
ExecutionEnvironment objects
Application software information
ApplicationEnvironment objects
Batch system resources information
ComputingShare objects
Authorization information such as VO
MappingPolicy objects

GLUE2 mapping of relevant info: Queues, Limits. Static

Cluster information	Sample value	GLUE2 object	GLUE2 attribute	Sample GLUE2 value
Authorized VO names	atlas	MappingPolicy	PolicyRule	vo:atlas
Batch system name	slurm	ComputingManager	ComputingManagerProductName	slurm
Batch system queue name	batch	ComputingShare	ComputingShareMappingQueue	batch
Name of **allocation** of batch queue for atlas	(created by A-REX)	ComputingShare	ComputingShareName	batch_atlas
Maximum allowed CPU time	7 days	ComputingShare	ComputingShareMaxCPUTime	604800 (seconds)
Maximum Wall time	7 days	ComputingShare	ComputingShareMaxWallTime	604800 (seconds)
Maximum usable memory for a job	16GB	ComputingShare	ComputingShareMaxVirtualMemory	16384 (MB)
The maximum allowed jobs for this share	5000	ComputingShare	ComputingShareMaxTotalJobs	5000
The maximum allowed number of jobs per grid user in this share	10000	ComputingShare	ComputingShareMaxUserRunningJobs	10000

GLUE2 mapping of relevant info: jobs statistics, dynamic

Cluster information	Sample value	**GLUE2 object**	GLUE2 attribute	Sample GLUE2 value
Jobs scheduled by A-REX not yet in the batch system	1	ComputingShare	ComputingSharePreLRMSWaitingJobs	1
Jobs scheduled by A-REX for which A-REX is performing data staging (downloading or uploading data)	1	ComputingShare	ComputingShareStagingJobs	1
Number of jobs waiting to start execution, in queue in the batch system, both non-grid and grid	262	ComputingShare	ComputingShareWaitingJobs	262
Queued non-grid jobs	67	ComputingShare	ComputingShareLocalWaitingJobs	67
Total number of jobs currently running in the system, non-grid and grid	1458	ComputingShare	ComputingShareRunningJobs	1458
Running non-grid jobs	1025	ComputingShare	ComputingShareLocalRunningJobs	1025
Total number of jobs in any state (running, waiting, suspended, staging, PreLRMSWaiting) from any interface both non-grid and grid	1721	ComputingShare	ComputingShareTotalJobs	1721

1721 = 1+262+1458

GLUE2 mapping of relevant info: job statistics, dynamic

Cluster information	Sample value	**GLUE2 object**	GLUE2 attribute	Sample GLUE2 value
Number of free slots available for jobs	1140	ComputingShare	ComputingShareFreeSlots	1140
Number of free slots available with their time limits	1140	ComputingShare	ComputingShareFreeSlotsWithDuration	1140:604800 #slots:duration [#slots:duration], a list
Number of slots required to execute all currently waiting and staging jobs	686	ComputingShare	ComputingShareRequestedSlots	686
Batch system overview of grid jobs	1774	ComputingManager	ComputingManagerSlotsUsedByGridJobs	1774
Batch system overview of non-grid jobs	8233	ComputingManager	ComputingManagerSlotsUsedByLocalJobs	8233

Formats and adding information

- The existing formats can be converted to e.g. json by Harvester ARC plugin
- Or: format of information can easily be provided as a json file on ARC side.

- Information needed by ATLAS which is not already provided:
  - Can be put on top of the glue2.0 schema, but:
    - will “pollute” the glue2.0 schema and must be used sparingly and after careful discussion and agreement

Summary

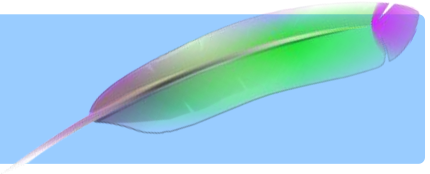
- ARIS is a powerful local information service:
  - *speaks* multiple dialects (NorduGrid, Glue1, GLUE2)
  - Creates LDIF and XML renderings
    - ARC is the only middleware that renders both
  - Offers information via both LDAP and Web Services
- Information consumers:
  - are easy to write as they would be based on well known standard technologies
  - Can get very accurate information directly from ARIS Resource level.
- Aiming at a **full GLUE2 support**
  - Hoping to phase-out of old technologies like LDAP or Gridftpd

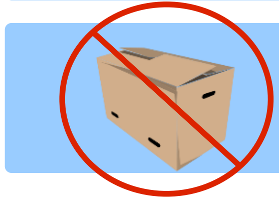
Backup – general and additional info

The ARC Information system key concepts

Information must be **as fresh as possible**.
The main source is therefore the
local or resource level

Index must be as lightweight as possible:
EGIIS contains just URLs.





No caching of information:
it only makes sense when fresh

Being able to ***speak*** all the possible
information system ***dialects***

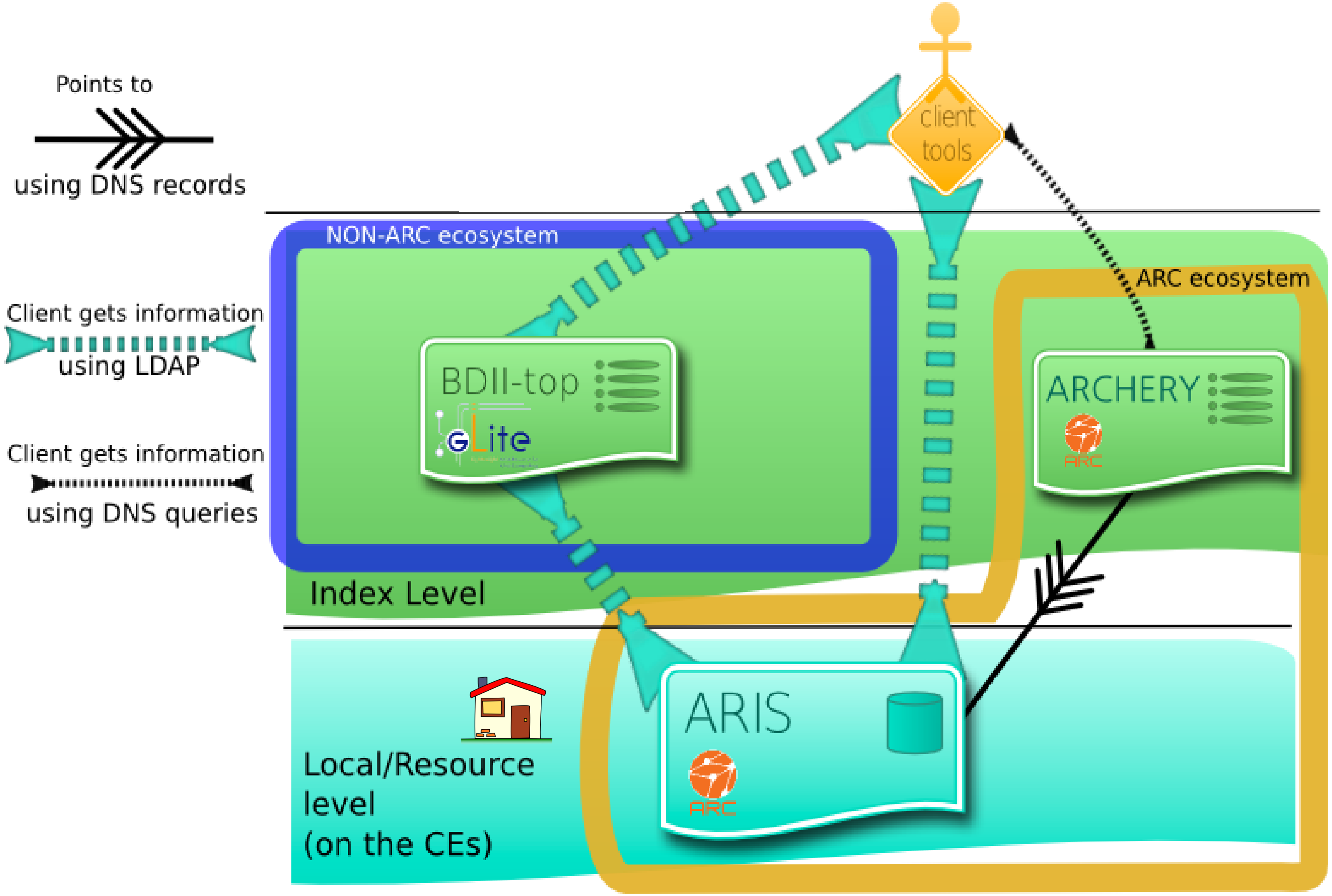
As a consequence, **clients** accessing the resources need to **discovery and query resources on their own**.

Technologies

- A collection of Perl scripts, the **infoproviders**, collecting information
  - LDAP:
    - Backend consisting of a
    - **schema** and the
    - **EGIIS** index
    - A dynamically populated **OpenLDAP** server and a collection of ldap Berkeley Database update scripts
  - **Web Services:**
    - Web Services Container: HED
    - ARCHED, serves info via **WSRF** based on **XML**

performance only depends on **slapd** and optimization of **perl scripts**

ARC IS: extended to "speak" other dialects



Points to

using DNS records

Client gets information

using LDAP

Client gets information

using DNS queries

client tools

NON-ARC ecosystem

BDII-top

gLite

ARC ecosystem

ARCHERY

ARC

Index Level

ARIS

ARC

Local/Resource level (on the CEs)

What GLUE2 VO info looks like

- One Share for the actual queue, without any VO information
- A Share for each VO, reporting specific VO numbers. Share names are of the form

  queuename_voname
- A MappingPolicy for each Share, should be used by clients for discovery

The ComputingShare and MappingPolicy concepts

- For each queue that is serving a VO, there is a ComputingShare and a MappingPolicy for such VO to represent the status of resources of that queue allocated to that VO.
  - Example: the queue called **batch** serving the VO **atlas** will be represented by a ComputingShare "called" **batch_atlas**
- **NOTE:** the name of a Share should **not** be used for discovery. It is written that way just for humans to be able to read at first glance.
- GLUE2 is a distributed database and using the attributes and unique IDs of its objects are the proper way to do it:
  1. Search for ID of the MappingPolicy associated with a VO
  2. Find the ComputingShare that serves that Policy ID.

1.Note that the use of unique IDs allows for large scale scheduling across clusters.

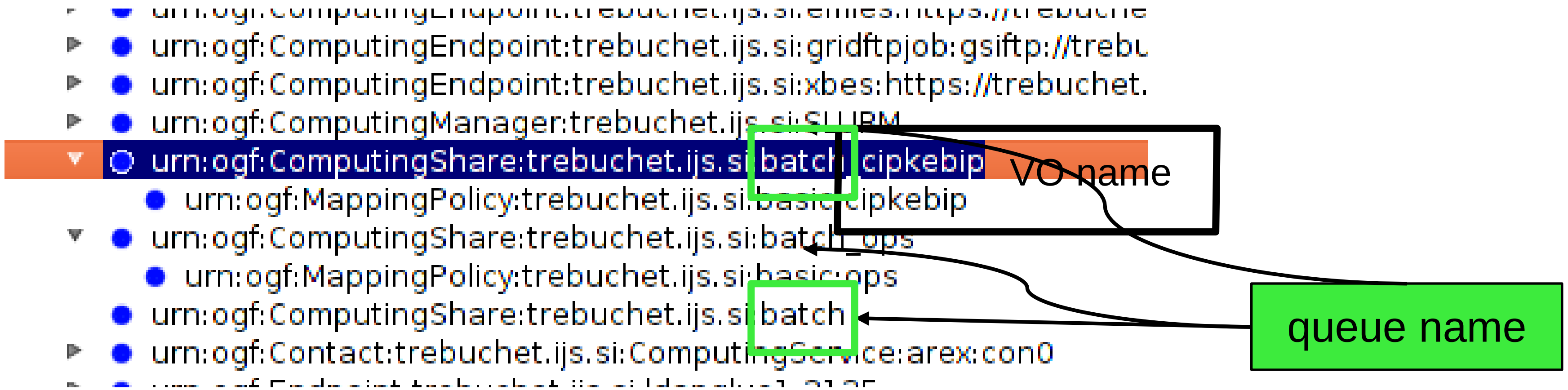
The ComputingShare and MappingPolicy concepts

- A **Share** is a set of resources. A **ComputingShare** is typically an allocation on a batch system queue, but can be anything that reserves resources.
- Every Share has a related **MappingPolicy**, that represents some kind of authorization associated to that share. So far there are no special authorization rules but the simple “Belonging to a VO” information
  - In Glue1 this information was somewhere in VoView
- In ARC there is a special ComputingShare for each batch system's queue with **NO MappingPolicy,** that represents the status of every queue as a whole, regardless of policy restrictions and resource allocations

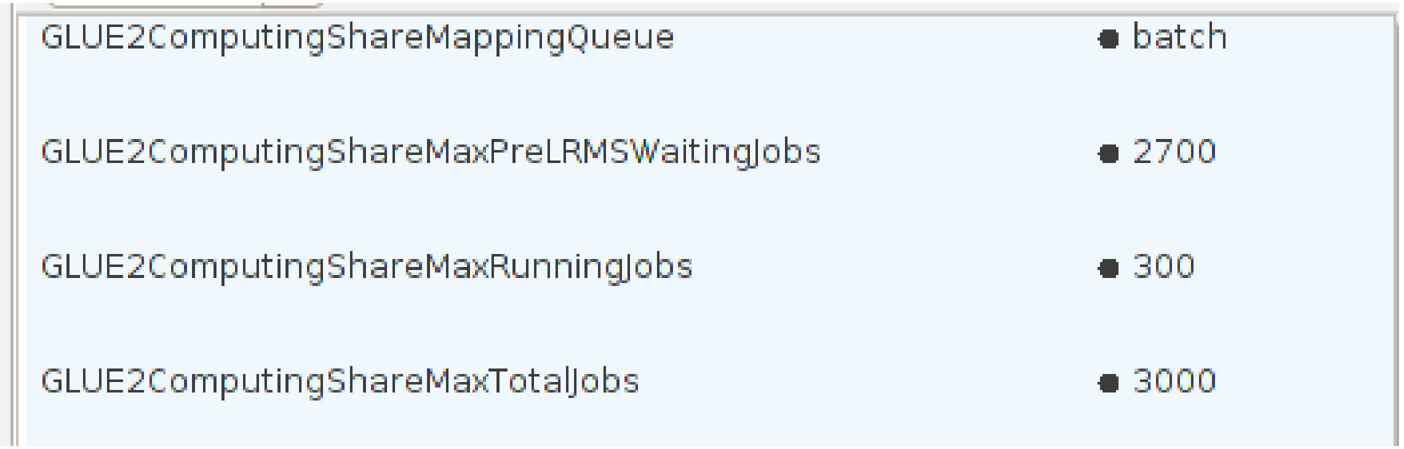
Ongoing developments related to available information

- VO support: based on GLUE2 ComputingShares
  - –Generic solution: gets VO info from user certificate. Sysadmin should just specify what are the expected VO names per queue/cluster

What VO info looks like



```
▸ ● urn:ogf:ComputingEndpoint:trebuchet.ijs.si:emies:https://trebuche
▸ ● urn:ogf:ComputingEndpoint:trebuchet.ijs.si:gridftpjob:gsiftp://trebu
▸ ● urn:ogf:ComputingEndpoint:trebuchet.ijs.si:xbes:https://trebuchet.
▸ ● urn:ogf:ComputingManager:trebuchet.ijs.si:SLURM
▾ ○ urn:ogf:ComputingShare:trebuchet.ijs.si:batch_cipkebip
     ● urn:ogf:MappingPolicy:trebuchet.ijs.si:basic:cipkebip
▾ ● urn:ogf:ComputingShare:trebuchet.ijs.si:batch_ops
     ● urn:ogf:MappingPolicy:trebuchet.ijs.si:basic:ops
  ● urn:ogf:ComputingShare:trebuchet.ijs.si:batch
▸ ● urn:ogf:Contact:trebuchet.ijs.si:ComputingService:arex:con0
▸ ● urn:ogf:Endpoint:trebuchet.ijs.si:ldapglue1:2135
```



GLUE2ComputingShareMappingQueue	● batch
GLUE2ComputingShareMaxPreLRMSWaitingJobs	● 2700
GLUE2ComputingShareMaxRunningJobs	● 300
GLUE2ComputingShareMaxTotalJobs	● 3000

Future developments

- VO:
  - MAY get VO info from LRMS that support it, but still not planned. Condor solution (ClassAds) far from being generic